PhD Thesis

presented in order to obtain the degree of:

Docteur de l’Ecole Nationale Supérieure des Télécommunications

Speciality: Signals and Images

STYLIANOU Ioannis

Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification

Committee of Referees

Christian WELLEKENS Chairman
Christophe d’ALESSANDRO Referees
Mohamed NAJIM
George CARAYANNIS Examiners
Xavier RODET
Geneviève BAUDOIN
Eric MOULINES
Olivier BOEFFARD Visitor
Mémoire de Soutenance de Thèse

présenté pour obtenir le titre de
Docteur de l’Ecole Nationale
Supérieure des Télécommunications.

Spécialité : Signal et Image

STYLIANOU Yannis

Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification.

SOUTENU LE 25/01 1996 DEVANT LE JURY COMPOSÉ DE

<table>
<thead>
<tr>
<th>Nom</th>
<th>Prénom</th>
<th>Titre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christian</td>
<td>WELLEKENS</td>
<td>Président</td>
</tr>
<tr>
<td>Christophe</td>
<td>d’ALESSANDRO</td>
<td>Rapporteurs</td>
</tr>
<tr>
<td>Mohamed</td>
<td>NAJIM</td>
<td></td>
</tr>
<tr>
<td>George</td>
<td>CARAYANNIS</td>
<td>Examineurs</td>
</tr>
<tr>
<td>Xavier</td>
<td>RODET</td>
<td></td>
</tr>
<tr>
<td>Geneviève</td>
<td>BAUDOIN</td>
<td></td>
</tr>
<tr>
<td>Eric</td>
<td>MOULINES</td>
<td></td>
</tr>
<tr>
<td>Olivier</td>
<td>BOEFFARD</td>
<td>Visiteur</td>
</tr>
</tbody>
</table>

Ecole Nationale Supérieure des Télécommunications
Enseignement Supérieur des Télécommunications
Acknowledgements

I wish to thank my main advisor Prof. Eric Moulines for his support throughout the course of this research. His high standard of excellence and integrity has been a great source of inspiration for me. Also thanks to Prof. Jean Laroche without whom this work could not have even been made. His willing to share with me his enthusiasm for research, his unlimited availability and his substantial help in my work are deeply appreciated.

I also wish to thank Dr. Olivier Cappé with whom I was fortunate to have the opportunity to collaborate on the voice conversion problem.

I have to acknowledge Prof. Christophe d’Alessandro and Mohamed Najim for having accepted to be examiners of my thesis despite their heavy schedule. Special thanks go to Prof. Christian Wellekens for accepting to be the president of the defense commitee and to Prof. George Carayannis, Prof. Xavier Rodet, Prof. Geneviève Baudoin et Dr. Olivier Boeffard for their participation in the jury. I appreciate the helpful comments and suggestions in the writing of my thesis from Prof. Geneviève Baudoin and Prof. Christophe d’Alessandro.

Also, thanks for the many good times to Vincent Doutaut, Alexis Le Pichon, Christophe Lambour, Karim Abed Meraim and Andrei Popescou.

I equally thank Simon and P. Weyer Brown for having read the text and for providing useful feedback and Olivier Vennard and François Yvon for checking the French summary.

I dedicate this piece of work to the memories of my father and of my friend Simeon. It is also dedicated to my mother and to my fiancée Katerina.
Contents

1 Introduction .......................................................... 1
   1.1 General background and problem definition ................. 1
   1.2 Scope of Thesis .................................................. 6

I SPEECH MODIFICATION ........................................... 11

2 Sinusoidal Models .................................................. 13
   2.1 Introduction ..................................................... 13
   2.2 Speech production model ...................................... 14
      2.2.1 The linear prediction formulation ...................... 14
      2.2.2 The sinusoidal formulation .............................. 14
      2.2.3 Time-scale and pitch-scale definition ................ 16
         2.2.3.1 Time-scale modification ............................ 16
         2.2.3.2 Pitch-scale modification ........................... 17
   2.3 Parameters estimation ........................................ 19
   2.4 Sinusoidal models: An overview ............................ 20
   2.5 Sinusoidal plus Noise Models .............................. 21
   2.6 TD-PSOLA ...................................................... 22
      2.6.1 The drawbacks of TD-PSOLA ........................... 23
   2.7 Conclusion ...................................................... 25
3 A Harmonic plus Noise Model, HNM

3.1 Introduction .................................................. 27
3.2 Description of HNM ............................................. 28
3.3 Estimation of HNM parameters .............................. 30
  3.3.1 Pitch, voicing and maximum voiced frequency estimation .... 30
    3.3.1.1 Initial pitch estimation ............................... 31
    3.3.1.2 Voiced/Unvoiced decision ............................ 33
    3.3.1.3 Maximum voiced frequency estimation ............... 33
    3.3.1.4 Refining the initial pitch estimation ................ 35
    3.3.1.5 The pitch detection algorithm ...................... 37
  3.3.2 Amplitude and phase estimation ........................ 39
  3.3.3 Estimation of the noise part parameters ............... 42
3.4 Phase and amplitude envelopes estimation ................. 43
  3.4.1 Phase envelope estimation .............................. 43
  3.4.2 Amplitude envelope estimation .......................... 45
3.5 Synthesis without modifications ............................ 50
  3.5.1 Synthesis of the harmonic part ........................ 50
  3.5.2 Synthesis of the noise part ............................ 52
3.6 An alternative method for the noise part .................... 53
3.7 Prosodic modifications .................................... 56
  3.7.1 Computation of synthesis time-instants ................ 56
    3.7.1.1 Time-scale modification ............................ 56
    3.7.1.2 Pitch-scale modification ............................ 57
    3.7.1.3 Joint time-scale pitch-scale modification .......... 59
    3.7.1.4 Mapping the synthesis time-instants to the analysis ones 59
  3.7.2 Computation of amplitudes and phases at the modified harmonic frequencies ........... 60
3.7.2.1 Synthesis using the modified sinusoids parameters .......................... 61
3.7.2.2 Intensity modification ............................................................................. 61
3.7.2.3 Other types of modification .................................................................... 61
3.7.2.4 The synthesis scheme for prosodic modifications ................................. 62
3.8 Example and discussion of the HNM application on speech ...................... 63
3.9 Conclusion .................................................................................................. 72

4 New decomposition techniques of the speech signal ........................................ 75
  4.1 Introduction ................................................................................................. 75
  4.2 Description of the DSM .............................................................................. 76
  4.3 Estimation of DSM parameters .................................................................. 78
    4.3.1 Choice of the weighting window $w(t)$ .................................................. 79
    4.3.2 The conditioning problem ..................................................................... 80
  4.4 Synthesis using the DSM ........................................................................... 80
  4.5 Example of synthesis using the DSM ......................................................... 81
  4.6 Refining the pitch estimation and passage from DSM to HNM .................... 83
  4.7 Description of the HSM ............................................................................ 85
  4.8 Estimation of HSM parameters .................................................................. 87
  4.9 Comparison of the three harmonic plus noise models ............................... 89
    4.9.1 Comparison of the residual signal variance .......................................... 91
    4.9.2 Comparison of the residual signal ....................................................... 91
  4.10 Conclusion ................................................................................................ 96

II SPEAKER MODIFICATION ............................................................................. 99

5 Background for speaker modification ............................................................... 101
  5.1 Introduction ................................................................................................. 101
5.2 Speaker modification ................................................. 102
  5.2.1 What is speaker modification? ................................. 102
  5.2.2 Purpose .......................................................... 102
5.3 Acoustical feature parameters for speaker identification .......... 103
5.4 Features selection .................................................... 105
5.5 Spectral transformation techniques .................................. 106
  5.5.0.1 Spectral transformation through vector quantization .... 107
  5.5.0.2 Speech spectrum transformation by speaker interpolation . 108
  5.5.0.3 Spectral transformation using Linear Multivariate Regression 109
  5.5.0.4 Spectral transformations using a Dynamic Frequency Warping (DFW) ................. 110
  5.5.0.5 Spectral transformations using Neural Networks .......... 111
5.6 Conclusion ............................................................ 112

6 Statistical methods for speaker modification 115

  6.1 Introduction ......................................................... 115
  6.2 The Gaussian Mixture Model ....................................... 116
  6.3 GMM parameter estimation ......................................... 117
    6.3.1 Maximum likelihood estimation ............................. 118
    6.3.2 Parameter estimation equations .............................. 120
      6.3.2.1 Mixture weights ...................................... 121
      6.3.2.2 Component density means ............................. 121
      6.3.2.3 Component density covariance matrices ............... 122
    6.3.3 The EM algorithm ............................................ 123
  6.4 Conversion functions .............................................. 124
    6.4.1 Optimization of the conversion function ................ 127
      6.4.1.1 Full conversion .................................... 127

vi
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4.1.2 Diagonal conversion</td>
<td>129</td>
</tr>
<tr>
<td>6.4.1.3 VQ-type conversion</td>
<td>130</td>
</tr>
<tr>
<td>6.5 Training of the conversion function</td>
<td>130</td>
</tr>
<tr>
<td>6.5.1 Practical issues on spectral vectors computation</td>
<td>131</td>
</tr>
<tr>
<td>6.5.2 The main steps for the transformation function learning</td>
<td>132</td>
</tr>
<tr>
<td>6.5.3 Corrective filters for the conversion of the noise part</td>
<td>134</td>
</tr>
<tr>
<td>6.6 The voice conversion system</td>
<td>135</td>
</tr>
<tr>
<td>6.7 Results and discussion</td>
<td>136</td>
</tr>
<tr>
<td>6.8 Conclusion</td>
<td>143</td>
</tr>
<tr>
<td>7 Summary of results and future research directions</td>
<td>145</td>
</tr>
<tr>
<td>7.1 Summary of results</td>
<td>145</td>
</tr>
<tr>
<td>7.2 Future research directions</td>
<td>149</td>
</tr>
<tr>
<td>A Another expression for the harmonic part</td>
<td>151</td>
</tr>
<tr>
<td>B Sound examples</td>
<td>155</td>
</tr>
<tr>
<td>C Résumé détaillé en français</td>
<td>161</td>
</tr>
<tr>
<td>C.1 Introduction</td>
<td>161</td>
</tr>
<tr>
<td>C.2 Le modèle HNM pour la parole</td>
<td>163</td>
</tr>
<tr>
<td>C.2.1 Estimation des paramètres des modèles</td>
<td>165</td>
</tr>
<tr>
<td>C.2.2 Comparaison des variances du signal résiduel</td>
<td>168</td>
</tr>
<tr>
<td>C.2.3 Comparaison des signaux résiduels dans le domaine temporel</td>
<td>168</td>
</tr>
<tr>
<td>C.2.4 Comparaison de l’erreur de modélisation</td>
<td>169</td>
</tr>
<tr>
<td>C.2.5 Le modèle pour la partie bruitée</td>
<td>171</td>
</tr>
<tr>
<td>C.2.6 Estimation des paramètres de la partie bruitée du HNM</td>
<td>172</td>
</tr>
<tr>
<td>C.2.7 Synthèse sans modification</td>
<td>172</td>
</tr>
</tbody>
</table>
C.2.8 Synthèse avec modifications prosodiques ........................................... 174
   C.2.8.1 Modification de l'échelle temporelle ........................................... 175
   C.2.8.2 Modification de la fréquence fondamentale ................................... 175
   C.2.8.3 Modification de la fréquence fondamentale et de l'échelle temporelle .................................................. 176
   C.2.8.4 Elimination des instants d'analyse répétés ..................................... 176
   C.2.8.5 Calcul des amplitudes et des phases aux nouvelles harmoniques 176
C.2.9 Exemple d'application ................................................................. 180
C.3 Modification du locuteur ............................................................. 185
   C.3.1 Applications .................................................................................. 186
   C.3.2 État de l'art .................................................................................... 187
   C.3.3 Description du système proposé ..................................................... 188
      C.3.3.1 Procédure d'apprentissage ....................................................... 188
      C.3.3.2 Classification probabiliste des données de référence ................... 189
      C.3.3.3 Fonction de transformation ....................................................... 191
      C.3.3.4 Optimisation de la fonction de transformation ........................... 191
      C.3.3.5 Transformation de la partie bruitée .......................................... 194
      C.3.3.6 Schéma de principe du système de modification du locuteur 195
List of Figures

2.1 Original and time stretched pitch contour by 0.6 ........................................ 16
2.2 Original and pitch contour modified by 0.6 ..................................................... 18
2.3 Pitch-scale modification with the TD-PSOLA method. Upper panel: original signal along with the analysis pitch- marks. Middle panel: three short-time synthetic signals. The mapping between these short-time signals and the analysis pitch- marks are indicated by the arrows. Lower panel: pitch-scale modified waveform along with the synthetic pitch-marks. The signal is a vowel / i /, uttered by a male speaker (pitch frequency around 100 Hz). The pitch-scale modification factor is equal to 0.8 (from [ML95]) ................................................................. 23
2.4 Time-scale modification with the TD-PSOLA method. Upper panel: original signal along with the analysis pitch- marks. Middle panel: three short-time analysis signals. The mapping between these analysis signals and the associated analysis pitch-marks are indicated by the arrows. Lower panel: time-scale modified waveform along with the synthetic pitch-marks. Same signal as in Fig.2.3. The time-scale modification factor is set to 2; the intermediate short-time signal is obtained by linearly interpolating the two adjacent waveforms (from [ML95]) ................................................................. 24
3.1 (a) Original speech signal (b) the magnitude spectrum of the speech signal (sampling frequency 16000 Hz) (c) the time-domain signal corresponding to frequency range 0 - 5000 Hz and (d) the time-domain signal corresponding to frequency range 5000 Hz - 8000 Hz ................................................................. 28
3.2 (a) Error function for a voiced frame. (b) Error function for an unvoiced frame. .... 32
3.3 (a) Speech signal overlayed by manually taken voiced/unvoiced decisions. (b) Voiced/unvoiced decisions produced by the algorithm. ......................................................... 34
3.4 Cumulative amplitude definition. ................................................................. 34
3.5 (a) Maximum voicing frequency estimation for a voiced frame (b) a voiced speech segment and (c) Maximum voicing frequency estimation for the voiced speech segment in (b)  .......................................................... 36

3.6 Spectrogram of the voiced speech signal in Fig.3.5(b). The x-axis is in msec and the y-axis in Hz. .................................................. 36

3.7 (a) Original et synthetic spectrum for the initial pitch estimation. (b) Original et synthetic spectrum for the refined pitch value. ........................................ 37

3.8 The pitch analysis algorithm. ................................................. 38

3.9 The time-domain envelope used for the time-behaviour of the noise part. The $t^i_a$ and $t^{i+1}_a$ represent two successive analysis time-instants. Typical values for $l_1$ and $l_2$ are : $l_1 = 0.15(t^{i+1}_a - t^i_a)$ and $l_2 = 0.85(t^{i+1}_a - t^i_a)$ .................................................................................. 42

3.10 Phase envelope estimation. a) for the first frame of a voiced segment and b) for the following frames of the same segment. ........................................ 44

3.11 Example of phase envelope estimation and phase unwrapping. ........ 44

3.12 Specified harmonic amplitudes (33 circles) and estimated spectral envelope for a cepstrum of order 20. (a) Direct estimation. (b) Regularized estimation with $\lambda = 2 \times 10^{-4}$ (from [CLM95]). .................................................................................. 47

3.13 Correspondence between the usual normalized frequency scale (in x-axis) and the normalized Bark scale (in y-axis) for sampling frequency 16 kHz. ............................... 48

3.14 Estimation of the spectral envelope for a voiced frame with maximum voiced frequency $F_M = 4000 \text{Hz}$ using a) linear frequency scale and b) normalized Bark scale ........ 49

3.15 For synthesis without speech modification the synthesis and the analysis time-instants coincide. ................................................. 50

3.16 Block diagram of the synthesis of the harmonic part for $t \in [t^i_a, t^{i+1}_a)$. ............... 52

3.17 Block diagram of the synthesis of the noise part for $t \in [t^i_a, t^{i+1}_a)$. ............... 54

3.18 Block diagram of the (a)analysis and (b) synthesis of the noise part using the alternative method ............................................ 55

3.19 Computation of the synthesis pitch-marks for time-scale modification. We have $P'(D(t)) \approx P(t)$ and $t^{i}_a = D(t^i_a)$. ............................................................... 57

3.20 Calculation of the synthesis pitch-marks for pitch modifications. ..................... 58

3.21 Elimination of repeated analysis time-instants ........................................ 59
3.22 Pitch modification in the frequency domain. (a) Original speech spectrum and (b) pitch modified spectral magnitude by a factor of 1.5 ................................................. 60
3.23 Block diagram of synthesis step using time-scale and pitch-scale modification .................. 63
3.24 Fundamental frequency and maximum voiced frequency estimation for a speech signal uttered by a male speaker. (a) the original speech signal (b) the estimated pitch contour (c) the estimated maximum voiced frequency .................................................. 64
3.25 (a) Harmonic part and (b) noise part synthesized from HNM. ........................................ 65
3.26 (a) Original speech signal and (b) the corresponding synthetic signal. ................................. 66
3.27 (a) Time-varying time-scale modification factors and (b) original pitch contour (dashed line) overlayed with the time-scaled pitch contour (solid line). ................................. 67
3.28 (a) Time-varying pitch-scale modification factors and (b) original pitch contour (dashed line) overlayed with the pitch-scaled pitch contour (solid line). ................................. 67
3.29 Wideband spectrograms of (a) the original speech signal and (b) the time-varying time-scaled synthetic one (the spectrogram has been broken into two overlapped parts). The x-axis is in sec and the y-axis in Hz. .................................................. 68
3.30 Narrowband spectrograms of (a) the original speech signal and (b) the time-varying time-scaled synthetic one. The x-axis is in sec and the y-axis in Hz. ................................. 69
3.31 Narrowband spectrogram of the simultaneously time-scaled and pitch-scaled synthetic signal. The x-axis is in sec and the y-axis in Hz. .................................................. 70
4.1 Variance of the least-square residual signal: no weighting window (solid line), Hamming weighting window (dashed line). .................................................. 80
4.2 Original speech signal (top) and the corresponding deterministic signal (bottom) ................. 81
4.3 Fourier spectrum of the stochastic part ................................................................................. 82
4.4 Stochastic part (top signal) of the speech signal of the preceding figure. The bottom signal is the synthesized stochastic part. .................................................. 83
4.5 Unwrapped phase of the original sinusoid (solid line), of the synthetic sinusoid without linear complex amplitude (dotted line) and of the synthetic sinusoid with linear complex amplitude (dashed line). .................................................. 84
4.6 Variance of the least-squares residual signals from: HNM (solid line), DSM (dashed line) and HSM (dash dotted line). The weighting window was a typical hamming window. .................................................. 91
4.7 (a) a fricative voiced of an original speech signal and the residual error signals from (b) HNM, (c) DSM and (d) HSM. ................................. 92
4.8 Magnitude of the Fourier transform of the residual signal (solid line) from (a) DSM and (b) HSM. The magnitude of the Fourier transform of the original speech signal has been also included (dashed line). ................................. 94
4.9 The same as in Fig.4.8 with an additive noise. ................................. 94
4.10 Comparison of the third order with the first order of the phase polynomial for the HSM. ............................................................... 95
4.11 Modelling error in dB using the three hybrid models. ....................... 96

6.1 Typical behaviour of the log-likelihood function for several iterations of EM algorithm for GMM training. ........................................... 133
6.2 Block diagram of the learning step. \( x_t \) represents the source spectral vectors and \( y_t \) the target spectral vectors. .................................................. 134
6.3 Block diagram of the corrective filters estimation for the conversion of the noise part. 135
6.4 Block diagram of the conversion step. \( t_1^s \): analysis time-instants, \( t_1^s \): synthesis time-instants. .................................................. 136
6.5 Normalized spectral distortion to target as a function of the number of components and the amount of training data in minutes. The solid line indicates using the Full-type conversion function and the dashed line using the Diagonal-type conversion function. The test data was 2000 spectral vectors (about 0.3 minutes). The first database was used. ............................................................... 138
6.6 Corrective filters for noise part transformation. The solid line is the psd of the source speaker while the dashed line is the psd of the target speaker. ............... 139
6.7 Normalized spectral distortion to source (left) and to target (right) as a function of the number of components. Using Full-type conversion function (solid line), Diagonal-type conversion function (dotted line), VQ-type (dashed line) and VQ-mapping (dashdot line). .................................................. 140
6.8 Short-term log-spectral distance between source and target data (solid line) as well as between transformed and target data for 100 short-term frames (\( \sim 1 \text{sec} \)) using 32 and 128 components and the three types of the conversion function: Full-type (dashed line), Diagonal-type (dashdot line) and VQ-type (dotted line). ....................... 141
6.9 Example of spectral conversion using (a) Full-type, (b) Diagonal-type and (c) VQ-type conversion function: Solid line for the source envelope, dashed line for the target envelope and dashdot line for the transformed envelope. 142
6.10 Fig.6.9(b). 142
6.11 Fig.6.9(c). 143
C.1 La variance du signal résiduel obtenue par: HNM₁ (trait continu), HNM₂ (trait interrompu) et HNM₃ (trait pointillé). La fenêtre de pondération était une fenêtre de Hamming. 168
C.2 (a) Une trame voisée fricative d’un signal de parole et les signaux résiduels obtenus par (b) HNM₁, (c) HNM₂ and (d) HNM₃. 169
C.3 Le spectre en dB des amplitudes de la transformée de Fourier des signaux résiduels pour HNM₂ (en haut) et HNM₃ (en bas). Le spectre en dB des amplitudes de la transformée de Fourier du signal original est aussi inclu (trait discontinu). 170
C.4 Le spectre en dB des amplitudes de la transformée de Fourier des signaux résiduels pour HNM₂ (en haut) et HNM₃ (en bas) avec un bruit additif. Le spectre en dB des amplitudes de la transformée de Fourier du signal original est aussi inclu (trait discontinu). 170
C.5 Erreur de modélisation en dB utilisant les trois modèles. 171
C.6 L’enveloppe utilisée pour la structure temporelle de la partie bruitée. Les $t_i^n$ et $t_{i+1}^n$ représentent deux instants d’analyse successifs. Des valeurs typiques pour $l_1$ et $l_2$ sont : $l_1 = 0.15(n_{i+1}^n - n_i^n)$ et $l_1 = 0.85(n_{i+1}^n - n_i^n)$. 173
C.7 Diagramme de la synthèse de la partie harmonique pour $n \in [n_i^n, n_{i+1}^n]$. 173
C.8 Diagramme de la synthèse de la partie bruitée pour $n \in [n_i^n, n_{i+1}^n]$. 174
C.9 Exemple de calcul des instants de synthèse pour une modification de la fréquence fondamentale d’un facteur 1.5 avec une élimination des instants d’analyse redondants. 176
C.10 Exemple de l’enveloppe spectrale utilisant une échelle fréquentielle (a) linéaire, (b) Bark. Fréquence maximale de voûtement $F_M = 4000Hz$. 178
C.11 Estimation de l’enveloppe spectrale. a) pour la première trame d’un segment voisé b) pour les trames suivantes. 179
C.12 Exemple de l’estimation de l’enveloppe de la phase et de sa déroulement. 179
C.13 Diagramme de la synthèse avec des modifications prosodiques. 180
C.14 Estimation de la fréquence fondamentale et de la fréquence maximale de voûtement sur un signal de parole prononcé par un locuteur masculin. (a) le signal original, (b) l'estimation du contour de la fréquence fondamentale et (c) l'estimation de la fréquence maximale de voûtement. ............................................. 181

C.15 (a) Partie harmonique (b) Partie bruitée. ......................................................... 182

C.16 (a) Un segment du signal original et (b) le segment correspondant du signal synthétique (partie harmonique + partie bruitée). .......................................................... 183

C.17 (a) Exemple de facteurs de modification de l'échelle temporelle, variants dans le temps. (b) Contour mélodique original (trait interrompu) superposé avec le contour mélodique du signal synthétique après la modification de l'échelle temporelle (trait solide). ... 183

C.18 (a) Exemple de facteurs de modification de la fréquence fondamentale (b) Contour mélodique original (trait interrompu) superposé avec le contour mélodique du signal synthétique après la modification de la fréquence fondamentale. ....................................... 184

C.19 (a) Signal original. (b) Signal synthétique après une modification de l'échelle temporelle avec un facteur de 1.3. ................................................................. 184

C.20 (a) Signal original. (b) Signal synthétique après une modification de la fréquence fondamentale avec un facteur de 1.3 ................................................. 185

C.21 Diagramme pour l'apprentissage de la fonction de transformation. ............... 194

C.22 Diagramme pour l'estimation de filtres correctifs. ........................................ 195

C.23 Filtres correctifs pour la transformation de la partie bruitée. En trait solide est la densité spectrale de puissance (dsp) de la référence et en trait interrompu la dsp de la cible. ......................................................... 195

C.24 Diagramme d'étape de conversion $t_a^i$: instants d'analyse, $t_s^i$: instants de synthèse. . . 196
Abbreviations

The following abbreviations will be used throughout the dissertation:

AR: Autoregressive
CELP: Code Excited Linear Prediction
DFW: Dynamic Frequency Warping
DSM: Deterministic plus Stochastic Model
DTW: Dynamic Time Warping
DFT: Discrete Fourier Transform
EM: Expectation Maximization
FD-PSOLA: Frequency Domain Pitch-Synchronous Overlap-Add
FFT: Fast Fourier Transform
GMM: Gaussian Mixture Model
HMM: Hidden Markov Model
HNM: Harmonic plus Noise Model
HSM: Harmonic plus Stochastic Model
LMR: Linear Multivariate Regression
LPC: Linear Prediction Coefficients
MFCC: Mel-Frequency Cepstrum Coefficients
PARCOR: Partial correlation
PSOLA: Pitch-Synchronous Overlap-Add
SOR: Successive Over-Relaxation
STFT: Short Time Fourier Transform
TD-PSOLA: Time Domain Pitch-Synchronous Overlap-Add
TTS: Text-to-Speech Systems
VQ: Vector Quantization

Fig. Figure
Hz Hertz
kHz Kilohertz
Mb Megabyte
pdf Probability density function
Mathematical Notation

The following mathematical notation will be used throughout the dissertation:

$(\cdot)^T$: Matrix or vector transpose
$(\cdot)^H$: Hermitian (complex conjugate) transpose
$Re(\cdot)$: Real part of a complex scalar
$Im(\cdot)$: Imaginary part of a complex scalar
$|\cdot|$: Modulus of a scalar
$\|\cdot\|$: 2-norm of a vector
$tr(\cdot)$: Trace operator
$E(\cdot)$: Statistical expectation
$\langle \cdot \rangle$: Integer closest to a real scalar
$\delta_{n,k}$: Kronecker delta
$\nabla_{\theta}(J)$: The gradient of function $J(\theta)$ with respect to $\theta$
$*$: Convolution operator
Chapter 1

Introduction

1.1 General background and problem definition

Although speech is one of the most important ways for people to communicate, this way is almost absent from the communication environments between man and machine. The main reason for this absence is that voice processing techniques related to the problem of man-machine communication do not yet provide high-quality communication. As computers become more prevalent, the need for communication between humans and computers is increasing. As the man-machine communication area becomes more and more necessary in activities such as information retrieval from speech databases, eyes-free hands-free communication or control and aid for handicapped persons, the usefulness of establishing a man-machine communication of better quality is also increasing. The areas of voice processing technology which are concerned with the above problem are speech recognition and speech synthesis. The first one consists of extracting the message information in a speech signal so as to control the actions of a machine in response to spoken commands, whereas the second one is the process of creating a synthetic replica of a speech signal so as to transmit a message from a machine to a person, with the purpose of conveying the information in the message [Rab94].

Our attention is restricted to the speech synthesis area. The goal of speech synthesis is to enable a machine to transmit orally information to a user[Rab94]. Three key factors influence the use of speech synthesis systems for different applications and determine also the 'ideal' synthesis system [All91] [Kla87] [Rab94]:

1
CHAPTER 1. INTRODUCTION

- the quality of the synthesized speech as measured in terms of intelligibility and naturalness
- the fluency of the spoken output i.e., the ability to create messages with a wide range of vocabulary, emphasis, intonation, speed, etc.
- the complexity of the synthesis hardware as measured in terms of both storage and computation.

Unfortunately there is no practical system with a performance that even comes close to the ideal.

Speech synthesis systems can be realized as either simple concatenation systems or as full text-to-speech systems. The first class has a stored vocabulary of pre-recorded and digitized speech messages (limited vocabulary) and it can be used as announcement machine in applications such as local time, weather, etc. These systems have low complexity and high-quality synthesis since they use pre-recorded words and phrases. However, they have low fluency since they can only speak the pre-recorded messages or minor combinations of them. When the messages are often changed in the course of a day (unlimited vocabulary) then we need the full text-to-speech systems (TTS). Some of the applications which require this type of system are the voice servers for reading e-mail and FAX messages over phone lines, flight information, standard database access services such as voice banking, stock price quotations, sports scores, and finally services that require the ability to speak unlimited, unconstrained text as found in medical textbooks, legal volumes and encyclopaedias. Although, full TTS systems provide low to medium quality at relatively high complexity[Rab91] they have high fluency since they can use unlimited vocabulary. For this reason the full TTS are more interesting than the first class (of limited vocabulary).

A full TTS system maps a sequence of characters of a text into another sequence of numbers representing the samples of an acoustic waveform. There are three main steps to achieve this [LC91] [Rab94]:

- Linguistic representation of the message encoded in the original text. This step includes the analysis of the text and the word pronunciation. The text analysis involves the division of the input letter sequence into words, dividing the word sequence into sentences, assigning parts of speech to the word tokens, and parsing them into intonation phrases. Word pronunciation translates spelling (orthography) into a phonological representation of the sort found in most dictionaries, i.e.: distinguish /diˈstɪŋɡwɪʃ/. English, French and Chinese languages pose problems at this step as opposed to Spanish, Italian, Greek and Japanese where spelling is related to pronunciation in a more regular way [LC91].
1.1. **GENERAL BACKGROUND AND PROBLEM DEFINITION**

- **Phonetic interpretation.** This phase of a TTS system assigns quantitative phonetic values (prosody markers) to the various aspects of this linguistic representation: duration of phonetic segments, fundamental frequencies target values for pitch accents, and so forth (intonation).

- **Signal generation.** This step uses the above detailed phonetic specification to produce time functions of the control parameters for an acoustic speech synthesis model, which are then used to calculate the samples of the speech waveform.

In order to improve the quality in terms of naturalness of current full TTS systems, three areas must be addressed [Cok76],[Fla76] [FR73] [Rab94] :

- Improved linguistic analyses
- Improved prosody rules
- Improved speech synthesis models

Only the last of the above areas makes use of digital speech (or general signal) processing techniques, and therefore it is in our interest.

For full TTS, there are two main categories of speech synthesis techniques :

1. Formant synthesis or synthesis by rule and

2. Synthesis using concatenated stored parametric units such as diphones or longer units.

The second method provides a better quality synthesis than the first one; using diphones [PWS58] or larger units [Siv61] minimizes the coarticulatory influences at the acoustic centre of phoneme as opposed to the first method where the transition from one phoneme to another is given by rules (we do not yet know the optimal rules to do this [Kla87]). In order to preserve the naturalness of synthetic speech, the text-to-speech systems based on concatenated units require smoothing at the units' boundaries, changing the duration of the units and finally imposing a fundamental frequency contour different from that originally recorded (prosodic modification or speech modification).

The advent of linear prediction (LPC)[ARS82] [AS75] speech analysis/re-synthesis techniques opened up the possibility of automated procedures for the creation of a diphone inventory and offered the possibility of changing the prosodic features of the recorded units [OS76] [OL85]. However, the buzziness inherent in LPC degrades perceived voice quality. In order to improve the voice quality of the TTS systems based on acoustical units concatenation a family of *pitch-
synchronous waveform processing techniques have been proposed. These algorithms rely on a pitch-synchronous overlap-add (PSOLA) approach for modifying the speech prosody and concatenating speech waveforms. The modifications of the speech signal are performed either in the frequency domain (FD-PSOLA) using the Fast Fourier Transform (FFT), or directly in the time domain (TD-PSOLA) [CS86] [MC90] [ML95] [Mou89] [HMC89]. The TD-PSOLA provides very efficient solutions for the real time implementation of synthesis systems. It works with short-time signals extracted directly from the original speech signal. Today it is perhaps the most important and widely used method for text-to-speech based on acoustical units concatenation [Duf94][DL93] [DL92] [dGSP94][BBC+93].

However, some limitations in the TD-PSOLA approach appeared after an extensive use of the method; these limitations are basically related to the non-parametric structure of the TD-PSOLA; when the prosodic features of the original acoustic units are changed by large modification factors (this is sometimes the case for the modifications needed in practice) a tonal noise in the 'synthetic' signal is perceived. The reason for this tonal quality is that the TD-PSOLA works by eliminating/duplicating short-time waveforms also for the noise-like signals (as unvoiced and fricative voiced speech segments). This produces an artificial long-time autocorrelation in the output signal, perceived as some sort of periodicity [ML95]. Thus, the TD-PSOLA method is recommended only for moderate modification factors (for example when the time-scale factor is less than 2).

Another problem arises when it is desired to alter the voice of the pre-recorded spoken messages. It seems reasonable that the solution is to record another speaker. However, this is not appropriate, as large databases will be created when the same problem (the creation of a third voice, etc.) re-occurs. For example, for the French language the acoustical units which are used for TTS are about 1500 elementary signals : 1200 diphones and 300 larger acoustical units which means about 10Mb at 16kHz sampling frequency [Mou98]. Also the recording period is long and very tiresome; the database must all be recorded by a speaker who can control (hold constant) voice quality so that there are not sudden changes in the source spectrum in the middle of syllables. The TTS systems which adapt the above solution are not flexible and can be viewed as limited voice database TTS systems (like TTS systems with limited and unlimited vocabulary).

A more flexible, simple and efficient way to bring the variety needed in TTS systems and to obtain an unlimited voice database TTS system is the voice conversion (or speaker modification) approach. Voice conversion consists in modifying the characteristics of the speech signal uttered by a speaker in such a way that a human listener could believe that the transformed speech is produced by another speaker (target speaker) which could be a specific person or not. When the target speaker is a specific one, then this approach needs some common utterances of few minutes (typically three to four minutes) from the pre-recorded speaker (source speaker) and
1.1. GENERAL BACKGROUND AND PROBLEM DEFINITION

the target speaker.

Speaker modification is in general a very difficult task and the problem has only been posed the last decade. For this kind of modification almost all researchers agreed that the prosodic modifications should be combined with spectral modifications. Unfortunately, the TD-PSOLA cannot be used in this case. In an attempt to give solution to the speaker modification problem, Valbret et al. [VMT92] proposed the combination of the LPC with the TD-PSOLA; prosodic modifications are applied on the excitation signal using the TD-PSOLA scheme [Mou89] and then the modified excitation signal is filtered through a filter which aims to model the vocal tract of the new speaker. In order to obtain this filter, spectral modifications are performed on the source spectrum using Linear Multivariate Regression (LMR) and Dynamic Frequency Warping (DFW) [VMT92] [Val92]. Other methods for spectral transformation have been also proposed in the literature [ANSK88][ANSK90], [SN91], [KS95][IS94] [IS95] [MA94] [MA95] (but not all of them have been combined with LPC). However, the obtained voice quality suffers from artefacts (bubbles and other oddities) which could be attributed both to the analysis/synthesis system used (i.e.: LPC), and to the method used for the spectral transformation.

This thesis will focus on the problem of high-quality speech and speaker modification. The main application is text-to-speech synthesis based on acoustical units concatenation. This means that:

- First, the proposed system must be able to extract appropriate parameters from a speech signal and then, using the stream of these parameters to produce a high quality synthetic signal which should be indistinguishable from the original one. In an application like TTS using concatenation of acoustical units it is necessary to control the prosodic features of the speech in order to obtain a synthetic signal characterized by naturalness. This also means that the proposed system must provide a flexible manner to change the prosodic parameters (pitch, rhythm, intensity) and also to allow time-varying prosodic modifications yielding the quality of the modified synthetic speech signal.

- Second, in TTS using concatenation of acoustical units is often desirable to alter the pre-recorded voice. As the short-term spectral characteristics of speech play a predominant role in speaker individuality, the adapted voice conversion approach should be able to control the spectral envelopes of speech in a robust and efficient way.

Obviously, such a system which is able to modify the speech and/or the speaker could be also applied to areas others than text-to-speech synthesis. For example, foreign language learning, audio monitoring or film/soundtrack post-synchronization, psychoacoustic research, voice individuality disguise for secure communications, voice individuality restore for interpreting telephony, ... . Concluding this section, we summarize: the object is speech and speaker
CHAPTER 1. INTRODUCTION

modification. Our purpose is to provide a parametric model which will be able in a flexible and robust way:

1. to extract from a speech signal appropriate parameters for our purpose.
2. to use the above parameters to obtain high-quality speech synthesis.
3. to modify the prosodic parameters for speech modification.
4. to control the spectral envelopes for spectral transformations.
5. to combine the prosodic and spectral modifications for speaker modification.

1.2 Scope of Thesis

The main objective of this thesis is to develop a system for high-quality speech and speaker modification. To achieve this objective a new speech model based on a harmonic plus noise representation of the speech signal is introduced. The proposed model is referred to in the following as the Harmonic plus Noise Model or HNM. It will be demonstrated that this model is capable of producing high quality speech synthesis and modification. The HNM combined with statistical methods for the spectral envelopes control will also be shown to be a robust and flexible system for speaker modification.

The major contributions of the work presented in this thesis are:

- A parametric pitch-synchronous model based on a harmonic plus noise representation for the speech signal. This twofold representation allows to apply different modification methods to each part (harmonic and noise part), yielding more natural resynthesis.

- A joint time-domain and frequency-domain representation of the noise part of the speech signals. As we will see, high frequencies (for example from $4000\,Hz$ to $8000\,Hz$ (sampling frequency : $16000\,Hz$)) extracted from speech signals, exhibit a specific time-domain structure in terms of energy localization which is synchronized with the glottal pulses in voiced fricatives. Taking into account this time-domain structure of the noise part, the quality of the synthetic signal is improved.

- A study of faster techniques for solving the linear set of equations suggested themselves when harmonic amplitudes and phases were to be estimated.
1.2. **SCOPE OF THESIS**

- The proposal of a new model called the Deterministic plus Stochastic Model which decomposes the speech signal into two parts: a deterministic and a stochastic part. This model makes use of harmonically related sinusoids with linearly varying complex amplitudes for the representation of the deterministic part.

- The proposal of an original model called the Harmonic plus Stochastic Model, HSM, which aimed also to decompose the speech signal using harmonically related sinusoids. The HSM makes use of a third order polynomial of real coefficients for the harmonic amplitudes in order to represent the harmonic part efficiently.

- The proposal of a pitch and maximum voiced frequency estimation technique adapted to sinusoidal models. The estimation of the above parameters is critical for the quality of the speech synthesis.

- Exploration of the pitch-synchronous scheme of the proposed system to develop a flexible technique for time-varying pitch and time-scale modification inspired by PSOLA methods.

- A novel technique for frequency and time-domain phase unwrapping needed for pitch modification. This leads to a reliable re-computation of the phases for the modified pitch harmonics.

- Application and evaluation of the Gaussian Mixture Model, GMM, to the description of the acoustic space of a speaker in order to increase the robustness of the speaker modification. In contrast to Vector Quantization, VQ, the GMM provides continuous 'smooth' classification indexes which avoid unnatural discontinuities. This feature of the GMM naturally overcomes the interpolation problem associated with the VQ approach and also provides a more accurate description of the spectral envelopes distribution.

- The proposal of new spectral transformation functions which exploit the soft segmentation of a speaker acoustic space, derived from the Gaussian Mixture Model. This gives robust spectral transformations. The proposed functions are linear within each acoustic class and they are aimed at modifying the posterior mean and covariance of each component of the mixture.

- The proposal of a voice conversion system which combines the capability of the HNM for high-quality speech modifications with the robustness of the proposed spectral conversion functions. The aim being to convert the voice of a speaker, called the source speaker, to another speaker, called the target speaker.
CHAPTER 1. INTRODUCTION

The thesis is organized as follows:

• PART I

Chapter 2 provides a background review of the sinusoidal models and the hybrid models and introduces notation and nomenclature used throughout, mainly in the first part of the thesis. The chapter begins with a description of the basic speech production model on which the sinusoidal model is based on. Next, the pitch and time-scale modification (based on the above speech production model) are defined. This is followed by an outline of some of the major sinusoidal and hybrid models used in speech synthesis and modification. Finally, the TD-PSOLA system is briefly presented and its limitations for speech modification are discussed.

Chapter 3 introduces the Harmonic plus Noise Model, HNM, for the speech signals and pitch and time-scale modifications using the HNM are presented. The HNM is first described and an example is given to promote its use for speech modelling. The pitch and maximum voiced frequency estimation is then described. This is followed by the estimation of the harmonic amplitudes and phases using a purely time-domain criterion, which allows the estimation of harmonic parameters on short-time frames (typically two pitch periods). It is shown that writing the harmonic part as a sum of exponential functions gives the basic matrix of the system that we have to solve for amplitudes and phases estimation, a Toeplitz structure. Thus, the system can be solved efficiently using the Levinson algorithm. Next, the estimation of the spectral and phase envelope is presented followed by the synthesis without and with speech modification. Finally, the experimental evaluation of the HNM on a large database is discussed.

Chapter 4 presents two original models for speech, aimed at decomposing the speech signal into two parts: a deterministic or harmonic part and a stochastic part. The first model, called DSM makes use of linearly varying complex amplitudes to model the deterministic part. The stochastic part, is then the residual signal resulting from the subtraction from the original speech signal of the deterministic part. The stochastic part is represented both in time-domain and frequency domain. The DSM is first presented and this is followed by the estimation of DSM parameters. Next the second model called HSM is described and then the estimation problem of HSM parameters is addressed. The HSM makes use of a third order polynomial of real coefficients for the harmonic amplitudes and a first or third order polynomial for the harmonic phases. The representation of the stochastic part of the HSM is exactly the same with the DSM. Finally a comparison between the three hybrid models proposed in this first part of the dissertation, the HNM, the DSM and the HSM is presented.
1.2. SCOPE OF THESIS

• PART II

In Chapter 5 a background review of the area of speaker modification is given as well as the notation and nomenclature used in the second part of the thesis. The purpose of speaker modification (voice conversion) and the factors involved in voice individuality are discussed. Some known techniques used for speaker modification are presented including the VQ approach, LMR and DFW approaches and finally speaker modification using neural networks.

Chapter 6 introduces the GMM and the spectral transformation function as a robust method for speaker modification. The Gaussian mixture density is first described and the proposed function is given. Next, three types of the transformation function are presented and their properties are discussed. This is followed by the description of the training procedure which consists of two steps: 1) the parameter estimation of the GMM using the Expectation Maximization (EM), algorithm and 2) the parameters estimation of the three types of the transformation function using the GMM and a probability classification of the spectral envelopes. Finally the experimental evaluation of the proposed method is presented and it is compared with the VQ approach.

Chapter 7 summarizes the major results and conclusions of the thesis and suggests future directions for research based on this work.

In Appendix A it is shown that writing the harmonic part as a sum of sinusoidal terms then the system to solve in order to estimate the amplitudes and phase of the sinusoids (problem addressed in Chapter 3), can be split into two sub-systems with matrices which are dominated by their main diagonal allowing the use of iterative approaches like Gauss-Seidel or the more efficient SOR (Successive Over-Relaxation) method to solve them.

In Appendix B some sound examples are given to support our conclusions. A floppy disk (or a dat cassette) with these examples is available from: ENST-Paris, Signal Department, 46, Rue Barrault, 75013 Paris. The format of the sounds is: short 16 bits without header, binary. Questions regarding the sound examples may be directed to the author at:

ENST-Paris, Signal Department
46, RUE BARRAULT
75013 PARIS
FRANCE
(331) 45.81.77.27
email: styliano@sig.enst.fr

or
CHAPTER 1. INTRODUCTION

AT&T Labs-Research
2D-457
600 Mountain Ave
P.O. Box 636
Murray Hill, NJ 07974-0636
USA
(1)908-582-3144
email : styliano@research.att.com

Finally, in Appendix C an extended summary in French of the work presented in this thesis is provided.
Part I

SPEECH MODIFICATION
Chapter 2

Sinusoidal Models

2.1 Introduction

The purpose of this chapter is to provide a background review of speech synthesis and pitch and time-scale modifications based on sinusoidal models and their extension, sinusoidal plus noise models. The notation and nomenclature used throughout the first part of the thesis is also introduced. Beginning with a parametric model for the production of speech signals we define precisely pitch and time-scale modifications. Some of the most recently proposed representations of speech based on the parametric sinusoidal and sinusoidal + noise models are then discussed. Although it is not a parametric model, the Pitch-Synchronous Overlap-Add (PSOLA) [MC90] synthesis technique is briefly presented. The particular strengths and weaknesses of these models when applied to speech modification are also discussed.

This chapter is organized as follows. In section 2.2 a basic speech production model is introduced. This model, which sinusoidal models are based on, helps to define precisely pitch and time-scale modifications. This is followed in Section 2.3 by the problem of identifying sinusoidal parameters. Section 2.4 presents an overview of some of the most known sinusoidal models proposed recently. Sinusoidal plus noise models are then presented in Section 2.5. Finally, Section 2.6 briefly presents the TD-PSOLA approach and discusses some of the sources of the synthetic signal degradation obtained by TD-PSOLA when speech modification with large scale factors is applied.
CHAPTER 2. SINUSOIDAL MODELS

2.2 Speech production model

When designing time-scale and pitch-scale modification algorithms, it is often convenient to refer to a parametric model for the production of speech signals (even when this model is not used explicitly, see TD-PSOLA, for analysis/synthesis purposes). One approach to the problem of representation of speech signals is to use a speech production model in which speech is viewed as the result of passing a glottal excitation waveform through a time-varying linear filter that models the resonant characteristics of the vocal tract.

2.2.1 The linear prediction formulation

The LPC [MG76] approach is perhaps the most widely used where the time-varying filter excited by either quasi-periodic pulses (during voiced speech), or noise (during unvoiced speech). In an attempt to design high-quality speech coders at midband rates, generalizations of the binary excitation model have been developed. One such approach is what is known as the multipulse LPC [AR82] which uses more than one pitch pulse to model voiced speech and a possibly random set of pulses to model unvoiced speech. The amplitudes and locations of the pulses are chosen to minimize the difference between original and synthetic speech signals (using an appropriately weighted criterion to take into account the perceptual characteristics on the human ear). Code excited linear prediction (CELP) [SA85] is another representation which models the excitation as one of a number of random sequences or "codewords" superimposed on periodic pitch pulses.

2.2.2 The sinusoidal formulation

In the sinusoidal framework the excitation waveform is represented as a sum of sine waves. The motivation of this sine wave representation of the excitation is that the voiced excitation (periodic pulse train) can be represented by a Fourier series decomposition in which each harmonic component of the decomposition corresponds to a single sine wave [MQ86].

For voiced speech, the excitation waveform $e(t)$ is represented as a sum of harmonically related complex exponentials with unit amplitudes, zero initial phase, and a slowly varying fundamental frequency function $t \rightarrow 2\pi/P(t)$, where $P(t)$ is the local pitch-period (the function $t \rightarrow P(t)$ is referred to as the pitch contour). In mathematical terms, this is expressed as

$$e(t) = \sum_{k=1}^{K(t)} exp(j\Phi_k(t))$$

(2.1)
2.2. SPEECH PRODUCTION MODEL

where $\Phi_k(t)$ is the instantaneous excitation phase of the $k$-th harmonic and $K(t)$ is the number of harmonics which may vary in time.

In a "pure" harmonic coder, the instantaneous frequencies, $f_k(t)$, are related in a harmonic manner:

$$f_k(t) = kf_0(t)$$  \hspace{1cm} (2.2)

The time derivative of the instantaneous phase is referred to as the instantaneous frequency.

$$\dot{\Phi}_k(t) = 2\pi kf_0(t)$$  \hspace{1cm} (2.3)

The time-varying linear filter that models the resonant characteristics of the vocal tract approximates the combined effect of: (1). the transmission characteristics of the supra-glottal cavities (including the radiation at the mouth opening) and (2). the glottal pulse shape. Its time-varying transfer function can be written as

$$H(f; t) = G(f; t)exp[j\psi(f; t)]$$  \hspace{1cm} (2.4)

where $G(f; t)$ and $\psi(f; t)$ are respectively referred to as the time varying amplitude and phase of the system. If the time-varying impulse response of the vocal tract is $h(\tau; t)$, then the speech signal $s(t)$ can be viewed as the convolution of $h(\tau; t)$ with $e(t)$:

$$s(t) = \int_0^t h(t - \tau; t) e(\tau) d\tau$$  \hspace{1cm} (2.5)

Note that the amplitudes of the pitch-harmonics in the excitation signal have been assumed constant: $G(f; t)$ alone accounts for the magnitude of the speech signal’s spectrum. Similarly, the pitch-harmonics in the excitation signal have a null initial phase: the system phase $\psi(f; t)$ alone accounts for the phases of the signal’s pitch-harmonics.

Assuming that the glottal excitation parameters in (2.1) are constant over the duration of the impulse response of the vocal tract filter in effect at time $t$, then using (2.1) and (2.4) in (2.5) results in the speech model

$$s(t) = \sum_{k=1}^{K(t)} A_k(t) e^{j\theta_k(t)}$$  \hspace{1cm} (2.6)

where $f_k(t) = kf_0(t)$. The harmonic amplitudes $A_k(t)$ of the $k$-th harmonic is the system amplitude $G(f_k(t); t)$. The phase $\theta_k(t)$ of the $k$-th harmonic is the sum of the excitation phase $\Phi_k(t)$ and the system phase $\psi(f_k(t); t))$:

$$\theta_k(t) = \Phi_k(t) + \psi(f_k(t); t))$$  \hspace{1cm} (2.7)

$\theta(t)$ is often referred to as the instantaneous phase of the $k$-th harmonic.
2.2.3 Time-scale and pitch-scale definition

The definition of time-scale and pitch-scale modifications given in [ML95] is applied directly to the sinusoidal speech production model and now is briefly reviewed as it is the definition used in the next chapter. Firstly, the time/pitch-scale warping function is described and the effect of time/pitch scale modifications on the sinusoidal model parameters is presented.

2.2.3.1 Time-scale modification

The goal of time-scale modification is to change the apparent rate of articulation without affecting the perceptual quality of the original speech. This requires the pitch-contour to be stretched or compressed in time, and the formant structure changed at a slower or faster rate than the rate of the input speech, but otherwise not modified. Fig. 2.1 shows an example of time stretched where the pitch-contour is slowing down but it is not modified.

![Original pitch contour](image1)

![Time stretched pitch contour by 0.6](image2)

Figure 2.1: Original and time stretched pitch contour by 0.6

For an arbitrary time-scale modification, the time $t$ in the original signal (corresponding
2.2. SPEECH PRODUCTION MODEL

to the original articulation rate) is mapped to \( t' \) in the modified signal through the mapping \( t \rightarrow t' = D(t) \), where \( D(t) \) is referred to as the time-scale warping function which is defined as follow

\[
D(t) = \int_0^t \beta(\tau) \, d\tau
\]  

(2.8)

in which \( \beta(\tau) > 0 \) is the time-varying time-modification rate. The case \( \beta(\tau) > 1 \) corresponds to slowing down the rate of articulation by means of a time-scale expansion, while the case \( \beta(\tau) < 1 \) corresponds to speeding up the rate of articulation by means of time-scale compression. Note that for a fixed rate \( \beta(\tau) = \beta \) the time-scale warping function (2.8) is reduced to a linear relation \( D(t) = \beta t \) while for time-varying time-scale modification this is a non-linear function.

The time-scale warping function \( D(t) \) specifies that speech events which take place at a time \( t' \) in the new time scale will have occurred at a time \( t = D^{-1}(t') \) in the original time scale, where \( D^{-1}(\cdot) \) denotes the inverse mapping from the new time scale back to the original time scale.

With reference to the voiced speech production model introduced above, the speech parameters should be transformed the following way:

- the pitch contour \( t' \rightarrow P'(t') \) of the modified signal is time-scaled by the warping function \( D \):

\[
P'(t') = P(D^{-1}(t'))
\]

- the system amplitudes and phases are time-scaled:

\[
\begin{align*}
A_k'(t') &= G\left(D^{-1}(t'), f_k(D^{-1}(t'))\right) \\
\theta_k'(t') &= \Phi_k'(t') + \psi\left(D^{-1}(t'), f_k(D^{-1}(t'))\right)
\end{align*}
\]

- the instantaneous frequency of the \( k \)-th harmonic in the modified signal at time \( t' \) corresponds to the instantaneous frequency in the original signal at time \( D^{-1}(t') \)

\[
f_k'(t') = f_k(D^{-1}(t'))
\]

2.2.3.2 Pitch-scale modification

The goal of pitch-scale modification is to alter the fundamental frequency in order to compress or to expand the spacing between harmonic components in the spectrum while preserving the short-time spectral envelope (the locations and bandwidths of the formants) as well as the time evolution. In that case obviously the pitch contour is modified (see Fig.2.2).

The first step is to specify a pitch-modification factor \( \alpha(t) > 0 \) which is in general a time-varying factor (for an arbitrary pitch-scale transformation), with an associated change in the pitch-contour

\[
P'(t) = \frac{P(t)}{\alpha(t)}
\]
Fig. 2.2 shows an example of pitch modification by constant pitch-scale factor (0.6): the time evolution is preserved and the pitch contour is scaled by 0.6.

![Pitch Contours](image)

**Figure 2.2:** Original and pitch contour modified by 0.6

When $\alpha(t) > 1$ the local pitch frequency is increased by a factor $\alpha(t)$ (the pitch period is multiplied by a factor $1/\alpha(t)$ while when $\alpha(t) < 1$ the pitch frequency is lowered by a factor $\alpha(t)$ (the pitch period is changed accordingly). With reference to the voiced speech production model introduced above, the speech parameters should be transformed the following way:

- the pitch-contour is scaled by the time-varying factor $\alpha(t')$:
  \[
  P'(t') = \alpha(t') P(t')
  \]

- to preserve the shape of the short-time spectral envelope, the system amplitudes and phases are computed at the shifted pitch-harmonic frequencies by sampling in frequency the amplitude and phase envelope.
  \[
  A_k'(t') = G(t', a(t') f_k(t')) \\
  \theta_k'(t') = \Phi_k' + \psi(t', a(t') f_k(t'))
  \]
2.3. PARAMETERS ESTIMATION

- the derivation of the instantaneous excitation phase, $\Phi'_k(t')$, is equal to the scaled $k$-th harmonic frequency

$$\Phi'_k(t') = 2\pi k\alpha(t)f_0(t)$$

Although the modification system described above is for voiced speech segments, it can be applied also to unvoiced frames. Using the same modification scheme both for voiced and unvoiced frames avoids voiced/unvoiced decisions. However, applying the same method to unvoiced segments generates artifacts for relatively large modification factors [QM92]: a certain "unnatural" quality is present in the unvoiced sounds and for pitch change greater than 20%, some hoarseness is present in the synthesized signal. Time-scaled unvoiced speech segments acquire a "tonal" quality, a result of the phase coherence between successive synthesis short-time spectra [ML95]. To maintain the quality of the transformed signal also on the unvoiced and voiced fricatives segments, different modification schemes must be applied. This possibility is offered by the proposed hybrid sinusoidal models for speech (see below).

2.3 Parameters estimation

The general identification problem in which the speech signal is to be represented by multiple sine waves is a difficult one to solve analytically. One would wish to estimate the sinusoidal parameters by the optimization of a cost function between the original and synthetic speech signal as it is expressed in (2.6). This approach poses a high-dimension non-linear optimization problem. To simplify the problem, frequency estimation is usually decoupled from the estimation of amplitudes and phases and the stationary hypothesis is applied within a frame (we suppose short duration speech frames, about 10ms).

For the case of frequencies restricted to multiples of the fundamental frequency a pitch detection scheme is proposed together with the speech model. This approach leads to very good frequency estimates provided that the fundamental frequency is evaluated accurately. As the frequencies are known the amplitudes and phases can be evaluated by a squared error criterion leading to a set of linear equations [AS84]. This method can be extended also to unvoiced frames using a predefined number of sinusoids uniformly distributed from zero to half the sampling frequency.

In the case where the frequencies are not multiples of the fundamental frequency a peak-picking algorithm is usually proposed for the parameters estimation [MQ86]: the frequencies of the underlying sine waves correspond to the location of the peaks of the periodogram, and the estimates of the amplitudes and phases are obtained by evaluating the short-time Fourier transform (STFT) at the frequencies of the peaks. The spectral sampling algorithm is less
optimal than the time-domain approach described above. The STFT-based methods would require long frames (at least three pitch-periods) to achieve a sufficient separation of the spectral lines. This is a shortcoming of these methods which becomes important when it is necessary to model segments where the pitch-period, amplitudes and phases vary rapidly. In terms of computational effort both approaches have comparable complexities since the linear set of equations in the time-domain case can be solved efficiently (see the next Chapter).

2.4 Sinusoidal models: An overview

The phase vocoder [FG66] was perhaps the first attempt to represent the speech waveform by a set of narrowband functions. The assumption is that if a speech signal is passed through a set of fixed bandpass filters and then the outputs of the filter bank recombined, the signal is not substantially degraded. The frequency deviation of the sine wave from the center frequency of each band is estimated via the phase derivative. A new formulation was introduced by Portnoff [Por81] by representing each sine wave component by excitation and vocal tract contributions. The sine wave frequencies in the model were constrained to be harmonically related. Another refinement of the phase vocoder was developed by Malah [Mal79], who also assumed a harmonic model with known pitch and made the filter bank pitch-adaptive, thus ensuring roughly one sine wave per filter.

A different approach was taken by Hedelin [Hed81], who proposed a pitch-independent sine wave model for use in coding the baseband signal for speech compression. The amplitudes and phases are estimated using Kalman filtering techniques, and each sine wave phase is defined to be the integral of the associated instantaneous frequency. As in the phase vocoder, absolute phase information is lost. Another sine wave-based speech system has been developed by Almeida and Tribollet in [AT83][AT82] in order to achieve low bit-rate and good-quality speech coding. Almeida and Silva [AS84] proposed to synthesize the speech signal in the time domain as the superimposition of harmonics whose instantaneous frequency varies continuously along an interpolation curve within each frame, in order to track fast pitch variations. The amplitude evolution along the segment is given by a first degree polynomial while the phase evolution is given by a polynomial of third degree. The more integrated approaches described in [MQ86] and [MAT89] apply the same sinusoidal representation to both voiced and unvoiced regions, the only difference being the use of a higher number of sinusoids in the latter. The sinusoidal model proposed by McAulay and Quatieri [MQ86] uses also a first degree polynomial for the amplitudes. The idea of applying a third degree polynomial to interpolate the phase between frame boundaries was independently proposed by Almeida and Silva [AS84] for use in their harmonic sine wave synthesizer. While the first model proposed by McAulay and Quatieri
2.5 **SINUSOIDAL PLUS NOISE MODELS**

[MQ86] uses sine waves with non harmonically related sinusoids, for the development of a low-rate sine wave coder a harmonic model was used [MQ85][MQ91].

Finally, George and Smith [GS87][GS92] proposed also an overlap-add sinusoidal model for analysis, synthesis and modification of musical tones.

2.5 Sinusoidal plus Noise Models

Sinusoidal models are known to be an efficient, very high quality representation of voiced speech. However, when applied to unvoiced speech, they become very inefficient if high quality is to be maintained. In many of the models presented above the difficulty of sinusoids to deal with unvoiced fricatives and plosive sounds is apparent. Moreover for application as speech modification the sinusoidal models are not appropriate. Thus another approach to modelling unvoiced and voiced fricative speech segments in the context of the sine wave model is to generate noise explicitly via the linear filtering of white noise whenever unvoiced speech components are detected in different bands. This approach, developed by Griffin and Lim [GL88], is referred to as the "multiband excitation vocoder". In this model, the short-time spectrum of speech is modelled as the product of an excitation spectrum and a spectral envelope. The spectral envelope is a smoothed version of the speech spectrum and the excitation spectrum is represented by a fundamental frequency, a voiced/unvoiced decision for each harmonic of the fundamental, and the phase of each harmonic declared voiced. Due to the division of the spectrum into multiple frequency bands with a binary voiced/unvoiced parameter for each band, they have termed this model the Multiband Excitation Model (MBE). Marques and Almeida [MA88a] [MA88b] proposed to represent unvoiced segments using stochastic narrow-band basis functions instead of sinusoids while Abrantes et al. [AMT91] have developed an hybrid sinusoidal modelling of speech: the proposed model represents the speech signal as a sum of sinusoids and bandpass random signals. The frequencies of the sinusoids are chosen as multiples of the fundamental frequency at the frame boundaries and they are interpolated within the frame. In this model the speech signal considered to be the output of the vocal tract excited by two different types of excitation: a quasi-periodic excitation produced by the vibration of the vocal cords and a random excitation produced by the turbulence of the air flow in the vocal tract.

Smith and Serra [SS90], and Serra [Ser89], have developed a system for sound analysis/transformation/synthesis based on a deterministic (sum of sinusoids) and a residual or a stochastic part. This model has been developed initially for musical use. The sound representation is obtained by restricting the sinusoids to modelling only the deterministic part of the sound, leaving the rest of the spectral information in the residual component. Disregarding the phase
both on deterministic and residual signal leads to the deterministic plus stochastic model. This twofold decomposition of the sound has been found to be the key factor for high-quality modifications as pitch-scale and time-scale. Also for musical and voice applications, X.Rodet and P.Depalle proposed to decompose a sound in order to obtain high-quality synthesis and modifications [Dep91] [PRD88] [Rod89] [RDP87] [RDP88] [DRP90]. Recently, a new additive synthesis method based on spectral envelopes and inverse Fast Fourier Transform, FFT, called $\text{FFT}^{-1}$, has been proposed by the team of synthesis at IRCAM (X.Rodet, P.Depalle and G. García) [RDG95] [Rod92] [DR92] [FRD93]. A different approach of speech decomposition has been proposed by C. d'Alessandro [d'A90] which is based on an elementary waveform representation in order to modify the speech signal in the time-frequency domain. The use of elementary waveforms for musical synthesis has been also proposed by X.Rodet [Rod80]. Recently a novel method for decomposition of the speech signal into a deterministic and a stochastic part has been proposed in [dBd95] [BdD95]. The method aims at decomposing the excitation signal (linear prediction residual) rather than the original speech signal. The stochastic part (or aperiodic part [BdD95]) is estimated using an iterative algorithm; the deterministic part is obtained by subtracting the estimated stochastic part from the excitation signal. An analysis/synthesis/modification algorithm for representing the aperiodic component of the excitation source in speech signals is proposed in [Gd94] [Gd96]. Also a study on the significance of the deterministic and the stochastic components for the speech signal can be found in [dBd95a] [DdB95c] [dBdB95b]. Finally, the decomposition of the speech signal in deterministic and stochastic components has been also explored for voice analysis [Fuj68] [Kli87] [dK93].

2.6 TD-PSOLA

All the above methods are parametric and based on the sinusoidal production model. In this section a non-parametric method will be briefly reviewed: the Time Domain Pitch Synchronous OverLap Add method or TD-PSOLA [MC90]. This method relies heavily on the speech production model described in Section 2.2.2 but the parameters of this model are not estimated explicitly. The TD-PSOLA method is characterized by simplicity and low computational complexity allowing high-quality speech modifications. It is now widely adopted for text-to-speech synthesis based on concatenation of acoustic units.

The first step of the analysis process consists of decomposing the speech waveform into a stream of short-time analysis signals. These short-time signals are obtained by multiplying the signal waveform by a sequence of time-translated analysis windows. The analysis windows are located around the analysis time-instants which are set at a pitch-synchronous rate on the
voiced portions of speech and at a constant rate on the unvoiced portions. In the PSOLA context, the analysis time-instants are also referred to as the analysis pitch_marks. The length of the analysis window is proportional to the local pitch-period.

In the second step the synthesis pitch_marks (synthesis time-instants) are determined from the analysis pitch-marks according to the desired time-scale and pitch-scale modification factors. Along with the stream of synthesis pitch-marks, a mapping between the synthesis and the analysis pitch-marks is determined, specifying which short-time analysis signal should be selected for any given synthesis pitch-mark.

The final step consists of synthesizing the speech signal by combining the synthesis waveforms synchronized on the stream of synthesis pitch-marks. An example of time-scale and pitch-scale modifications using the TD-PSOLA approach is given in Fig.2.3 and Fig.2.4 respectively (from [ML95]).

![Figure 2.3: Pitch-scale modification with the TD-PSOLA method. Upper panel: original signal along with the analysis pitch- marks. Middle panel: three short-time synthetic signals. The mapping between these short-time signals and the analysis pitch- marks are indicated by the arrows. Lower panel: pitch-scale modified waveform along with the synthetic pitch-marks. The signal is a vowel /i/, uttered by a male speaker (pitch frequency around 100 Hz). The pitch-scale modification factor is equal to 0.8 (from [ML95])](image)

### 2.6.1 The drawbacks of TD-PSOLA

As we have seen, in order to achieve high-quality pitch and time-scale modification, TD-PSOLA eliminates/duplicates short-time waveforms extracted from the original speech signal
by windowing. Unfortunately, when this approach is applied on unvoiced fricatives, this process introduces a tonal noise because the repetition of segments of a "noise-like" signal produces an artificial long-time autocorrelation in the output signal, perceived as some sort of periodicity [ML95]. A simple solution to this problem consists of reversing the time-axis whenever the TD-PSOLA needs to duplicate unvoiced short-time signals [ML95]. This solution reduces the undesirable correlation in the output signal but the tonal quality does not completely disappear. Moreover, this solution can not be applied when the time-scale factor is greater than 2.

When voiced fricative (but also just voiced) frames are processed the same problem of tonality is present. This can be explained by the fact that the voiced fricatives exhibit a strong noise-like energy. Obviously the above solution can not be applied even for time-scale factors less than 2.

Also, because it is a non-parametric scheme, TD-PSOLA, does not allow complex modifications of the signal, such as increasing the degree of friction, or changing the amplitude and phase relationships between the pitch-harmonics.
2.7 Conclusion

The sinusoidal models have been found to be an efficient representation of voiced speech. For a flexible representation of the unvoiced sounds and for high-quality pitch and time-scale modification the hybrid models are more suitable. Experience in low-bit rate coding shows that the harmonic representation of voiced frames provides a simple and efficient representation of speech signals given that the fundamental frequency is evaluated accurately. The TD-PSOLA approach allows high-quality speech modification but only for moderate modification factors. The non-parametric scheme of this method is also one of its major shortcomings. However, the pitch-synchronous processing provided by TD-PSOLA allows the use of a flexible technique for pitch and time-scale modification.

From the above discussion the model that is suitable for high-quality speech synthesis/modification has the following characteristics: pitch-synchronous parametric harmonic plus noise (hybrid) model. One approach to this ideal model is discussed in the next chapter.
CHAPTER 2. SINUSOIDAL MODELS
Chapter 3

A Harmonic plus Noise Model, HNM

3.1 Introduction

In the previous chapter the sinusoidal model has been presented as an effective model for speech signals. However, there are many problems that must be addressed in applying the sinusoidal model to speech modification (time-scale and pitch-scale). The models addressing these problems are the hybrid models which decompose the speech signal into a deterministic and a stochastic part. This decomposition makes it possible to apply different modification methods to each part, increasing the quality of the modified synthetic signal.

This chapter presents HNM, a new analysis/modification/synthesis model based on a harmonic plus noise representation of the speech signal. The chapter is organized as follows. The development and description of the HNM is given first. This is followed in Section 3.3 by the problem of estimating the model parameters. Section 3.4 deals with the phase and amplitude envelope estimation needed for pitch-scale modification. Section 3.5 is concerned with the synthesis step when no speech modifications are applied. An alternative analysis and synthesis system for the noise part is discussed in Section 3.6. In Section 3.7 time-scale and pitch-scale speech modifications are addressed. Finally, in Section 3.8 an example of analysis, synthesis and modification of speech signal is given and the results of the application of the HNM to a larger database of speech signals, including male and female voices are discussed. The chapter ends with a summary of the system, a presentation of some sound examples, and conclusions.
3.2 Description of HNM

Fig. 3.1a shows the waveform of a voiced speech segment uttered by a female speaker (the sampling frequency is 16000Hz). The magnitude of the Fourier transform of the waveform is shown in Fig. 3.1b. The periodicity characteristic of the waveform is easily seen in the frequency domain; peaks at frequencies which could be supposed to be multiples of a fundamental frequency are observed. However, this frequency-domain periodicity does not cover all the frequency range from 0 to half of the sampling frequency. For this example, there is a periodicity up to 5000Hz and the frequency range from 5000Hz up to 8000Hz is dominated by a noise-like energy. The waveforms which correspond to the frequency range 0 – 5000Hz and 5000 – 8000Hz are plotted in Fig.3.1c and Fig.3.1d respectively.

![Waveform illustrations](image)

Figure 3.1: (a) Original speech signal (b) the magnitude spectrum of the speech signal (sampling frequency 16000Hz) (c) the time-domain signal corresponding to frequency range 0 - 5000Hz and (d) the time-domain signal corresponding to frequency range 5000Hz - 8000Hz

The important point to note is the special time-domain structure of the second waveform (which corresponds to a high-pass signal): it presents noise bursts which are synchronized with
3.2. DESCRIPTION OF HNM

the pitch period. For voiced frames, this noise-like signal is owed to turbulences of the glottal airflow principally observed on the voiced fricatives. The delimit of the two frequency bands is in general a time-varying parameter. For example, the voiced fricatives like /v/, /z/, have a noise part which covers more frequency range than in the above example. Also, very often the two frequency parts, periodic and noise-like part, are not completely separated. In order to obtain a high-quality speech synthesis we should take into account that the speech signal for the voiced frames, is composed of two distinct components: a component which reflects the periodicity of the speech signal and the other for the friction noise, turbulences of the glottal airflow, etc. The speech models which make this kind of decomposition of the speech signal are often referred to as hybrid or sinusoidal + noise models. If now the frequencies of the sinusoids are restricted to be multiples of a fundamental frequency the models are called harmonic + noise models (see below).

Harmonic + noise models have attracted a lot of research efforts in recent years, stimulated by the pioneering works by Griffin and Lim [GL88] and Abrantes et al. [AMT91]. The proposed HNM (Harmonic plus Noise Model) assumes the speech signal to be composed of a harmonic part $h(t)$ and a noise part $n(t)$. For a voiced speech signal, the spectrum is divided into two bands delimited by the so-called maximum voiced frequency $F_m(t)$ a time-varying parameter. The lower band of the spectrum (below the maximum voiced frequency) is represented by the harmonic part (low-pass signal), while the upper band by the noise part (high-pass signal). Thus, the harmonic part accounts for the periodic (voiced) structure of the speech signal and in the present discussion this part designates sums of harmonically related sinusoidal components with continuously time-varying amplitudes and phases.

\[
h(t) = \sum_{k=1}^{K(t)} a_k(t) \cos \phi_k(t) \tag{3.1}\]

where $a_k(t)$ and $\phi_k(t)$ are respectively the amplitude and phase at time $t$ of the $k$-th harmonic. Note that $K(t)$ represents the time-varying number of pitch-harmonics included in the harmonic part. These parameters are updated at specific time-instants denoted $t_i$. The time interval between two successive time-instants $t_i$ and $t_{i+1}$ is a frame. The derivative of the phase is defined as the instantaneous frequency which in HNM is the time-varying harmonic frequency $2\pi k f_0(t)$

\[
\dot{\phi}_k(t) = 2\pi k f_0(t) \tag{3.2}\]

The noise part accounts for the unvoiced frames and for the friction noise, the period-to-period fluctuations produced by the turbulences of the glottal airflow, etc. For voiced frames, the noise part is a high-pass signal. As we have stated above, this high-pass signal exhibits a specific frequency content as well as a time-domain structure in energy localization. One can remark that the noise bursts are synchronized with the glottal pulses in voiced fricatives.
The HNM follows this observation: the frequency contents of the noise part is described by a time-varying AR envelope; its time-domain structure is represented by a piecewise linear energy-envelope function.

The noise part \( n(t) \) is therefore supposed to have been obtained by filtering a white Gaussian noise \( u(t) \) by a time-varying, normalized all-pole filter \( h(t, \tau) \) and multiplying the result by an energy-envelope function \( e(t) \)

\[
n(t) = e(t) \left[ h(t, \tau) \ast u(t) \right]
\]

(3.3)

The filter \( h(t, \tau) \) is evaluated at each time-instant \( t_i \).

The final synthetic speech signal \( s(t) \) is supposed to be the superposition of the harmonic and the noise part: \( s(t) = h(t) + n(t) \).

### 3.3 Estimation of HNM parameters

In theory, the HNM parameters could be estimated by an analysis-by-synthesis technique i.e., by the optimization of a cost function between the original and the synthetic speech signal. However, this approach poses a high-dimension non-linear optimization problem and it is difficult to solve it analytically. To simplify the problem, the parameters for the noise and the harmonic part are estimated separately. The fundamental frequency estimation is isolated from the estimation of amplitudes and phases and the interdependence of the parameters in neighbouring frames is alleviated through the hypothesis of the quasi-stationary signal. Thus, the first step of the analysis process consists of estimating the fundamental frequency and the maximum voiced frequency for the voiced frames.

#### 3.3.1 Pitch, voicing and maximum voiced frequency estimation

All the harmonic models proposed for speech synthesis and coding propose also a pitch detection algorithm [TMR90]. However, estimating correctly this parameter, the pitch or otherwise the fundamental frequency, is not a simple task [RS78]. Many algorithms have been proposed in the literature [Hes83][RCRM76][SR79][Sen78][MQ85][GH87][Her88][DC89][MYC91], but we found that the design of a pitch and voicing detection algorithm adapted to our model was necessary in order to obtain a synthetic speech signal free of oddities.

The quality of the synthetic speech generated by HNM is essentially influenced by the quality and the faultlessness of the pitch measurement as well as the voicing detection (voiced/unvoiced decision). An algorithm for pitch, voicing and maximum voiced frequency estimation is
3.3. ESTIMATION OF HNM PARAMETERS

proposed [Sty96d]. The first step of the algorithm consists of an initial estimation of the pitch, based on an autocorrelation domain approach. The initial pitch estimation is used to detect the voicing, the maximum voicing frequency and finally to refine the initial pitch. All these detections are made in the frequency-domain.

3.3.1.1 Initial pitch estimation

The initial pitch estimation has to be based on a criterion which takes into account how close the synthesized speech will be to the original speech. A reasonable normalized criterion is:

\[
\mathcal{E} = \frac{\int_{-1/2}^{1/2} [ |S_w(f)| - |\hat{S}_w(f)| ]^2 df}{\int_{-1/2}^{1/2} |S_w(f)|^2 df} \tag{3.4}
\]

where \( S_w(f) \) is the Fourier transform of a windowed speech segment of the speech signal \( s(t) \) and \( \hat{S}_w(f) \) the synthetic speech spectrum generated by a fundamental frequency \( f_0 \).

To obtain the maximum sensitivity to regions of the spectrum containing pitch harmonics when large regions of the spectrum contain noise-like energy, the expected value of the error given by (3.4) should not vary with the pitch period for a spectrum consisting entirely of noise-like energy [GL88]. However, since the spectral envelope is sampled more densely for longer pitch periods, the expected error is smaller for longer pitch periods. To counter this, Griffin and Lim [GL88] proposed to multiply the error criterion in (3.4) by a pitch period dependent correction factor:

\[
\mathcal{E} = \frac{\int_{-1/2}^{1/2} [ |S_w(f)| - |\hat{S}_w(f)| ]^2 df}{\int_{-1/2}^{1/2} |S_w(f)|^2 df \left[ 1 - P \cdot \sum_{\ell = -\infty}^{\infty} w^4(t) \right]} \tag{3.5}
\]

By replacing integrals of continuous functions by summations of samples of these functions, Griffin and Lim [GL88] proposed an efficient method for computing (3.5) and it is the error function which is used for the initial pitch estimation:

\[
E(P) = \frac{\sum_{\ell = -\infty}^{\infty} s^2(t)w^2(t) - P \cdot \sum_{\ell = -\infty}^{\infty} r(l \cdot P)}{\left[ \sum_{\ell = -\infty}^{\infty} s^2(t)w^2(t) \right] \left[ 1 - P \cdot \sum_{\ell = -\infty}^{\infty} w^4(t) \right]} \tag{3.6}
\]

where \( s(t) \) is the speech signal, \( w(t) \) is the analysis window which is subject to the constraint

\[
\sum_{\ell = -\infty}^{\infty} w^2(t) = 1.0 \tag{3.7}
\]
The function $r(k)$ is defined as:

$$r(k) = \sum_{t=-\infty}^{\infty} s(t) w^2(t) s(t+k) w^2(t+k)$$

(3.8)

The above function is evaluated for periods $P$ in the set $[\frac{f_{\text{fmin}}}{f_{\text{fmax}}}, \ldots, \frac{f_{\text{s}}}{f_{\text{fmin}}}]$, where the $f_{\text{fmin}}$ and $f_{\text{fmax}}$ are the minimum and the maximum searched fundamental frequencies. Typical range values of minimum and maximum fundamental frequencies are 60 to 230 Hz for a male voice and 180 to 400 Hz for a female voice.

The error function in (3.6) has been normalized so that the minimum is near zero for a purely periodic signal and is near one for a noise signal. Fig. 3.2(a) shows the error function for a voiced frame and Fig. 3.2(b) for an unvoiced frame.

![Figure 3.2: (a) Error function for a voiced frame. (b) Error function for an unvoiced frame.](image)

Note, that minimizing (3.6) over $P$ is equivalent to maximizing

$$\Psi(P) = P \cdot \sum_{l=-\infty}^{\infty} r(l \cdot P)$$

(3.9)

This technique is similar to the autocorrelation method, but considers the peaks at multiples of the pitch period instead of only the peak at the pitch period. This suggests a computationally efficient method for maximizing $\Psi(P)$ over all integer pitch periods by computing the autocorrelation function using the fast Fourier transform (FFT) and then summing samples spaced by the pitch period. It should be noted that, in practice, the summations of (3.9) are finite due to the finite length of the window $w(t)$.

In order to eliminate gross pitch errors, a pitch tracking method proposed by [GL88] is used: the pitch track continues through the current frame by looking forward and backward two frames (corresponding to 20 ms). The minimum error paths from two frames in the past to the
current frame, and from two frames in the future to the current frame are found. The error along a path is determined by summing the errors at each pitch period through which the path passes. We then determine which of these paths has the smallest error, and the initial pitch period estimate is chosen as the pitch period in the current frame in which this smallest error path terminates. The permissible frame-to-frame pitch period deviation is set to 0.32 msec.

### 3.3.1.2 Voiced/Unvoiced decision

Using the initial fundamental frequency, we generate a synthetic signal, \( \hat{s}(t) \) as the sum of harmonically related sinusoids with amplitudes and phases estimated by the DFT algorithm. Denoted as \(|\hat{S}(f)|\) the synthetic spectrum and \(|S(f)|\) the original spectrum, the voiced/unvoiced decision is made by comparing the normalized error over the first four harmonics of the estimated fundamental frequency to a given threshold (-15 dB is typical).

\[
E = \frac{\int_{0.7f_0}^{4.3f_0} (|S(f)| - |\hat{S}(f)|)^2}{\int_{0.7f_0}^{4.3f_0} |S(f)|^2}
\]

(3.10)

where \( \hat{f}_0 \) is the initial fundamental frequency. If the error \( E \) is below the threshold, this frame is marked as voiced. Otherwise, as unvoiced. An example of voiced/unvoiced decision based on the above method, is shown in the Fig. 3.3. In Fig. 3.3(a) the speech signal is overlayed by voiced/unvoiced decisions which are taken manually. The Fig. 3.3(b) shows the voiced/unvoiced decisions produced by the algorithm. In the y-axis, the pitch period is presented in samples. On the unvoiced frames the pitch period is set to zero.

### 3.3.1.3 Maximum voiced frequency estimation

The first step in estimating the maximum voiced frequency, \( F_m \), is to search for the largest sine-wave amplitude (peak) in the frequency range \([\hat{f}_0/2, 3\hat{f}_0/2]\). For this peak, two amplitudes and a frequency location are determined. The location is defined simply as the frequency at which the actual peak occurs and it is denoted here by \( f_c \). The first amplitude is defined as the magnitude of the sample at the peak, while the second amplitude is defined as the nonnormalized sum of the amplitudes of all of the samples from the previous valley to the following valley (see Fig.3.4). This second amplitude is called cumulative amplitude and it is denoted by \( Amc \) in contrast to the simple amplitude : \( Am \). Using the cumulative amplitudes also resulted in a better separation between true peaks and spurious peaks than would a simple amplitude at the peak [Sen78]. The peaks which are in the frequency range \([\hat{f}_0 - \hat{f}_0/2, f_c + \hat{f}_0/2]\) are
Figure 3.3: (a) Speech signal overlayed by manually taken voiced/unvoiced decisions. (b) Voiced/unvoiced decisions produced by the algorithm.

Figure 3.4: Cumulative amplitude definition.
3.3. ESTIMATION OF HNM PARAMETERS

also searched and the two types of the amplitudes are calculated for each peak. Denote the frequencies of these neighbouring peaks and the mean value of their cumulative amplitudes, by \( f_i \) and \( Amc(f_i) \) respectively, the following "harmonic tests" are applied to the largest peak:

\[
\frac{Amc(f_c)}{Amc(f_i)} > 2
\]  
(3.11)

or

\[
Am(f_c) - \max\{Am(f_i)\} > 13\text{db}
\]  
(3.12)

then, if

\[
\frac{|f_c - Lf_0|}{Lf_0} < 20\%
\]  
(3.13)

then the frequency \( f_c \) is declared voiced; otherwise this frequency is declared unvoiced. \( L \) is the number of the nearest harmonic to the \( f_c \).

Having classified the frequency \( f_c \) as voiced or as unvoiced, then the interval \( [f_c + \frac{\hat{f}_c}{2}, f_c + 3\frac{\hat{f}_c}{2}] \) is searched for its largest peak and the same "harmonic tests" are applied. The process is continued throughout the speech band. In many cases the voiced regions of the spectrum are not clearly separated from the unvoiced ones. To counter this, a vector of binary decisions is formed making the convention that the frequencies declared as voiced will be noted as 1 and the unvoiced as 0. Filtering this vector by a three-point median smoothing filter, the two regions are separated. Then, the last detected one in the filtered vector provides the maximum voiced frequency.

An example of the application of the above algorithm is shown in the Fig. 3.5(a). The detected voiced frequencies are denoted by stars. Fig. 3.5(c) shows the estimated maximum voiced frequency from about 600 frames of the speech segment depicted in Fig. 3.5(b). In practice the maximum voiced frequency varies rapidly from frame to frame. To smooth these estimations a time domain median smoothing filter can be used but we should take into account the delays of the nonlinear smoother. In Fig. 3.5(c) a median filter of order 5 has been used. Fig. 3.6 shows the spectrogram of the speech segment in order to compare it with the detected maximum voiced frequency. The x-axis is in sec. and the y-axis in Hz.

3.3.1.4 Refining the initial pitch estimation

As the pitch values estimated by the pitch detection algorithm will be used in the harmonic modelling of the speech, an accurate pitch estimation is necessary. Using the initial pitch estimation, \( f_0 \), and the frequencies \( f_i \) classified as voiced from the previous step, the refined
Figure 3.5: (a) Maximum voicing frequency estimation for a voiced frame (b) a voiced speech segment and (c) Maximum voicing frequency estimation for the voiced speech segment in (b)

Figure 3.6: Spectrogram of the voiced speech signal in Fig.3.5(b). The x-axis is in msec and the y-axis in Hz.
3.3. ESTIMATION OF HNM PARAMETERS

pitch, \( \hat{f}_0 \), is defined as the value which minimizes the error:

\[
E(\hat{f}_0) = \sum_{i=1}^{L_n} |f_i - i \cdot \hat{f}_0|^2
\]

(3.14)

where \( L_n \) is the number of the voiced frequencies \( f_i \). In the Fig. 3.7(a) the original spectrum is overlayed with the synthetic spectrum for the initial pitch, while the Fig. 3.7(b) shows the original spectrum overlayed with the synthetic spectrum for the refined pitch.

![Graphs showing frequency vs. magnitude in dB for initial and refined pitch estimates.](image)

Figure 3.7: (a) Original et synthetic spectrum for the initial pitch estimation. (b) Original et synthetic spectrum for the refined pitch value.

### 3.3.1.5 The pitch detection algorithm

The algorithm that we use for pitch detection, voiced/unvoiced decisions and maximum voiced frequency estimation consists of the following steps (see Fig. 3.8):

1. Define the interval of the possible pitch values.

2. Window a speech segment with the analysis window. The duration of the window must be at least three times the maximum pitch period. A typical Blackman window is used.

3. Compute the error criterion of (3.6) versus the possible pitch periods. For a 16 kHz sampling rate this error is typically computed for all integer pitch periods from 40 to 260 samples.
4. Using (3.10) make the voiced/unvoiced decision in the frequency domain.

5. If the frame is declared voiced, separate the ‘voiced’ frequencies from the ‘unvoiced’ ones using the peak picking algorithm described above and the ‘harmonic test’ defined by (3.11) (3.12) (3.13). The last voiced frequency determines the maximum voiced frequency of the current frame. The size of the FFT used for this purpose is $N = 4096$.

6. The initial pitch as well as the frequencies classified as voiced are used to refine the pitch. The minimization of the error criterion of (3.14) provides the final value (refined) of the pitch.

7. Return to step 2 and continue the process throughout the speech signal. Typical value for frame rate is $10\text{msec}$.

![Figure 3.8: The pitch analysis algorithm.](image)

Using the stream of pitch values estimated by the above procedure, the analysis time-instants, $t^i_a$, are then set at a pitch synchronous rate on the voiced portions of speech, and at a fixed rate ($10\text{ms}$) on unvoiced segments

$$t^{i+1}_a = t^i_a + P(t^i_a)$$  \hspace{1cm} (3.15)

where $P(t^i_a)$ is the pitch period at time-instant $t^i_a$. 
3.3.2 Amplitude and phase estimation

It can be assumed that the amplitude of the harmonics and the pitch period are nearly constant around the time-instant \( t_a^i \)

\[
a_k(t) = a_k(t_a^i) \\
P(t) = P(t_a^i)
\]  

(3.16)

for small \(|t - t_a^i|\). Then, the instantaneous phase of \( k \)-th harmonic \( \phi_k(t) \) may be developed in the neighborhood of \( t_a^i \) as

\[
\phi_k(t) = \phi_k(t_a^i) + k \frac{2\pi f_0(t_a^i)}{P(t_a^i)} (t - t_a^i)
\]

(3.17)

for small \(|t - t_a^i|\). \( f_0(t_a^i) = 1/P(t_a^i) \) is the fundamental frequency in Hz, at the time-instant \( t_a^i \) (the pitch period, \( P(t_a^i) \), is expressed in sec). It is also supposed that the maximum voiced frequency \( F(t) \) is constant around the time-instant \( t_a^i \) and equal to \( F(t_a^i) \).

Working with real amplitudes, the estimation problem can be solved efficiently as it can be split into two sub-problems which are solved using an iterative approach (see Appendix A). However, the use of complex amplitudes is more attractive as we will show that the system to solve is characterized by a Toeplitz matrix. Hence, using the above so-called stationary conditions and complex amplitudes, the harmonic part can be written as a sum of exponential functions:

\[
\hat{h}(t) = \sum_{k=-L}^{L} A_k(t_a^i) e^{j2\pi k f_0(t_a^i)(t - t_a^i)}
\]

(3.18)

where \( L \) is the number of harmonics included in the harmonic part : \( L = F(t_a^i)/f_0(t_a^i) \) and \( A_k(t_a^i) \) is the complex amplitude of the \( k \)th harmonic with \( A_{-k} = A_k^* \) (where \( ^* \) denotes conjugate operation).

The harmonic complex amplitudes are evaluated by use of a weighted least-squares method aiming at minimizing the following criterion with respect to \( A_k \)

\[
\epsilon = \sum_{t=t_a^i-N}^{t_a^i+N} w^2(t) \left( s(t) - \hat{h}(t) \right)^2
\]

(3.19)

where \( s(t) \) is the original signal, \( \hat{h}(t) \) is the harmonic part defined as in (3.18), \( w(t) \) is the weighting function and \( N \) is the integer closest to local pitch period \( P(t_a^i) : N = \langle P(t_a^i) \rangle \), where \( \langle . \rangle \) represents the "round to nearest integer" operator. Note that the analysis frame is centered around the analysis instant \( t_a^i \) and the length of the analysis window is \( M = 2N + 1 \).

The weighting function \( w^2(t) \) is needed in order to provide a better match with the signal portion located at the center of the analysis frame.
At this point in the analysis, we find it convenient to switch to matrix notation\(^1\). In particular, we may rewrite the harmonic part simply as

\[ \hat{\mathbf{h}} = \mathbf{B} \mathbf{x} \]  

(3.20)

where \( \mathbf{B} \) is a \((2N + 1)\)-by-\((2L + 1)\) matrix defined by :

\[ \mathbf{B} = \begin{bmatrix} \mathbf{b}_{-L} & \mathbf{b}_{-L+1} & \mathbf{b}_{-L+2} & \cdots & \mathbf{b}_L \end{bmatrix} \]  

(3.21)

where \( \mathbf{b}_k \) is a \((2N + 1)\)-by-1 vector corresponding to \(k\)-th harmonic and it is defined by :

\[ \mathbf{b}_k = \begin{bmatrix} e^{j2\pi k f_0 (t_0^L - N)} & e^{j2\pi k f_0 (t_0^L - N + 1)} & e^{j2\pi k f_0 (t_0^L - N + 2)} & \cdots & e^{j2\pi k f_0 (t_0^L + N)} \end{bmatrix}^T \]  

(3.22)

where the symbol \(^T\) denotes the transpose operator.

The solution to least-squares problem is then given by the normal equations [LH74]

\[ (\mathbf{B}^T \mathbf{W}^T \mathbf{B}) \mathbf{x} = \mathbf{B}^T \mathbf{W}^T \mathbf{W} \mathbf{s} \]  

(3.23)

where \( \mathbf{W} \) is a \((2N + 1)\)-by-\((2N + 1)\) diagonal matrix with diagonal elements the weight vector \( \mathbf{w}^T = [w(-N) w(-N + 1) \cdots w(N)] \) which is a typical Hamming window. \( \mathbf{s} \) is a \((2N + 1)\)-by-1 vector which contains the original data \( s^T = [s(-N) s(-N + 1) s(-N + 2) \cdots s(N)] \) and \( \mathbf{x} \) is a \((2L + 1)\)-by-1 vector which contains the sought parameters and is defined by

\[ \mathbf{x} = [A_{-L} A_{-L+1} A_{-L+2} \cdots A_L]^T \]  

(3.24)

The (3.23) can be written as :

\[ \mathbf{R} \mathbf{x} = \mathbf{b} \]  

(3.25)

where \( \mathbf{R} = (\mathbf{B}^T \mathbf{W}^T \mathbf{W} \mathbf{B}) \) and \( \mathbf{b} = \mathbf{B}^T \mathbf{W}^T \mathbf{W} \mathbf{s} \).

Note that \( \mathbf{R} \) is a \((2L + 1)\)-by-\((2L + 1)\) matrix with elements \([r_{ik}]\) given by :

\[ r_{ik} = \sum_{t=t_{-N}^L}^{t_{L}+N} w^2(t) e^{j2\pi (i-L) f_0 t - j2\pi (k-L) f_0 t} \]  

(3.26)

with \( i = 1, \cdots, 2L + 1 \) and \( k = 1, \cdots, 2L + 1 \) and that \( \mathbf{b} \) is a \((2L + 1)\)-by-1 vector with \(k - th\) element given by :

\[ b_k = \sum_{t=t_{-N}^L}^{t_{L}+N} w^2(t) s(t) e^{-j2\pi (k-L) f_0 t} \]  

(3.27)

\(^1\)Note, that we will use \(t\) both for continuous and discrete-time domain, where the sampling frequency is assumed normalized to unity.
3.3. ESTIMATION OF HNM PARAMETERS

We show now that $\mathbf{R}$ is Toeplitz matrix. It suffices to show that $r_{i+p,k+p} = r_{i,k}$ for all $i, k, p$ (such that the indices are within the bounds). In fact,

$$
r_{i+p,k+p} = \sum_{t=t_L-N}^{t_L+N} w^2(t) e^{j2\pi (i-L-1)f_0 t + j2\pi p f_0 t} e^{-j2\pi (k-L-1)f_0 t} e^{-j2\pi p f_0 t} 
$$

(3.28a)

$$
= \sum_{t=t_L-N}^{t_L+N} w^2(t) e^{j2\pi (i-L-1)f_0 t - j2\pi (k-L-1)f_0 t} 
$$

(3.28b)

$$
= r_{i,k} 
$$

(3.28c)

Hence all the matrix $\mathbf{R}$ is only defined by the $2L + 1$ elements of its first column. Because $\mathbf{R}$ is a Toeplitz matrix, fast algorithms like the Levinson’s algorithm, can be used to solve the linear system of equation in (3.25).

Other properties of the matrix $\mathbf{R}$ is that it is Hermitian:

$$
\mathbf{R}^* = \mathbf{R}^T
$$

(3.29)

In fact,

$$
r_{i,k}^* = \sum_{t=t_L-N}^{t_L+N} w^2(t) e^{j2\pi (L-i+1)f_0 t - j2\pi (L-k+1)f_0 t} 
$$

(3.30a)

$$
= \sum_{t=t_L-N}^{t_L+N} w^2(t) e^{j2\pi (k-L-1)f_0 t - j2\pi (i-L-1)f_0 t} 
$$

(3.30b)

$$
= r_{i,k} 
$$

(3.30c)

Note also that $\mathbf{R}$ verifies $r_{L+1+i,L+1+k} = r_{L+1-k,L+1-i}$:

$$
r_{L+1+i,L+1+k} = \sum_{t=t_L-N}^{t_L+N} w^2(t) e^{j2\pi i f_0 t} e^{-j2\pi k f_0 t} 
$$

(3.31a)

$$
= \sum_{t=t_L-N}^{t_L+N} w^2(t) e^{-j2\pi k f_0 t} e^{j2\pi i f_0 t} 
$$

(3.31b)

$$
= r_{L+1-k,L+1-i} 
$$

(3.31c)

Note that if the interaction among the harmonics is insignificant that is $r_{ik} = 0$ for $i \neq k$, $\mathbf{R}$ becomes diagonal and the solution of (3.25) is straightforward; the complex amplitudes of the $k$-th exponential is given by

$$
A_k = \frac{\sum_{t=t_L-N}^{t_L+N} w^2(t) s(t) e^{-j2\pi k f_0 t}}{\sum_{t=t_L-N}^{t_L+N} w^2(t)} 
$$

(3.32)
which is equivalent to the amplitude estimated by the peak picking algorithm (which is based on FFT) proposed by [MQ86]. One important point to note is that the estimation of the parameters is done entirely in the time-domain. This allows the use of short time-frames (typically two pitch-periods). FFT-based methods would require longer frames (three to four pitch-periods). This feature of HNM model appears to be important to model segments where the pitch-period and/or the pitch-harmonics amplitudes and phases vary rapidly.

### 3.3.3 Estimation of the noise part parameters

The last step of the analysis consists of estimating the parameters of the noise part. In each analysis frame, the spectral density function of the original signal (full band) is modelled by a $p$th-order all-pole filter, $H(t_i, z)$, by use of a standard correlation based method [Kay88]. The correlation function is estimated from $40\text{msec}$ of signal located around the center of the analysis frame. Also, over the same duration, the variance of the original signal is estimated, representing the gain of the filter. For a signal sampled at $16k\text{Hz}$ the order of the AR model is set to 15.

In [LSM93b] [LSM93a] the time-domain behaviour of the noise part was imposed by an energy distribution function. The major shortcoming of this function appears when it becomes necessary to stretch it (in the case of pitch and time modification). To overcome this limitation a parametric envelope must be estimated in each frame. Unfortunately, determining such an envelope is not an easy task. However, we found that a triangular-like time-domain envelope gives, in general, satisfactory results (see Fig.3.9). A parametric method for the estimation of the time-envelope is one of the objects of our future work.

![Figure 3.9: The time-domain envelope used for the time-behaviour of the noise part. The $t_i$ and $t_{i+1}$ represent two successive analysis time-instants. Typical values for $l_1$ and $l_2$ are: $l_1 = 0.15(t_{i+1} - t_i)$ and $l_2 = 0.85(t_{i+1} - t_i)$.](image-url)
3.4 Phase and amplitude envelopes estimation

The phase and amplitude envelopes estimation can be viewed as an intermediary step between analysis and synthesis. These envelopes are used when for example pitch modification is done. In this case, the estimation of amplitudes and phases at frequencies not necessarily corresponding to a pitch-harmonic in the original signal is required.

For the phase envelope the estimated principal values of the phase must be unwrapped in the frequency and in the time domain. To this purpose, we have developed a new algorithm which guarantees phase continuity in the frequency-domain as well as in the time-domain by using the frequency derivative of the phase. For the amplitude envelope it is desirable to use a method that leads to an envelope which passes through the measured harmonic-wave amplitudes. Such a technique has already been developed in [CLM95]; it provides a continuous frequency-envelope when the value of this envelope (here the harmonics amplitudes) are specified only at discrete frequencies (in our case the pitch-harmonic frequencies). The method makes use of discrete cepstral coefficients and is based on a frequency-domain least-squares criterion combined with a regularization technique to increase the robustness of the estimation procedure. In this section only the application of the method to HNM will be given.

3.4.1 Phase envelope estimation

Since the analysis time-instants are obtained from a stream of pitch periods which are derived from a fractional pitch estimator, they are defined in the continuous time-domain. Hence, working in the discrete time-domain we have to correct the estimated phase values (estimated at the discrete time-domain analysis time-instants) to the fractional analysis time-instants. The estimation procedure of the phase envelope makes use of the corrected values of the phase. The phase envelope estimation is obtained as follows. At the first voiced frame of a voiced portion of signal, the phase is unwrapped in the frequency domain by adding integer multiples of $2\pi$ in order to keep the variation of the "phase slope" $d\phi_k$ as smooth as possible. The phase slope $d\phi_k$ being defined as

$$d\phi_k = \phi_{k+1} - \phi_k$$  \hspace{1cm} (3.33)

where $k$ denotes the $k$th harmonic. For example, using the first phase slope $d\phi_1 = \phi_2 - \phi_1$, $M$ integer multiples of $2\pi$ are added to $\phi_3$ so that $d\phi_2 = (\phi_3 + M \cdot 2\pi) - \phi_2 \approx d\phi_1$. Next the phase $\phi_4$ is unwrapped by using the slope $d\phi_3$, and so on, until the last harmonic.

In the next voiced frame, the phases are unwrapped by using the phase slopes from the previous frame (and not the phase slopes of the current one as was the case for the first frame of the voiced portion). This guarantees a phase continuity in the frequency-domain as well as in the
time-domain. The principle of the phase envelope estimation is shown in Fig. 3.10. For a new voiced segment the phase of the first frame is unwrapped using the current phase slopes and for the following frames the same procedure as above is applied.

![Diagram of phase envelope estimation](image)

**Figure 3.10:** Phase envelope estimation. a) for the first frame of a voiced segment and b) for the following frames of the same segment.

An example of phase unwrapping for 10 successive voiced frames is shown in Fig. 3.11. Note the phase continuity both in frequency and time domain.

![Phase unwrapping](image)

**Figure 3.11:** Example of phase envelope estimation and phase unwrapping.
3.4. PHASE AND AMPLITUDE ENVELOPES ESTIMATION

3.4.2 Amplitude envelope estimation

The amplitude envelope is computed from the estimated amplitudes of the harmonics by a discrete regularized cepstrum method using a warped frequency scale. Given a set of \( L \) values of harmonic amplitudes \( a_k \) measured at the normalized harmonic frequencies of the fundamental frequency \( f_0, f_k = k f_0 / f_s \) where \( f_s \) is the sampling frequency, the discrete cepstrum is obtained by minimizing the following squared error in the log-spectral domain

\[
\epsilon = \sum_{k=1}^{L} \| \log a_k - \log |S(f_k; \mathbf{c})| \|^2,
\]  
(3.34)

The spectrum magnitude \( |S(f; \mathbf{c})| \) is related to the real cepstrum coefficients by

\[
\log |S(f; \mathbf{c})| = c_0 + 2 \sum_{i=1}^{p} c_i \cos(2\pi f i).
\]  
(3.35)

where \( \mathbf{c} = [c_0, \ldots, c_p]^T \) are the real cepstrum coefficients, and \( p \) is the order of the cepstrum.

The least-squares solution is

\[
\mathbf{c} = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{a}
\]  
(3.36)

where \( \mathbf{a} = [\log(a_1) \ldots \log(a_L)]^T \) are the specified log-magnitude values, and matrix \( \mathbf{M} \) is defined as

\[
\mathbf{M} = \begin{bmatrix}
1 & 2 \cos(2\pi f_1) & 2 \cos(2\pi f_2) & \ldots & 2 \cos(2\pi f_p) \\
\vdots & \vdots & \vdots & & \vdots \\
1 & 2 \cos(2\pi f_L) & 2 \cos(2\pi f_L) & \ldots & 2 \cos(2\pi f_L)
\end{bmatrix}.
\]  
(3.37)

The problem with the solution given by (3.36) is that the matrix \( \mathbf{M}^T \mathbf{M} \) is in general ill-conditioned when \( p \) approaches \( L \) (singular when \( p \geq L \)). As the number \( L \) is a time variable parameter (\( L \) is the number of harmonics included into the harmonic part), the order \( p \) must be chosen as small as possible in order to avoid the ill-conditioned problem. However, very small orders (less than 9) lead to not a good fit at the specified frequency points. Regularization techniques are well-known for obtaining well-behaved solutions to over-parameterized estimation problems [TA77]. In [CLM95] a regularization technique applied to the discrete cepstrum estimation is presented and we briefly reviewed now.

The discrete cepstrum is now obtained by minimizing the following composite error criterion

\[
\epsilon_r = \sum_{k=1}^{L} \| \log a_k - \log |S(f_k; \mathbf{c})| \|^2 + \lambda \mathcal{R}[S(f; \mathbf{c})].
\]  
(3.38)

The first term is the error criterion as it is given in (3.34). The penalty functional \( \mathcal{R}[S(f; \mathbf{c})] \) assesses the physical plausibility of the spectral envelope. \( \lambda \) is the regularization parameter.
Larger values of \( \lambda \) produce ‘more regular’ estimators. The functional \( \mathcal{R}[S(f;c)] \) that we use is:

\[
\mathcal{R}[S(f;c)] = \int_{-1/2}^{1/2} \left[ \frac{d}{df} \log |S(f;c)| \right]^2 df.
\]  

(3.39)

This is a classical smoothness constraint that penalizes rapid variations in the spectral envelope [CLM95].

If we use the (3.35), we may write (3.39) into the form

\[
\mathcal{R}[S(f;c)] = \int_{-1/2}^{1/2} \left[ \frac{d}{df} \left( c_0 + 2 \sum_{i=1}^{p} c_i \cos(2\pi f) \right) \right]^2 df
\]  

(3.40)

\[
= 16\pi^2 \int_{-1/2}^{1/2} \left[ \sum_{i=1}^{n} c_i \sin(2\pi f i) \right]^2 df
\]  

\[
= 16\pi^2 \sum_{i=1}^{p} \sum_{j=1}^{p} c_i c_j \int_{-1/2}^{1/2} \sin(2\pi f i) \sin(2\pi f j) df
\]  

\[
= 8\pi^2 \sum_{i=1}^{p} \sum_{j=1}^{p} c_i c_j \int_{1/2}^{1/2} \left[ \cos(2\pi f (i - j)) - \cos(2\pi f (i + j)) \right] df
\]

The first part of the integral is not zero only for \( i = j \). Thus the regularization criterion can be written as

\[
\mathcal{R}[S(f;c)] = 8\pi^2 \sum_{i=1}^{p} c_i^2 i^2
\]  

(3.41)

or

\[
\mathcal{R}[S(f;c)] = c^T R c
\]  

(3.42)

where \( R \) is a \( p \times p \) diagonal matrix

\[
R = \begin{bmatrix}
0 & \ldots & \ldots \\
8\pi^2 1^2 & \ldots & \ldots \\
\ldots & \ldots & \ldots \\
8\pi^2 p^2 & \ldots & \ldots \\
\end{bmatrix}
\]  

(3.43)

Writing the error criterion in matricial form

\[
\epsilon_r = (a - Mc)^T(a - Mc) + \lambda c^T R c
\]  

(3.44)

the solution can be shown to be

\[
c = [M^T M + \lambda R]^{-1} M^T a
\]  

(3.45)
3.4. PHASE AND AMPLITUDE ENVELOPES ESTIMATION

Fig. 3.12 displays the spectral envelopes obtained using the direct and the regularization technique for a segment of male voice where the amplitudes of the harmonics (33 harmonics, maximum voiced frequency $F_M = 4000Hz$, sampling frequency 16000Hz) have been estimated as it is discussed in the previous section. The comparison of Fig. 3.12(a) and 3.12(b) proves that the direct estimation problem is ill-conditioned while the regularization technique suppresses the large oscillations and diverging behaviour in the spectral envelope [CLM95].

However, the modelling error at the low frequencies is not small (below 1000Hz). This leads to a poor quality of the synthetic signal. Since the ear is less sensitive in the spectral envelope at higher frequencies, nonlinear frequency scaling should be used in order to fit better the amplitudes of the lower harmonics. For this purpose we use the Bark scale which is a standard perceptually based warping function. The relationship between Hz-Bark is given by

Figure 3.12: Specified harmonic amplitudes (33 circles) and estimated spectral envelope for a cepstrum of order 20. (a) Direct estimation. (b) Regularized estimation with $\lambda = 2 \times 10^{-4}$ (from [CLM95]).
the analytical expression proposed by Zwicker [ZT80]

\[
\begin{align*}
    f(\text{Bark}) &= 13 \arctan \left( \frac{0.76 f(\text{Hz})}{1000} \right) \quad \text{if } f(\text{Hz}) \geq 605 \\
    f(\text{Bark}) &= 8.7 + 14.2 \log_{10} \left( \frac{f(\text{Hz})}{1000} \right) \quad \text{if } f(\text{Hz}) < 605
\end{align*}
\]

(3.46)

In order to work with the bark scale we use the normalized bark scale where the frequency range is from 0 to 1/2. The normalized bark scale can be obtained as

\[
f(\text{normalized bark}) = \frac{f(\text{Bark})}{2 \times 21.52}
\]

(3.47)

where 21.52 is the value of 8000Hz expressed in Bark (8000Hz is half of the sampling frequency). Fig. 3.13 shows the correspondence between linear and Bark normalized scale.

![Hz to Bark (normalized)](image)

Figure 3.13: Correspondence between the usual normalized frequency scale (in x-axis) and the normalized Bark scale (in y-axis) for sampling frequency 16 kHz.

Other warping functions could also be used such as mel scale or the warping function proposed by McAulay and Quatieri ([MQ91]). Basically the function proposed in [MQ91] is linear in the low-frequency region and exponential in the high-frequency region.

Using the normalized bark scale the spectral envelope is estimated as is mentioned above with the only difference that previously, the harmonic frequencies are transformed using (3.46) to the normalized Bark scale.

Fig. 3.14 shows that the estimation of the spectral envelope without and with normalized Bark scale. This figure shows that passing through the Bark scale allows us to obtain a spectral
3.4. PHASE AND AMPLITUDE ENVELOPES ESTIMATION

evelope with a negligible error for the first harmonics. In contrast, the amplitudes between 0.3 and 0.4 normalized Bark frequencies (which correspond to 2000 up to 4000 Hz) are very dense. Consequently, the modelling error in this frequency range is notable. However, since the ear is less sensitive to details in the spectral envelope at higher frequencies the spectral envelope estimated using Bark scale provides a synthetic speech signal which is almost indistinguishable from the synthetic one obtained using the estimated amplitudes from the analysis step. An

![Linear frequency scale and Bark frequency scale](image)

Figure 3.14: Estimation of the spectral envelope for a voiced frame with maximum voiced frequency $F_M = 4000\text{Hz}$ using a) linear frequency scale and b) normalized Bark scale

important point to note is that the above quality of the synthetic signal is achieved using a small order of cepstrum (typically 16) compared to other approaches [MQ91]. It is worth noting that the spectral envelope can be proven to be useful not only for pitch-scale modification but also for spectral transformation and low rate speech coding.
3.5 Synthesis without modifications

The synthesis is performed in a pitch-synchronous way as the analysis step. When no speech modification is done the synthesis time-instants, \( t_s^i \) (or synthesis pitch-marks) coincide with the analysis time-instants \( t_a^i \) (or analysis pitch-marks) : \( t_s^i = t_a^i \forall i \). Using the synthesis time-

![Diagram of synthesis and analysis time-instants](image)

Figure 3.15: For synthesis without speech modification the synthesis and the analysis time-instants coincide.

instants the harmonic part is obtained by the additive synthesis technique while the noise part by the overlap-add method. An alternative method for the analysis and synthesis of the noise part will be discussed in the next section. Given the harmonic and noise part, the synthetic signal is simply obtained by adding the two parts.

3.5.1 Synthesis of the harmonic part

Since a set of amplitudes and phases and a fundamental frequency are estimated for each voiced frame, it might seem reasonable to generate the harmonic part of the \( i \)-th frame using the equation

\[
\hat{h}(t) = \sum_{k=0}^{L(t_s)} a_k(t_s^i) \cos(\phi_k(t_s^i) + k \ 2\pi f_0(t_s^i)t)
\]  

(3.48)

where \( t = 0, 1, \ldots, N \) and \( N \) is the length of the synthesis frame which is also the integer closest to pitch period at the time-instant \( t_s^i \). However, this leads to discontinuities at the frame boundaries. To counter this, the estimated parameters are smoothly interpolated between successive frames. Due to the harmonic structure of the HNM there is no need for a component matching stage as in [MQ86]: the \( k \)-th harmonic in frame \( i \) corresponds to the \( k \)-th harmonic in frame \( i + 1 \).

Let \([a_k^i, \phi_k^i, f_0^i]\) and \([a_k^{i+1}, \phi_k^{i+1}, f_0^{i+1}]\) denote the sets of parameters at synthesis time-instant \( t_s^i \) and \( t_s^{i+1} \) for the \( k \)-th harmonic respectively. The amplitudes and phases are obtained by sampling the phase and amplitude (spectral) envelope at the harmonics of the fundamental frequencies \( f_0^i \) and \( f_0^{i+1} \). The instantaneous amplitude \( a_k(t) \) is obtained by linear interpolation
3.5. SYNTHESIS WITHOUT MODIFICATIONS

of the estimated amplitudes at the frame boundaries:

\[ a_k(t) = a_k^i + \frac{a_k^{i+1} - a_k^{i+1}}{t_s^{i+1} - t_s^i} t \quad \text{for} \quad t_s^i \leq t < t_s^{i+1} \]  \hspace{1cm} (3.49)

In contrast to the third polynomial used in [MQ91][AS84], we use a first degree polynomial for the phase. First, the phase at the time-instant \( t_s^{i+1} \) is predicted from the estimated phase at \( t_s^i \) by

\[ \hat{\phi}_k^{i+1} = \phi_k^i + k 2\pi f_{0uv}(t_s^{i+1} - t_s^i) \]  \hspace{1cm} (3.50)

where \( f_{0uv} \) is the mean value of the fundamental frequencies at the two successive time-instants \( t_s^i \) and \( t_s^{i+1} \)

\[ f_{0uv} = \frac{f_0^i + f_0^{i+1}}{2} \]  \hspace{1cm} (3.51)

Next, the phase \( \phi_k^{i+1} \) is augmented by the term \( 2\pi M_k \) (\( M_k \) is an integer) in order to approach the predicted value. Therefore, the value of \( M_k \) is given by

\[ M_k = \left\lfloor \frac{1}{2\pi} (\hat{\phi}_k^{i+1} - \phi_k^i) \right\rfloor \]  \hspace{1cm} (3.52)

Then, the instantaneous phase \( \phi_k(t) \) is obtained by a simple linear interpolation

\[ \phi_k(t) = \phi_k^i + \frac{\hat{\phi}_k^{i+1} + 2\pi M_k - \phi_k^i}{t_s^{i+1} - t_s^i} t, \quad t_s^i \leq t < t_s^{i+1} \]  \hspace{1cm} (3.53)

Since the above analysis is applied to the voiced frames, it is necessary to specify the interpolation procedure also for unvoiced to voiced and voiced to unvoiced passages. In the first case where the frame \( i \) is unvoiced and the frame \( i + 1 \) is voiced, the amplitudes of the frame \( i \) are set to zero i.e., \( a_k^i = 0 \) \( \forall k \), while we maintain the same fundamental frequency i.e., \( f_0^i = f_0^{i+1} \). The phases are given by

\[ \phi_k^i = \phi_k^{i+1} - k 2\pi f_0^{i+1} N \]  \hspace{1cm} (3.54)

where \( N = t_s^{i+1} - t_s^i \).

In the second case, the frame \( i + 1 \) is unvoiced and now for the amplitudes we have \( a_k^{i+1} = 0 \) \( \forall k \) and for the frequency \( f_0^{i+1} = f_0^i \). The phases are defined as

\[ \phi_k^{i+1} = \phi_k^i + k 2\pi f_0^i N \]  \hspace{1cm} (3.55)

For the interpolation procedure described above, we have assumed that the number of harmonics in each frame is the same. In general, the two sets have different numbers of harmonics as, both fundamental frequency and maximum voiced frequency are time-varying parameters. To counter this, extra harmonics of the fundamental frequency of the frame corresponding to the set with the fewer harmonics are considered, with zero amplitude. The phase are defined by (3.54) or (3.55) depending on the case (if the set with the fewer harmonics corresponds to the
first or to the second frame respectively). Thus, if \( L_i \) and \( L_{i+1} \) are the number of harmonics of the frame \( i \) and \( i+1 \), the number of harmonics that will be synthesized is \( L = \max \{ L_i, L_{i+1} \} \).

Having determined the instantaneous values of the harmonic amplitudes and phases the final harmonic part will be given by

\[
\hat{h}(t) = \sum_{k=0}^{L} a_k(t) \cos(\phi_k(t))
\]  

(3.56)

where \( a_k(t) \) is given by (3.49) and \( \phi_k(t) \) by (3.53).

The Fig.3.16 shows the block diagram of the synthesis of the harmonic part for the frame between the synthesis time-instants \( t^i_s \) and \( t^{i+1}_s \).

![Block diagram of the synthesis of the harmonic part](https://example.com/block_diagram.png)

**Figure 3.16:** Block diagram of the synthesis of the harmonic part for \( t \in [t^i_s, t^{i+1}_s] \).

### 3.5.2 Synthesis of the noise part

As discussed earlier, the maximum voiced frequency is in general a time-varying parameter. This means that to synthesize the noise part between two successive synthesis time-instants with different maximum voiced frequency, a time-varying high pass filtering should be used (as the lower frequencies have been synthesized by the harmonic part). However, this increases the computational cost of the synthesis (which should be as low as possible). In practice, the noise part is obtained by an Overlap-Add (OLA) synthesis technique.
3.6. AN ALTERNATIVE METHOD FOR THE NOISE PART

First, for each analysis time-instant $t_s^i$ a signal is synthesized by filtering a unit-variance white Gaussian noise through a normalized lattice filter. The length of the noise is equal to twice the local pitch period. In the context of autoregressive signal modelling the coefficients of the lattice filter (or $k$-parameters) are called reflection coefficients or PARCOR coefficients. Given the polynomial coefficients $\alpha$, of the $p$-th order all-pole filter estimated at the time $t_s^i$, the $k$-parameters are obtained by the recurrence formula given in [OS89][MG76]:

\[
k_i = a_i^{(i)}
\]
\[
\alpha^{(i-1)} = \frac{a_m^{(i)} - k_ia_0^{(i)}}{1 - k^2} \quad m = 1, 2, \cdots, (i - 1)
\]

(3.57)

for $i = p, p - 1, \cdots, 1$. The above recursion is initialized by $k_p = \alpha_p^{(p)}$. In practice this transformation is performed in the analysis step.

The output of the normalized lattice filter is multiplied by the estimated variance at the analysis time-instant $t_s^i$, $V(t_s^i)$. Hence, the synthesized noise signal at this point of the synthesis step has the same variance as the original signal. It should be noted that the synthesized noise signal covers the whole frequency range. However, when the frame to synthesize is voiced, the frequencies from 0 to the maximum voiced frequency have been synthesized by the harmonic part. Thus, and only if the frame is voiced, the synthesized noise signal is filtered by a high-pass filter with cutoff frequency equal to the local maximum voiced frequency. Obviously, if the frame is unvoiced the harmonic part is zero and the above step of the high-pass filtering is omitted. Denoted now by $n(t, t_s^i)$ the synthesized noise signal at $t_s^i$ and by $n(t, t_s^{i-1})$ the noise signal synthesized at the previous synthesis time-instant, $t_s^{i-1}$, the noise part for the current frame is obtained by overlapping and adding $n(t, t_s^i)$ with $n(t, t_s^{i-1})$. Finally, and also only if the frame to synthesize is voiced, the time-domain envelope is directly applied to the noise part.

Fig.3.17 shows the block diagram of the synthesis of the noise part for the frame between the synthesis time-instants $t_s^i$ and $t_s^{i+1}$.

### 3.6 An alternative method for the noise part

An alternative method for the analysis and synthesis of the noise part has also been developed. Using an analysis window of 20ms around each analysis time-instant, the harmonic amplitudes of a fixed fundamental frequency for all the analysis time-instants are estimated. This is done in the same manner as for the harmonic part, i.e.: solving the linear set of equations

\[
\begin{pmatrix} B^T & W^T & W \end{pmatrix} x = B^T W^T W s
\]

(3.58)
However, using a fixed fundamental frequency and a fixed analysis window, the solution can be efficiently obtained by computing at once and at the beginning of the analysis step the matrix:

\[
\left( B^T W^T W B \right)^{-1} B^T W^T W
\]  

(3.59)

The harmonic amplitudes for the noise part are then obtained by multiplying the above constant matrix with the signal vector, \( s \), which is defined by the 20ms samples around each analysis time-instant. In contrast to the harmonic part, for the noise part only the magnitude of the spectrum is of interest. So the phase information is rejected.

To avoid time-aliasing problems the fundamental frequency should be selected to be at least twice the time resolution (1/20ms = 50Hz). In practice we use 100Hz. This means that for each analysis time-instant we have 80 amplitudes for the noise part (sampling frequency 16kHz). In order to diminish the number of parameters needed to save for the noise part, the spectrum envelope procedure described above can be applied but this time the amplitudes of the noise part are the input parameters. The order of the discrete cepstrum in this case is set to 10. Hence, the number of parameters to save for the noise part is comparable with the order of the AR approach used before (see Section 3.3.3). Note that now the variance of the signal is not needed.

In the synthesis step the amplitudes for the noise part are obtained by sampling the spectrum envelope at the harmonic frequencies of 100Hz. Only the amplitudes of the frequencies which are within the frequency range from the local maximum voiced frequency to half of...
the sampling frequency are obtained. The corresponding phases are generated with a random number generator. Then the noise part is obtained by linearly interpolating the amplitudes and phases between two successive synthesis as was the case for the harmonic part. Obviously, if the frame to synthesize is voiced the time-domain envelope is applied on the noise part as previously. It is worth noting that the synthesis of the noise part carried out in this manner is faster than the technique used in the previous section: there is no need of high-pass filtering and only one pitch period is generated at each synthesis time-instant instead of the two periods needed for the OLA technique.

Fig. 3.18 shows the analysis and the synthesis step for the noise part, as it is described above. In Fig. 3.18, \( f_{0c} \) denotes the constant fundamental frequency for the noise part (for example, \( 100Hz \)) and \( R^{-1} \) the inverted matrix calculated in the beginning of the analysis step. The amplitude and phase index, \( k \), at the synthesis step is defined in the range \((k_m : k_a)\), where \( k_m f_{0c} > F_M(t^i_a) \) for the first frame \((t^i_a)\) and \( k_m f_{0c} > F_M(t^{i+1}_a) \) for the second frame \((t^{i+1}_a)\). Obviously, \( k_a f_{0c} < f_s / 2 \). Note that \( F_M \) denotes the maximum voiced frequency at the corresponding frame and \( f_s \) the sampling frequency.

![Block diagram](image)

Figure 3.18: Block diagram of the (a) analysis and (b) synthesis of the noise part using the alternative method
3.7 Prosodic modifications

The pitch-synchronous scheme of the HNM allows the use of a simple and flexible technique for time-scale and pitch-scale modifications inspired by [MC90]. The first step consists of finding out the synthesis time-instants \( t_s^i \) (or synthesis pitch-marks) according to the desired time-scale and pitch-scale modification factors. Using the new synthesis time instants the modified synthetic signal is obtained in the same manner as is discussed in the previous section (synthesis without modification). Note that in the case of pitch-scale modification the amplitudes and phases at the new harmonics have to be estimated. A full discussion of the modification procedure is given below.

3.7.1 Computation of synthesis time-instants

3.7.1.1 Time-scale modification

As was already mentioned in the previous chapter, the object of time-scale modification is to alter the apparent rate of articulation without affecting the spectral content: the pitch contour and the time-evolution of the formant structure should be time-scaled, but otherwise not modified [ML95].

From the stream of analysis time-instants \( t_a^i \) and the desired time-scale modification factor \( \beta(t) \), \( \beta(t) > 0 \) the synthesis time-instants \( t_s^i \) will be determined. The mapping \( t_a^i \rightarrow t_s(i) = D(t) \) is referred to as the time-scale warping function which is defined as the integral of \( \beta(t) \)

\[
D(t) = \int_0^t \beta(\tau) \, d\tau
\]  

(3.60)

Note that for a constant time-modification rate \( \beta(t) = \beta \), the time-scale warping function is linear: \( D(t) = \beta t \). The case \( \beta > 1 \) corresponds to slowing down the rate of articulation by means of a time-scale expansion, while the case \( \beta < 1 \) corresponds to speeding up the rate of articulation by means of a time-scale compression. Thus, speech events which take place at a time \( t_{or} \) in the original time scale will occur at a time \( t_{mo} = \beta t_{or} \) in the new (modified) time scale.

In the present discussion, to each analysis time-instant \( t_a^i \) is specified a time-scale modification factor \( b_a \). Thus, \( \beta(t) \) is piecewise constant function, i.e.: \( \beta(t) = \beta_a \), \( t_a^i \leq t < t_a^{i+1} \). It follows therefore that the time-scale warping function \( D(t) \) can be written as

\[
D(t) = D(t_a^i) + \beta_a(t - t_a^i), \quad t_a^i \leq t < t_a^{i+1}
\]  

(3.61)

with \( D(t_a^1) = 0 \). Having specified the time-scale warping function \( D(t) \), the next step consists of generating the stream of the synthesis time-instants \( t_s^i \), while preserving the pitch contour.
3.7. PROSODIC MODIFICATIONS

the pitch in the time-scaled signal at time \( t \) should be close to the pitch in the original signal at time \( D^{-1}(t) \). In other words, \( t \rightarrow P'(t) = P(D^{-1}(t)) \). We have now to find a stream of synthesis pitch marks (synthesis time-instants) \( t^i_s \), such that \( t^{i+1}_s = t^i_s + P'(t^i_v) \). To solve this problem, the use of a stream of virtual pitch-marks \( t^i_v \) in the original signal related to the synthesis pitch-marks by

\[
\begin{align*}
t^i_s &= D(t^i_v) \\
t^i_v &= D^{-1}(t^i_s)
\end{align*}
\]

is proposed in [ML95]. Assuming that \( t^i_s \) and \( t^i_v \) are known, we try to determine \( t^{i+1}_s \) (and \( t^{i+1}_v \)), such that \( t^{i+1}_s - t^i_s \) is approximately equal to the pitch in the original signal at time \( t^i_v \). This can be expressed as

\[
t^{i+1}_s - t^i_s = \frac{1}{t^{i+1}_v - t^i_v} \int_{t^i_v}^{t^{i+1}_v} P(t) \, dt \quad \text{with} \quad t^{i+1}_s = D(t^{i+1}_v)
\]

According to this equation, the synthesis pitch period \( t^{i+1}_s - t^i_s \) at time \( t^i_v \) is equal to the mean value of the pitch in the original signal calculated over the time-interval \( t^{i+1}_v - t^i_v \). Note that this time-interval \( t^{i+1}_v - t^i_v \) is mapped to \( t^{i+1}_s - t^i_s \) by the mapping function \( D(t) \).

The (3.63) is an integral equation but it is easily solved because \( D(t) \) and \( P(t) \) are piecewise linear functions. Fig. 3.19 illustrates an example of the synthesis pitch marks computation for time-scale modification by 1.5.

![Time-scale modification by 1.5](image)

Figure 3.19: Computation of the synthesis pitch-marks for time-scale modification. We have \( P'(D(t)) \approx P(t) \) and \( t^i_s = D(t^i_v) \).

3.7.1.2 Pitch-scale modification

The goal of the pitch-scale modification is to alter the fundamental frequency of a speaker while the spectral envelope of the speaker’s vocal tract system function is unchanged. As was
the case for the time-scale modification, the first step consists of computing the synthesis time
instants \( t_s^i \) from the stream of the analysis time instants \( t_a^i \) and the pitch-scale modification
factors \( a(t) \), with \( a(t) > 0 \). This is done in the following way:

The pitch-scale modification is applied only on voiced speech segments, where the analysis
time-instants are set at a pitch synchronous way, i.e. : \( t_s^{i+1} = t_s^i + P(t_a^i) \). As the synthesis
time-instants must also be positioned in a pitch synchronous way we have : \( t_s^{i+1} = t_s^i + P'(t_s^i) \),
where \( P'(t_s^i) \) is approximately equal to the scaled pitch period in the original signal around
the time \( t_s^i \) by \( 1/\alpha(t_s^i) \) :

\[
P'(t_s^i) = \frac{P(t_a^i)}{\alpha(t_a^i)}
\]  

(3.64)

Assuming now that the synthesis time-instant \( t_s^i \) is known, we try to determine the next
synthesis time-instant \( t_s^{i+1} \). This is easily done by taking that the synthesis pitch period
\( t_s^{i+1} - t_s^i \) is equal to the mean value of the scaled pitch period (by \( 1/\alpha(t_s^i) \)) in the original signal
calculated over the time-frame \( t_s^{i+1} - t_s^i \). This can be expressed as

\[
t_s^{i+1} - t_s^i = \frac{1}{t_s^{i+1} - t_s^i} \int_{t_s^i}^{t_s^{i+1}} \frac{P(t)}{\alpha(t)} dt
\]  

(3.65)

The above integral equation is easily solved as \( P(t) \) is piecewise linear function and \( \alpha(t) \) is a
piecewise constant function :

\[
\alpha(t) = \alpha(t_s^i) \quad \text{for} \quad t_s^i \leq t < t_s^{i+1}
\]  

(3.66)

![Figure 3.20: Calculation of the synthesis pitch-marks for pitch modifications.](image-url)
3.7. PROSODIC MODIFICATIONS

3.7.1.3 Joint time-scale pitch-scale modification

The modification procedure presented above can be further generalized to perform simultaneous time-scale and pitch-scale modification. In fact, specifying at each analysis time-instant a time-scale and pitch-scale factor and combining the equations (3.63) and (3.65), the synthesis time-instants can be obtained by solving the following integral equation

\[ t_{s}^{i+1} - t_{s}^{i} = \frac{1}{t_{u}^{i+1} - t_{u}^{i}} \int_{t_{u}^{i}}^{t_{u}^{i+1}} \frac{P(t)}{\alpha(t)} \, dt \]  \hspace{1cm} (3.67)

3.7.1.4 Mapping the synthesis time-instants to the analysis ones

The synthesis time-instants determined by the above procedure are not, in general, univocally associated with the analysis time-instants. It is clear from the Fig.3.19 that a virtual pitch-mark \( t_{v}^{i} \), does not necessarily correspond to an analysis time-instant. A similar observation can be made in the case of pitch-scale modification (Fig.3.20), where now the virtual pitch-marks are the synthesis time-instants. A simple solution consists of replacing the virtual pitch-marks by the nearest analysis time-instant. However, in adapting the above solution another problem arises many times: two or more successive frames could be the same and then the parameters of the sinusoids (amplitudes and phases) will not vary within the synthetic frame (or more synthetic frames). This does not lead to not a high-quality modified synthetic signal. It follows therefore that the repeated analysis time-instants should be eliminated. As an example, in Fig.3.21 the repeated analysis time instants in Fig.3.20 are eliminated.

Elimination of repeated analysis time–instants

![Diagram showing the elimination of repeated analysis time-instants](image)

Figure 3.21: Elimination of repeated analysis time-instants.
3.7.2 Computation of amplitudes and phases at the modified harmonic frequencies

Given the synthesis time-instants (which correspond to analysis time-instants) the next step for prosodic modification is the computation of the harmonic amplitudes and phases at the shifted (or original when only time-scale modification is done) harmonic frequencies. This is done by sampling the phase and amplitude envelopes at the corresponding harmonic frequencies. In the case of pitch modification, the phase and amplitude envelope should be sampled at the shifted harmonic frequencies $2\pi k f_0(t_a^i) = 2\pi k \alpha(t_a^i) f_0(t_a^i)$. Note that the above envelopes are sampled up to the current maximum voiced frequency $F_M(t_a^i)$. Thus, the pitch-modified harmonic part, $\hat{h}(t)$, occupies the same frequency range (0 up to $F_M(t_a^i)$ Hz) with the non pitch-scaled harmonic part of the same frame, $h(t)$, but the number of harmonics included in each part is different; when $\alpha(t_a^i) > 1$, $\hat{h}(t)$ includes fewer harmonics than $h(t)$, and when $\alpha(t_a^i) < 1$, $\hat{h}(t)$ includes more harmonics than $h(t)$. This means that the initial energy of the harmonic part will be changed. Therefore, the amplitudes of the shifted harmonics are normalized in such a way that the final energy of the pitch-modified harmonic part is equal to the energy of the unmodified one. An example of a pitch scaling by a factor of 1.5 is illustrated in Fig. 3.22, where it can be seen that the harmonic line spacing has increased while the spectral shape is maintained.

![Figure 3.22](image-url)

Figure 3.22: Pitch modification in the frequency domain. (a) Original speech spectrum and (b) pitch modified spectral magnitude by a factor of 1.5
3.7. PROSODIC MODIFICATIONS

3.7.2.1 Synthesis using the modified sinusoids parameters

Using the new synthesis time-instants and the modified set of parameters (amplitudes and phases) corresponding to each time-instant, the modified harmonic and noise part are obtained in the same manner as that discussed for the synthesis in the absence of time-scale and pitch-scale modification.

Note that the parametric time-domain envelope can be easily stretched, thus preserving on voiced frames the time-behaviour of the scaled-noise part. For example, for time-scaling by a factor of 2, two pitch periods are generated in the place of one. In terms of computational efforts the parametric time domain envelope is very attractive as it offers a simple way to scale it. However, this envelope has been determined regardless of the specific features of each speaker. As was mentioned above, the estimation of a parametric speaker-dependent time-domain envelope has to be explored. This remains as one of our future works.

3.7.2.2 Intensity modification

Speaking usually about prosodic modification, three parameters are considered: the pitch contour, the time evolution and the intensity. In the above analysis, we focused only on the modification of the two first parameters. The modification of the third parameter can be easily performed by associating two intensity-scale factors at each synthesis time-instant; the first one, \( c_h(t_i^1) \), for the harmonic part, and the second one, \( c_n(t_i^1) \), for the noise part. For the case of the harmonic part the harmonic amplitudes are multiplied by the square root of the current harmonic intensity-scale factor

\[
a'_k(t_i^1) = \sqrt{c_h(t_i^1)} a_k(t_i^1) \quad \text{for} \quad k = 1, \ldots, L
\]  

(3.68)

where \( L \) is the number of harmonics included in the harmonic part. These values are then linearly interpolated to generate the harmonic part.

For the noise part, the variance \( v(t_i^1) \), which represents the gain of the filter, is the parameter which is multiplied by the square root of the current noise intensity-scale factor \( c_n(t_i^1) \)

\[
v'(t_i^1) = \sqrt{c_n(t_i^1)} v(t_i^1)
\]  

(3.69)

3.7.2.3 Other types of modification

The HNM, as a full parametric model, allows for the application of a variety of other modifications than those presented above. For example, it is easy to modify the voiced/unvoiced quality of the speech signal by controlling the maximum voiced frequency. Another example is
frequency scaling. Frequency scaling compresses or expands the frequency contents of a speech signal such that both the spectral envelope and the spacing between harmonic components in the spectrum (and, thus, the pitch contour) are scaled. Note that this type of modification differs from the pitch modification discussed above in that the latter scales the pitch contour but preserves the short-time spectral envelope (see Fig.3.22). Frequency scaling can be viewed as a more general case of frequency axis modification. For large pitch-scale modification factors sometimes required in text-to-speech synthesis, it would be desirable to implement some sort of interaction between the modified harmonic structure and the spectral envelope. This can be achieved using frequency scaling. In contrast, as the pitch modification does not account for changes in the vocal tract spectral characteristics which may take place during human pitch modification [Sen82], only moderate modification factors can be used (from 0.6 to 2.0) in this case. Frequency scaling can be used for impaired hearing where it may be required to scale the spectrum nonlinearly in frequency. Another application is to speech conversion (subject of the second part of this dissertation) using methods like Dynamic Frequency Warping [VMT92].

3.7.2.4 The synthesis scheme for prosodic modifications

The main steps for synthesis with prosodic modification are given below

1. Using the analysis time-instants $t_a^i$, time-scale factors $\beta(t_a^i)$ and the pitch-scale factors $\alpha(t_a^i)$, compute the synthesis-time instants $t_s^i$. The equations used for this purpose are (3.63), (3.65) and (3.67) for time-scale, pitch-scale and joint time and pitch-scale modification respectively. Associate synthesis time-instants with analysis time-instants by replacing the virtual time-instants by the closest analysis time-instants. Eliminate repeated analysis time-instants.

2. Sample the phase and spectral envelope at the modified (when pitch-scale modification is done) or not harmonic frequencies of the current fundamental frequency, to obtain the phases and amplitudes at the corresponding harmonics.

3. Interpolate linearly the amplitudes and phases obtained form the previous step across successive frame boundaries as was described in Section 3.5.1 to generate the harmonic part. The frame boundaries are specified by the synthesis time-instants obtained by the first step.

4. Make synthesis of the noise part using the synthesis time-instants obtained by the first step and the overlap-add method.
5. Obtain the modified synthetic speech signal by adding the harmonic and the noise part Fig.3.23 shows the block diagram of the synthesis system of the HNM when prosodic modifications are applied.

![Block diagram of synthesis step using time-scale and pitch-scale modification](image)

Figure 3.23: Block diagram of synthesis step using time-scale and pitch-scale modification

### 3.8 Example and discussion of the HNM application on speech

In this section the results from the application of the HNM on speech is addressed. First, we give an example of speech analysis-modification-synthesis using HNM. Therefore, a step by step application of HNM will be presented. Next, the application of HNM on a speech database including male and female voices is discussed.

The word "wazi waza" pronounced by a male speaker was recorded at a sampling rate of $16k\,Hz$. This is a typical example of a voiced speech segment with voiced fricative sounds ('z'). The fundamental frequency and the maximum voiced frequency are estimated using the procedure discussed in the Section 3.3.1. Fig.3.24 shows the original speech signal the estimated fundamental frequency in $Hz$ and the maximum voiced frequency also presented in $Hz$. Note, that this step of analysis is performed in a constant frame rate of $10\,ms$. Using the stream of the estimated fundamental frequencies, the analysis time-instants $t_a^i$ are set at a
pitch synchronous rate on the voiced portions of speech, and at a fixed rate (10 ms) on unvoiced segments. Next, the amplitudes and phases of the harmonics included in the harmonic part are estimated (see Section 3.3.2). The number of samples used for estimating the parameters of a given frame is equal to twice the frame pitch period. For the noise part the spectral density function of the original signal (full band) is estimated by an all-pole filter of order 15 using Levinson’s algorithm. As this method makes use of the autocorrelation function to estimate the coefficients of the filter, a longer window (than that for the harmonic part) has to be used. The length of this window is fixed and equal to 40 ms. Over the same duration the variance of the original signal is estimated. In order to allow different possibilities for the synthesis of the noise part the coefficients of the all-pole filter are transformed to the k-parameters which are called reflection coefficients or PARCOR coefficients [MG76] [Mak75] [RS78]. Having estimated the parameters of the HNM, the next step consists of estimating the phase and spectral envelope (Section 3.4). At the end of the analysis step, a set of parameters including the local fundamental frequency, the local maximum voiced frequency and the local phase and spectral envelopes, is associated at each analysis time-instant.

When there is no speech modification, the synthesis time-instants coincide with the analysis time-instants. Using the stream of the parameters estimated at each analysis time-instant the
3.8. EXAMPLE AND DISCUSSION OF THE HNM APPLICATION ON SPEECH

The harmonic part is synthesized by the additive method described in Section 3.5.1 while the noise part using the overlap-add technique described in Section 3.5.1. Fig. 3.25 shows a part of the harmonic and the noise signals synthesized by HNM. The synthetic signal is obtained by adding the harmonic and the noise part. For comparison, Fig. 3.26 presents the above part of the original and the corresponding synthetic signal.

![Harmonic and Noise Parts](image)

Figure 3.25: (a) Harmonic part and (b) noise part synthesized from HNM.

An important property of the HNM is its capability of performing time-varying time and pitch-scale modifications. The crucial point in order to modify the time-scale and the pitch, is the determination of the stream of the synthesis time-instants from the analysis time-instants and the desired time-scale and pitch-scale modification. First, a stream of time-scale and pitch-scale modifications factors are associated at each analysis time-instant. For time-scale modification, the stream of synthesis time-instants is determined using the (3.63) while for pitch-scale modification using the (3.65). The next step is the association between the analysis and the synthesis time-instants followed by the elimination of the repeated analysis time-instants. The final step, before the synthesis, consists of computing the amplitudes and the phases at the harmonics of the local fundamental frequency (which is modified when pitch-scale modification is applied). The synthesis of the time-scaled or pitch-scaled synthetic signal is then performed in an identical manner as without modification.

Fig. 3.27 shows an example of time varying time-scale factors (top of the figure), and the resulted pitch-contour of the time-scaled synthetic signal (bottom). The important point to note is that the pitch-contour is time-scaled but otherwise not modified. In Fig. 3.28(a) an example of time-varying pitch-scale modification factors is depicted and in (b) the resulting modified
CHAPTER 3. A HARMONIC PLUS NOISE MODEL, HNM

Figure 3.26: (a) Original speech signal and (b) the corresponding synthetic signal.

pitch contour (solid line). For comparison, in the same figure the original pitch contour has also been depicted (dashed line). In Fig.3.29 the wideband spectrograms of the original speech signal and the time-scaled synthetic one are depicted (frame size : 4msec., window : Kaiser, hop size : 2msec.). The time-scale factors shown in Fig.3.27 are used for the generation of the time-scaled synthetic signal. Note that as the maximum voiced frequency does not exceed 5000Hz, the frequency axis is limited up to 5000Hz in order to observe better the formant structure. The important point to note is that the speech signal is time-scaled while the local pitch period in maintained and the formant structure is time-scaled but otherwise is respected.

Fig.3.30 shows narrowband spectrograms of original and pitch-scaled synthetic signal (frame size : 32msec., window : Kaiser, hop size : 4msec.). Note that the formant structure is preserved while the distance between the harmonics is increased. The pitch synchronous scheme of the HNM allows a simultaneous pitch and time-scale modification without additional complexity. In this case the synthesis time-instants are determined by (3.67). Once the synthesis time-instants are determined, the steps of the association of the analysis time-instants with the synthesis ones, the elimination of the repeated analysis time-instants and finally the computation of the amplitudes and phases from the spectral and phase envelope respectively, are applied. Fig.3.31 presents a narrowband spectrogram of simultaneously pitch and time-scaled synthetic signal using the previous time-varying pitch and time scale factors. The HNM has been tested on a speech database including 5 male and 5 female speech voices in French. The utterances are selected to comprise a high variety of situations occurring frequently in spoken French and also enough fricative voiced and fricative unvoiced, nasal and plosive sounds. The
3.8. EXAMPLE AND DISCUSSION OF THE HNM APPLICATION ON SPEECH

Figure 3.27: (a) Time-varying time-scale modification factors and (b) original pitch contour (dashed line) overlayed with the time-scaled pitch contour (solid line).

Figure 3.28: (a) Time-varying pitch-scale modification factors and (b) original pitch contour (dashed line) overlayed with the pitch-scaled pitch contour (solid line).
Figure 3.29: Wideband spectrograms of (a) the original speech signal and (b) the time-varying time-scaled synthetic one (the spectrogram has been broken into two overlapped parts). The x-axis is in sec and the y-axis in Hz.
Figure 3.30: Narrowband spectrograms of (a) the original speech signal and (b) the time-varying time-scaled synthetic one. The x-axis is in sec and the y-axis in Hz.
Figure 3.31: Narrowband spectrogram of the simultaneously time-scaled and pitch-scaled synthetic signal. The x-axis is in sec and the y-axis in Hz.
3.8. EXAMPLE AND DISCUSSION OF THE HNM APPLICATION ON SPEECH

The majority of frames were voiced in order to test the combination of the harmonic and noise part. The range of the pitch values for the male subjects was [65 – 220 Hz] and for the female [280 – 350 Hz]. The utterances were sampled at 16k Hz and in total was more than 100 seconds of speech.

The voicing detection (voiced/unvoiced decisions) was found to be one of the most important parameters, followed by the pitch estimation, for a synthesis without artefacts. It is more desirable, in the sinusoidal context, to detect an unvoiced frame as voiced, than the inverse. The voiced/unvoiced decisions taken by the voicing detection algorithm described in Section 3.3.1 was surprisingly free of errors. The pitch estimation was also "in general" free of errors; sometimes during voiced to unvoiced passage, an error of one octave would occur. The estimated maximum voiced frequency has been found to be rough with sharp frame-to-frame transitions. This problem has been eliminated by smoothing the estimations by a median filter (in a post-processing step) taking into account the delay introduced by the above non-linear filter. It is worth noting, that the pitch estimation algorithm was found to be robust even with additive background noise.

Informal listening tests have demonstrated the fidelity of the synthesis. The synthetic speech produced by the HNM is of very high quality, almost perceptually indistinguishable from the original. Not only does this validate the performance of the HNM, but it is also shows that if the phases could be efficiently coded, then the HNM could be also used for low-rate speech coding. The sounds which are not very well reproduced were the plosive sounds in the case where these sounds were detected as unvoiced.

Time-scale modifications are of high-quality, even for large modification factors, mainly because the harmonic and the noise part are processed by different techniques. The results are better than those obtained with TD-PSOLA, especially for voiced fricatives. For pitch-scale modifications the results are satisfactory when moderate modification factors are used (0.5 – 2.0). Although good quality has been achieved with the HNM applied for speech modifications, further work remains. Pitch modifications with larger factors may take account interactions between the modified harmonic structure and the spectral envelope.

The main computational load of the HNM is in the estimation of the amplitudes and phases of the harmonics and the generation of the harmonic components at the synthesis step. Other operations such as estimation of the parameters for the noise part, synthesis of the noise part, determination of the synthesis time-instants for speech modification, phase and amplitude envelope estimation add an insignificant amount to the overall computational load. For 1 second of speech the allocated time for pitch-voicing-maximum voiced frequency estimation, estimation of the parameters for harmonic and noise part and finally synthesis of these two parts, is about 2.8 sec in a Sparc-10 system. When speech modifications are desired
the allocated time is no more than 3 seconds per 1 second of speech. The overall system was implemented in the standard C language.

3.9 Conclusion

In this chapter an analysis/modification/synthesis system based on an harmonic plus noise, (HNM), representation of the speech signal has been presented. The HNM model we propose differs from previously reported models in many ways:

- The analysis and the synthesis are pitch-synchronous. This feature allows the use of a simple and flexible technique for time-scale and pitch-scale modifications inspired by PSOLA methods. The method makes time-varying, joint time-frequency speech modifications easy to implement and the results are free of artifacts often encountered in other modification procedures.

- Contrary to the multi-band excitation model proposed by Griffin [GL88], the spectrum is divided into only two bands: in the lower band, the signal is represented solely by harmonically related sinewaves with slowly varying amplitudes, frequencies and phase. The upper band contains the noise component.

- In voiced fricatives, the stochastic part exhibits a specific time-domain structure in terms of energy localization: the friction noise energy does not spread over the whole speech period; it is concentrated in the part of the pitch period where the glottis is open. The HNM model follows this observation: the frequency contents of the stochastic part is described by a time-varying AR model; its time-domain structure is imposed by a parametric time-domain envelope which modulates the noise component.

In HNM synthesis, the estimation of the parameters of the model is of crucial importance. It appears that the resulting quality is strongly influenced by the way the parameters are extracted from the data. Several improvements with respect to previously reported works are proposed:

- The estimation is done entirely in the time domain by maximizing an objective function derived from a statistical model of the speech signal (to date, most reported scheme use a mix of frequency-domain and time-domain methods; moreover, the objective function is not explicitly specified). An advantage of our method is that it allows the use of short time-frames (typically two pitch-periods), which appears to be important to model
segments where the pitch-period and/or the pitch-harmonics amplitudes and phases vary rapidly.

- For pitch modification, it is required to unwrap the phase in the frequency-domain. To this purpose, we have developed a new algorithm, which guarantees phase continuity in the frequency-domain as well as in the time-domain by using the frequency and time derivative of the phase. The harmonic amplitudes at the modified harmonic frequencies are computed from the spectral envelope estimated using a regularization technique for discrete cepstrum estimation described in [CM94].

The Harmonic plus Noise Model has been extensively tested on a large database of speech signals including male and female voices. In spite of its relative simplicity the HNM has given very satisfactory results. Time-scale modifications are of very high-quality, even for large modification factors. The results are better than those obtained with TD-PSOLA [MC90], especially when large modification factors are used and voiced fricatives are processed. For pitch-scale, the results are satisfactory when moderate modification factors are used. As already mentioned, for large pitch modification factors it would be desirable to implement some sort of interaction between the modified harmonic structure and the spectral envelope. Time-varying modifications are possible which can be used for speech-adaptive modification yielding a more natural modified synthetic signal.
CHAPTER 3. A HARMONIC PLUS NOISE MODEL, HNM
Chapter 4

New decomposition techniques of the speech signal

4.1 Introduction

In the previous chapters we have seen that the decomposition of the speech signal into two parts, one which accounts for the periodic structure of the speech signal and the other for the non-periodic structure, is the key factor for high-quality speech synthesis and modification. The HNM presented in the previous chapter decomposes the voiced frames of speech into two parts: the lower or "harmonic" band (from 0Hz up to maximum voiced frequency) which is supposed to include solely harmonically related sinusoids and the upper or noisy band which contains the noise part. However, the hypothesis that the speech signal has a perfect harmonic structure in the harmonic band is not (in general) true. Unfortunately, modelling this band as a sum of harmonically related sinusoids (which is a very simple and attractive approach), even if the fundamental frequency has been estimated accurately, introduces an unavoidable modelling error. Also, there is no reason to believe that low frequencies of speech signal do not contain noise information. It follows therefore that the decomposition performed by the HNM is a virtual decomposition of the speech signal, since the presence of this noise signal in the harmonic band is ignored.

In order to improve this problematic decomposition of the speech signal, this chapter introduces two original models for speech which, while preserving the attractive harmonic structure of the HNM (harmonically related sinusoids), achieve a better decomposition of speech than
HNM. These models are the Deterministic plus Stochastic Model, DSM [LSM93b][LSM93a], and the Harmonic plus Stochastic Model, HSM [Sty96a] [Sty96c] [Sty96b]. The DSM aims to decompose the original speech signal into a periodic part (deterministic part) and a stochastic part. The deterministic part, assumed to contain only harmonically related sinusoids with \textit{linearly varying complex amplitudes}, accounts for the periodic structure of the speech signal. The stochastic part is defined as the residual signal obtained by subtracting the sinusoidal components from the original speech signal. So, the stochastic part accounts for everything in the signal that is not described by the deterministic part. The HSM also decomposes the speech signal into a harmonic part and a stochastic part and it makes use of \textit{n}-th order polynomial with \textit{real} coefficients for the harmonic amplitudes and an \textit{l}-th order polynomial for the phase.

This chapter is organised as follows. In Section 4.2, the DSM model is described. In Section 4.3, we address the problem of estimating DSM parameters. Section 4.4 then presents the synthesis step of the DSM and an example of an application of the DSM on a speech signal is given in Section 4.5. An approach (different from the one presented in the previous Chapter) which is used to refine coarse pitch estimates using the DSM, is discussed next in Section 4.6. This is followed by a description of the HSM in Section 4.7. Section 4.8 describes the procedure for the estimation of the HSM parameters. Lastly, Section 4.9 is devoted to an extensive comparison of the three harmonic plus noise models which are presented in the first part of the thesis, the HNM, the DSM and the HSM, and then one of them is selected for the voice conversion system described in the second part of the dissertation.

4.2 Description of the DSM

This section presents an original model for speech that we have developed for speech synthesis and modification. In the literature [LSM93b] [LSM93a], this model was also called Harmonic + noise model; to distinguish its name from the HNM presented in the previous chapter, its name was changed to DSM.

The DSM assumes the speech signal to be composed of a deterministic and a stochastic component. The term "deterministic" designates sums of harmonically related sinusoidal components with \textit{piece-wise linearly varying complex amplitudes}. The stochastic part is defined as the \textit{residual signal} obtained by subtracting the sinusoidal components from the original speech signal. The stochastic part thus accounts for everything in the signal that is not described by the deterministic part. It includes the friction noise, the period-to-period fluctuations produced by the turbulences of the glottal airflow, and potential pitch-harmonic modelling errors. In DSM as in HNM, the stochastic part is represented both in time-domain and frequency domain:
the frequency contents of the stochastic part is described by a time-varying AR envelope; its time-domain structure is represented by a time-domain energy-envelope function.

The input signal $s(t)$ is therefore modelled as the sum of harmonic components, $\hat{d}(t)$, and a noise signal $q(t)$. The deterministic part, $\hat{d}(t)$, is given by

$$\hat{d}(t) = \sum_{k=-K(t)}^{K(t)} A_k(t) \exp(jk2\pi f_0(t)t)$$

where $A_k(t)$ is the complex harmonic amplitude at time $t$, $f_0(t)$ is the fundamental frequency and $q(t)$ is the stochastic component. These parameters are updated at specific time-instants denoted $t_a^i$. Within each time-frame $[t_a^i, t_a^{i+1}]$, the fundamental frequency $f_0(t) = f_0(t_a^i)$ is held constant; the complex amplitudes $A_k(t)$ are affine functions of time:

$$A_k(t) = a_k(t_a^i) + (t - t_a^i)b_k(t_a^i)$$

$a_k(t_a^i)$ is the original complex amplitude of the $k$th harmonic in time-frame $i$: it represents the original amplitude and phase of the harmonic at the time-instant $t_a^i$. Note that $b_k(t_a^i)$ is the complex slope of the harmonic; $b_k(t_a^i)$ reflects pseudo-linear variations of the harmonic amplitude and slight mis-adjustments of its instantaneous frequency. $K(t)$ represents the time-varying number of pitch- harmonics included in the deterministic part.

The stochastic part $q(t)$ is supposed to have been obtained by filtering a white Gaussian noise $u(t)$ by a time-varying, normalised all-pole filter $h(t, \tau)$ and multiplying the result by an energy-envelope function $e(t)$:

$$q(t) = e(t) [h(t, \tau) \ast u(t)]$$

where ‘$\ast$’ denotes the convolution operator. The filter $h(t, \tau)$ is evaluated at each time-instant $t_a^i$ and is interpolated on a sample by sample basis.

Based on this description of the DSM, it turns out that the DSM differs from the HNM both in its deterministic part (harmonic part for the HNM) and the stochastic part (noise part for the HNM). The harmonic part of the HNM can be obtained from the deterministic part of the DSM by setting the slope parameter to zero. The stochastic part, as it is defined by the subtraction of the deterministic part from the original speech signal is a full-band signal (from $0Hz$ to $f_s/2$ Hz, with $f_s$ the sampling frequency) in contrast to the noise part of the HNM which encloses only the frequencies from the maximum voiced frequency up to half of the sampling frequency. Also the time domain envelope for the stochastic part is determined in the analysis step and it varies from frame to frame. For the noise part of the HNM, this envelope has a parametric form, easily manipulated especially when speech modification is performed but it is imposed, that is to say, it is not determined in any optimal way and finally it does not vary from frame to frame or speaker to speaker.
4.3 Estimation of DSM parameters

The first step of the analysis process consists of estimating the fundamental frequency \( f_0(t) \), the maximum voiced frequency \( F_m(t) \), and separating the voiced from the unvoiced frames. This is achieved by use of the same pitch detector described for HNM (see Section 3.3.1). Because the DSM makes use of linear complex harmonic-amplitudes, it enables slight frequency deviations from the hypothesis of an "exact" harmonic structure. This means that the accuracy of the pitch estimate is not critical for high-quality synthesis (in contrast to HNM). Thus, the refining step of the pitch detector is not necessary and is eliminated. It will be shown that the DSM itself can refine the pitch estimation.

The second step of the analysis process consists, on voiced portions of speech, of estimating the values of the complex amplitudes and slopes of the pitch harmonics. This is done using a weighted least-squares method aiming at minimising the following criterion with respect to \( a_k(t_i^v) \) and \( b_k(t_i^v) \)

\[
\mathcal{E} = \sum_{t=t_i^v-N}^{t_i^v+N} w(t)(s(t) - \hat{d}(t))^2
\]

in which \( \hat{d}(t) \) is defined as in (4.1), \( w(t) \) is a weighting window and \( N \) is the integer closest to the local pitch period \( P(t_i^v) \). As was the case of the HNM, the above criterion has a quadratic form for the parameters and is solved by inverting an over-determined system of linear equations [LH74]. First, a \((2N + 1)\)-by-\((4L + 2)\) matrix \( \mathbf{B} \) is formed by concatenating the following matrices

\[
\mathbf{B} = [\mathbf{B}_1|\mathbf{B}_2|\mathbf{B}_3|\mathbf{B}_4]
\]

with

\[
(\mathbf{B}_1)_{il} = E^{(i-N)(l+1)} \quad (\mathbf{B}_2)_{il} = E^{-(i-N/2)(l+1)} \\
(\mathbf{B}_3)_{il} = (i - N/2)(\mathbf{B}_1)_{il} \quad (\mathbf{B}_4)_{il} = (i - N)(\mathbf{B}_2)_{il}
\]

where \( E = \exp(j2\pi f_0(t_i^v)) \) in which \( i \) ranges from 0 to \( 2N \) and \( l \) ranges from 0 to \( L = K(t_i^v) = F_m(t_i^v)/f_0(t_i^v) \). The \((4L + 2)\)-by-1 vector \( \mathbf{x} \) contains the sought parameters and is defined by

\[
\mathbf{x} = [a_1 \ a_2 \ldots a_L \ a_L^* \ a_{1}^{*} \ b_1 \ b_2 \ldots b_L \ a_1^* \ b_1^* \ b_2^* \ldots b_L^*]
\]

The solution to the least-squares problem is then

\[
\mathbf{x} = (\mathbf{B}^h \mathbf{W}_h \mathbf{W} \mathbf{B})^{-1} \mathbf{B}^h \mathbf{W}_h \mathbf{W} \mathbf{s}
\]

in which \( \mathbf{W} \) is a \((2N + 1)\)-by-\((2N + 1)\) diagonal matrix containing the weighting factors \( w(t) \), and \( \mathbf{s} \) is a vector formed from \( 2N + 1 \) samples of the original signal. Note that \(^h\) denotes the
4.3. ESTIMATION OF DSM PARAMETERS

transpose and conjugate. The deterministic part \( \hat{d}(t) \) is readily obtained by computing \( \hat{d} = Bx \) in which the vector \( \hat{d} \) contains the samples \( \hat{d}(t) \). The necessity of using a non-rectangular weighting window \( w(t) \) is discussed below.

The last step of the analysis consists of estimating the parameters relative to the noise component. In each analysis frame, the deterministic part is simply subtracted in the time-domain from the original signal yielding the residual signal \( q(t) \). The residual signal’s spectral density function is then modelled by fitting a \( p \)th order all-pole filter using a standard correlation-based method [Kay88].

The temporal energy distribution-function \( e(t) \) is calculated by squaring the residual signal, low-pass filtering it with a positive impulse-response filter and decimating by a factor \( R \). In practice, for a signal sampled at 16kHz, the AR model order is set to 15, and the decimating factor to 30.

4.3.1 choice of the weighting window \( w(t) \)

As mentioned above, the deterministic signal can be obtained by computing \( \hat{d} = Bx \). The difference between \( \hat{d}(t) \) and \( s(t) \), is the stochastic part, which is given by \( q = Bn - s \). When no weighting window is used, the energy of \( q(t) \) exhibits sharp peaks at \( t = t_a \). This can be explained by calculating the covariance matrix of the residual signal \( \hat{d}(t) - s(t) \) when no weighting window is used and when the input signal, \( s(t) \), is a white noise, \( u(t) \), with unit variance. Then, the covariance matrix is given by

\[
E(qq^h) = E\{(Bn - u)(Bn - u)^h\} \\
= E\{B(B^hB)^{-1}B^hu^hu^hB(B^hB)^{-1}B^h - uu^hB(B^hB)^{-1}B^h - B(B^hB)^{-1}B^hu^h + uu^h\} \\
= I - B(B^hB)^{-1}B^h
\]

where \( I = E(uu^T) \) is the identity matrix and \( E(.) \) denotes the expected value. The variance of the residual signal is the diagonal of the above matrix. Fig.4.1 shows the distribution of the variance of the residual signal for a synthetic harmonic sound with \( L = 20 \). The variance exhibits a strong peak at the centre of the analysis window.

Ideally, our least-squares analysis should generate a residual signal whose variance should be evenly distributed across the analysis frame, so as not to artificially modify the residual signal’s time-behaviour. This is obviously not the case when a rectangular window is used, and the residual signal (the stochastic part) exhibits strong peaks at time \( t_a \); a quite audible, artificial pitch is heard in the stochastic part. This problem can be solved in practice by using a non-rectangular weighting window. If a window is used that puts more weight on the frame-centre, then the variance of the error is decreased accordingly, and the distribution of the variance
becomes more uniform. Unfortunately, analytically determining the weighting window which makes the variance nearly constant is impossible in the general case. However, we found by trial and error that a Hamming weighting window gave satisfactory results, as can be seen in Fig. 4.1.

4.3.2 The conditioning problem

In theory, since $2K(t_s^i)$ complex parameters must be estimated, the analysis frame could be as short as $4K(t_s^i)$ samples. For example, in order to take into account sinusoids with frequencies up to $f_s/4$ (with $f_s$, the sampling frequency), one would choose $K(t_s^i) = P(t_s^i)/4$ where $P(t_s^i)$ is the pitch period in samples. This means that the length of the analysis frame could be as small as one pitch period. In practice however, it is best to take an analysis frame-length equal to twice the pitch period, to avoid problems related to the poor-conditioning of matrix $B^bB$.

4.4 Synthesis using the DSM

The synthesis is performed in a pitch-synchronous overlap-add technique. The deterministic and the stochastic components are synthesised separately and then added together to yield the synthetic signal.

The deterministic part is simply obtained by overlap-adding a stream of short-time signals $d_i(t)$ at the specific synthesis time-instants $t_s^i$ which are identical to the analysis time-instants $t_a^i$. The short-time signals are obtained from the harmonic parameters using the synthesis equation given by (4.1) and applying a Hamming window centred at $t_s^i$. The length of the synthesis window is twice the local pitch-period, and the overlap between successive short-term signals is one pitch-period.
4.5. **Example of Synthesis using the DSM**

The stochastic component is obtained by a direct application of the synthesis equation (4.3). We first synthesize a signal by filtering a unit-variance white Gaussian noise through a time-varying normalised lattice filter whose coefficients $k_i(t)$ are derived from the stream of $A(t_o^*, z)$ using a linear interpolation of the model coefficients on a sample-by-sample basis. Note that, since the lattice is normalised, the energy of the output signal is constant in time (and equal to unity). The time-domain energy envelope-function is then applied directly to the output of the lattice filter.

This section presents the application of the DSM model to a speech signal. The word ‘z/a/z/a/’ pronounced by a male speaker was recorded at a sampling rate of 16kHz. The stream of pitch-synchronous time-instants was first calculated, and the estimation of the parameters for the deterministic part is performed as described in the previous section. Fig.4.2 presents a mixed-voiced part of the original signal (top signal) and the corresponding deterministic part (bottom signal). Notice that both the pitch and the amplitude variations are correctly modelled. The deterministic signal was then simply subtracted from the original signal, in the time-domain, to yield the stochastic part. Fig.4.3 presents the frequency contents of a portion of the error signal. Note that the low-frequency sinusoidal components have been *entirely cancelled* during the subtraction.

![Figure 4.2: Original speech signal (top) and the corresponding deterministic signal (bottom)](image-url)
The stochastic part was analysed as described above for every analysis-frame, and an all-pole 15th-order filter was estimated. The energy-envelope function $e(t)$, see (4.3), was calculated by squaring the stochastic signal, low-pass filtering it with a Hanning window, and decimating by a factor 10. Fig.(4.4) presents the stochastic part (top signal) corresponding to the segment shown in the preceding figures. The bottom signal is the synthetic stochastic part.

The synthetic signal generated by DSM is indistinguishable from the original signal. However, as was mentioned above, the deterministic part has entirely cancelled the low frequencies (see Fig.(4.3)). Unfortunately, this means that the DSM does not separate the speech signal itself into a deterministic part and a stochastic part; in the deterministic part, a noisy signal will be included which by definition "belongs" to the stochastic part. Using the same algorithm which was used in the previous chapter for computing the synthesis time-instants when speech modifications were to be carried out, some synthesis time-instants were repeated. If the DSM is used to modify the prosody of a speech signal, for speech segments where the synthesis frames have to be repeated, the noise signal included into the deterministic part will be correlated from one frame to another (repeated frames). This leads to a synthetic sound with a slightly metallic quality (a shortcoming also of TD-PSOLA). For this reason, this model is not appropriate for speech modification. However, in the next section it will be shown that the DSM is more robust when it comes to pitch inaccuracies than the INM and that using the DSM it is possible to rectify errors in the pitch estimates. Then the complex slope can be omitted and the re-estimation of the model parameters -the complex amplitudes without the
4.6. REFINING THE PITCH ESTIMATION AND PASSAGE FROM DSM TO HNM

As discussed earlier, the accuracy of the pitch estimate is not critical for high-quality synthesis in the DSM. However, an accurate estimate of the pitch could lead to a better estimation of the amplitudes $a_k(t_a^i)$ and, assuming that the harmonic amplitudes vary linearly within a speech frame (as was for HNM), one can omit the complex slopes at the analysis step. This model is then essentially the HNM. Note that one could refine the pitch as described in Section 3.3.1. Alternatively, when a classic (and perhaps faster) pitch estimation procedure (see [Hes83]) was used for a coarse estimate of the pitch, it can be shown now that the calculated values of $a_k(t_a^i)$ and $b_k(t_a^i)$ can be used to refine this pitch estimate.

According to the (4.1) and to the (4.2), the instantaneous phase of the $k$th harmonic at time $t$ is given by

$$\theta_k(t) = k2\pi f_0(t_a^i)(t - t_a^i) + \arg\left(a_k(t_a^i) + (t - t_a^i)b_k(t_a^i)\right)$$ \hspace{1cm} (4.10)

in which one recognises a linear term $k2\pi f_0(t_a^i)(t - t_a^i)$ to which a corrective term is added.

Figure 4.4: Stochastic part (top signal) of the speech signal of the preceding figure. The bottom signal is the synthesized stochastic part.
If \( a_k(t_a^i) \) and \( b_k(t_a^i) \) have the same argument, then the corrective term has a constant phase which is equal to \( \theta_k(t_a^i) \). This happens when the pitch is constant and the estimate is correct. If however \( a_k(t_a^i) \) and \( b_k(t_a^i) \) do not have the same argument, then the phase of the corrective term varies with time between \( \arg(a_k(t_a^i) - b_k(t_a^i)N) \) and \( \arg(a_k(t_a^i) + b_k(t_a^i)N) \) where \( N \) is the integer closest to pitch period expressed in samples. This happens when the pitch estimate is biased or when the pitch varies. The time behaviour of the corrective phase is not linear and does not exactly correspond, therefore, to a correction of the frequency \( f_0(t_a^i) \).

Fig.4.5 is an example of such a case: The signal is composed of a single sinusoid, whose pitch has been severely underestimated. The original unwrapped linear phase is shown in Fig.4.5 (solid line). The least-square analysis described above was performed with an estimated pitch \( \hat{f}_0 \) 20% lower than the correct pitch \( f_0 \), and with a window length of twice the estimated pitch period. The dashed line displays the phase of the synthetic sinusoid, as given by (4.10), while the dotted line displays the phase of the synthetic sinusoid without linear complex amplitude \( k2\pi f_0(t_a^i)(t - t_a^i) \). Because the corrective phase in (4.10) is not linear, the dashed line oscillates around the original linear phase, but it is clear in Fig.4.5 that the linear complex-amplitude model makes it possible to approximately follow slow phase drifts and to refine the pitch estimate. This is done in the following way:

Using parameters \( a_k(t_a^i) \) and \( b_k(t_a^i) \), the instantaneous frequency of each harmonic \( \hat{f}_k(t) \) is calculated by derivative (4.10) with respect to \( t \). An optimal fundamental frequency \( \hat{f}_0(t_a^i) \) is found by minimising the cumulative weighted least-square error between \( \hat{f}_0(t_a^i) \) and \( f_k(t)/k \):

\[
A = \sum_{t=-N}^{N} \sum_{k=1}^{K} w(t)(\hat{f}_0(t_i) - f_k(t)/k)^2
\]

(4.11)
4.7. DESCRIPTION OF THE HSM

in which the same weighting window as in (4.4) was used. The solution is given by:

\[
\hat{f}_0(t_a^i) = \frac{\sum_{t=-N}^{N} \sum_{k=1}^{K(t_a^i)} w(t) f_k(t)/k}{\sum_{t=-N}^{N} \sum_{k=1}^{K(t_a^i)} w(t)}
\]  

(4.12)

Once the pitch has been refined, only the values of \(a_k(t_a^i)\) are re-computed by (4.8) in which the coarse pitch \(f_0(t_a^i)\) is replaced by \(\hat{f}_0(t_a^i)\). In that case, \(B_3\) and \(B_4\) are omitted from (4.8). Only one iteration is necessary to correct pitch period biases up to 1 sample (at 16kHz). The phase and the real amplitude of the kth harmonic at time \(t_a^i\) are then given by \(\arg(a_k(t_a^i))\) and \(|a_k(t_a^i)|\) respectively. The next analysis-time \(t_a^{i+1}\) is calculated according to the new pitch estimate as \(t_a^{i+1} = t_a^i + f_a/\hat{f}_0(t_a^i)\). This ensures the analysis remains pitch synchronous.

Using the amplitudes and the phases estimated as above, synthesis and speech modifications can be performed in a manner similar to that described for the HNM (see previous chapter).

4.7 Description of the HSM

From the above analysis, it turns out that the DSM, using linearly varying complex amplitudes, can model the speech signal accurately but it can not decompose the speech signal efficiently. This fact makes the DSM unsuited to speech modifications. To counter this, another original approach for speech, the HSM-Harmonic plus Stochastic Model, is now presented. The HSM, like DSM, aims to decompose the speech signal into two parts: one part accounting for the periodic structure of the signal and the other for the non-periodic structure. The first part is called the harmonic part and the second the stochastic (or residual). Accordingly, a speech signal, \(s(t)\), can be written as

\[
s(t) = h(t) + q(t)
\]  

(4.13)

where \(h(t)\) represents the harmonic part and \(q(t)\) the residual signal which accounts for everything in the signal that is not described by the harmonic part.

In this section we focus on an efficient determination of \(h(t)\). This means that \(h(t)\) should account for the periodic structure of the signal and for this alone; all the remaining information should be found into the stochastic part. It is worth noting that in this case the residual signal \(q(t)\) can be modelled in a manner similar to that described for the DSM. Therefore, the problem of \(q(t)\) modelling will not be treated here.
The harmonic part can be expressed in general as

\[ h(t) = \sum_{k=0}^{L(t)} a_k(t) \cos(\varphi_k(t)) \]  

(4.14)

where \(a_k(t)\) is a \(n\)-th order polynomial with real coefficients for the amplitude of the \(k\)-th order harmonic and \(\varphi_k(t)\) is a \(l\)-th polynomial of the phase also for the \(k\)-th harmonic. \(L(t)\) is the number of harmonics included in the harmonic part. In the present discussion the order, \(n\), of the amplitude polynomial is set to 3 and for the order, \(l\), of the phase polynomial we will discuss two possibilities: first and third order. The above parameters are updated at specific analysis time-instants denoted by \(t_a^i\), as was the case for HNM. When we denote the pitch period at \(t_a^i\) to be \(P(t_a^i)\) and \(N\) the integer closest to \(P(t_a^i)\), then for \(|t - t_a^i| \leq N\), the \(a_k(t)\) and \(\varphi_k(t)\) can be written as

\[ a_k(t) = \alpha_k + \beta_k (t - t_a^i) + \gamma_k (t - t_a^i)^2 + \delta_k (t - t_a^i)^3 \]

\[ \varphi_k(t) = \epsilon_k + 2\pi k \left[ \zeta (t - t_a^i) + \eta (t - t_a^i)^2 + \theta (t - t_a^i)^3 \right] \]

or

\[ \bar{\varphi}_k(t) = \epsilon_k + 2\pi k \zeta (t - t_a^i) \]

(4.15)

Then, the instantaneous frequency, \(f_0(t)\), within the analysis frame \(|t - t_a^i| \leq N\), is defined by

\[ f_0(t) = \frac{d}{dt} \frac{d \varphi_k(t)}{k 2\pi} \]

\[ = \zeta + 2\eta (t - t_a^i) + 3\theta (t - t_a^i)^2 \]

or

\[ f_0(t) = \zeta \]

(4.16)

where the second expression of the frequency corresponds to the first order phase polynomial, so in this case the instantaneous frequency is a constant (within the analysis frame). If we know the fundamentals frequencies at \(t = t_a^i - N, t = t_a^i, \) and \(t = t_a^i + N, \) and they are denoted respectively by \(f_0^{-1}, f_0^i, \) and \(f_0^{i+1} \), then the unknown coefficients of the frequency polynomial (which are also the same for the phase polynomial except the coefficients \(\epsilon_k\) are given by

\[ \zeta = f_0(t_a^i) \]

\[ = f_0^i \]

\[ \eta = \frac{f_0^{i+1} - f_0^{-1}}{4N} \]

\[ \theta = \frac{f_0^{i+1} + f_0^{-1} - 2f_0^i}{6N^2} \]

(4.17)

The estimation of the others coefficients for the phase and the amplitude polynomials will be discussed in the next section.
4.8. ESTIMATION OF HSM PARAMETERS

Given the harmonic part, the stochastic part is then defined as the residual signal given by (4.13)

\[ q(t) = s(t) - h(t) \]  

(4.18)

As discussed earlier, the residual signal has a specific time-domain structure in terms of energy localisation. Modelling this residual signal can be done in the same way as for the DSM: an AR model for the frequency aspect, and a piece wise linear envelope function for the time-behaviour. Throughout this chapter we will consider only the estimation problem of the harmonic part parameters as the estimation of the residual signal parameters was developed when the DSM was presented.

4.8 Estimation of HSM parameters

In this section, we focus on the coefficient estimation problem of the amplitude and phase polynomials, assuming that the values of the pitch and maximum voiced frequency at each analysis time-instant \( t_a^i \) have been already estimated using the pitch estimation procedure developed in the previous Chapter. Also the number of harmonics, \( L(t) \), within each analysis frame can be considered to be a constant parameter and equal to \( F(t_a^i)/f_0(t_a^i) \), where \( F(t_a^i) \) denotes the maximum voiced frequency at the time-instant \( t_a^i \), and \( f_0(t_a^i) \) the local fundamental frequency. Note that in this case, the coefficients \( \epsilon_k \) are the only unknown parameters of the phase polynomial. In the following, the estimation procedure for the \( \epsilon_k \) coefficients and for the coefficients of the amplitude polynomial is discussed.

To evaluate the unknown parameters of the new model, the same weighted least-squares criterion used for the other models will be used:

\[ \varepsilon = \sum_{t=t_a^i-N}^{t_a^i+N} w^2(t) (s(t) - h(t))^2 \]  

(4.19)

where \( s(t) \) is the original signal, \( h(t) \) is the harmonic part defined as in (4.14) and \( w(t) \) is the weighting function. However, as the coefficients \( \epsilon_k \) are unknown and the amplitude polynomial has an order greater than zero, it can be shown that the above criterion leads to nonlinear equations and then, that the solution must be calculated using relaxation methods. In order to avoid this iterative solution, the following analysis scheme was adapted. Denoting by \( \phi_k^i \), the phase value of the \( k \)-th harmonic at the time-instant \( t_a^i \), and evaluating the phase polynomial at \( t_a^i \), using (4.15), then the coefficients \( \epsilon_k \) are given by

\[ \epsilon_k = \phi_k^i(t_a^i) \]  

(4.20)
Note that $\phi_k(t_a)$ can be efficiently estimated using the HNM presented in the previous chapter. In fact, the HNM makes use of a zero order polynomial for the harmonic amplitudes and a first order polynomial for the phase. Also, in the previous chapter it was shown that in this case the system to solve is characterised by a Toeplitz matrix so that fast algorithms can be used in order to obtain the solution. Hence, using the HNM analysis step, we estimate the phase $\phi_k(t_a)$ and so the coefficients $\epsilon_k$ (4.20). Having determined all the coefficients of the phase polynomial, the next step consists of estimating the coefficients of the amplitude polynomial.

As $\varphi(t)$ is known for each $t \in [t_a - N, t_a + N]$, it turns out that the solution to the least-squares error defined by (4.19) is given by a linear set of equations. Specifically, within the analysis frame $[t_a - N, t_a + N]$ the harmonic part can be rewritten as

$$h(t) = \sum_{k=0}^{L} \alpha_k \cos(\varphi_k(t)) + (t - t_a) \sum_{k=0}^{L} \beta_k \cos(\varphi_k(t)) + (t - t_a)^2 \sum_{k=0}^{L} \gamma_k \cos(\varphi_k(t)) + (t - t_a)^3 \sum_{k=0}^{L} \delta_k \cos(\varphi_k(t))$$

(4.21)

where the number of harmonics, $L$, has been considered as constant parameter. In matrix notation, the above equations can be written as

$$\mathbf{h} = \mathbf{B} \mathbf{x}$$

(4.22)

where $\mathbf{B}$ is a $(2N + 1)$-by-$(4L + 4)$ matrix defined by

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 : \text{diag}(t) \mathbf{B}_1 : \text{diag}(t)^2 \mathbf{B}_1 : \text{diag}(t)^3 \mathbf{B}_1 \end{bmatrix}$$

(4.23)

where $\mathbf{t} = [-N, -N + 1, \cdots, N]^T$ denotes the $(2N + 1)$-by-1 time vector, $\text{diag}(t)$ denotes a $(2N + 1)$-by-$(2N + 1)$ diagonal matrix with $t$ the diagonal entries of this matrix and $\mathbf{B}_1$ is a $(2N + 1)$-by-$(L + 1)$ matrix given by

$$\mathbf{B}_1 = \begin{bmatrix} \mathbf{b}_1 : \mathbf{b}_2 : \cdots : \mathbf{b}_L : \mathbf{1} \end{bmatrix}$$

(4.24)

where $\mathbf{b}_k$ is a $(2N + 1)$-by-1 vector defined by

$$\mathbf{b}_k = [\cos(\varphi_k(-N)) \cos(\varphi_k(-N + 1)) \cdots \cos(\varphi_k(N))]^T \quad \text{for } 1 \leq k \leq L$$

(4.25)

and $\mathbf{1}$ denotes $(N + 1)$-by-1 unit vector: $\mathbf{1} = [1 1 1 \cdots 1]^T$. In (4.25), $\varphi_k(t)$ represents the first or the third order phase polynomial. Note that $N$ denotes the integer closest to the local pitch period and that for convenience of notation the centre of the analysis frame, $t_a$, was set to 0. The vector $\mathbf{x}$ denotes an $(4L + 4)$-by-1 vector which contains the unknown coefficients of the amplitude polynomial and is defined by

$$\mathbf{x} = [\alpha_0 \alpha_1 \cdots \alpha_L \beta_0 \beta_1 \beta_2 \cdots \beta_L \gamma_0 \gamma_1 \gamma_2 \cdots \gamma_L \delta_0 \delta_1 \delta_2 \cdots \delta_L]^T$$

(4.26)
4.9. **COMPARISON OF THE THREE HARMONIC PLUS NOISE MODELS**

If \( s \) denotes the \((2N + 1)\)-by-1 vector of the original speech samples of the current frame

\[
s = [s(-N) s(-N + 1) \cdots s(N)]^T
\]  
(4.27)

and \( W \) denotes the \((2N + 1)\)-by-\((2N + 1)\) diagonal matrix with diagonal entries the weight vector

\[
w = [w(-N) w(-N + 1) \cdots w(N)]^T
\]  
(4.28)

which is a typical Hamming window, then the solution to the least-squares problem is given by the normal equations

\[
(B^T W^T B) x = B^T W^T W s
\]  
(4.29)

or

\[
x = (B^T W^T B)^{-1} B^T W^T W s
\]  
(4.30)

Note that the harmonic part is readily obtained by computing \( \hat{h} = W^{-1} B x \) and then the residual signal \( q(t) \) is obtained by \( q(t) = s(t) - \hat{h}(t) \).

The application of the HSM on speech signals will be presented in the next section with an extensive comparison to the HNM and the DSM.

### 4.9 Comparison of the three harmonic plus noise models

After the above analysis and presentation of the three harmonic plus noise models, HNM, DSM and HSM, it seems reasonable to compare them in order to conclude this first part of the thesis. For convenience, we repeat at this point the three basic mathematical expressions of the harmonic part of each model:

- **HNM**: Sum of exponential functions without slope

  \[
h_1(t) = \sum_{k=-L}^{L} a_k(t^i_a) e^{j2\pi k f_0 (t-a)}
\]  
(4.31)

- **DSM**: Sum of exponential function with complex slope

  \[
h_2(t) = \sum_{k=-L}^{L} A_k(t) e^{j2\pi k f_0 (t-a)}
\]  
(4.32)

where

\[
A_k(t) = a_k(t^i_a) + (t - t^i_a) b_k(t^i_a)
\]  
(4.33)

with \( a_k(t^i_a), b_k(t^i_a) \) to be complex numbers (amplitude and slope respectively).
• HSM : Sum of sinusoids with time-varying real amplitudes

\[ h_3(t) = \sum_{k=0}^{L} a_k(t) \cos(\varphi_k(t)) \]  

(4.34)

where

\[ a_k(t) = \alpha_k + \beta_k (t - t_a^k) + \gamma_k (t - t_a^k)^2 + \delta_k (t - t_a^k)^3 \]

\[ \varphi_k(t) = \epsilon_k + 2\pi k[\zeta (t - t_a^k) + \eta (t - t_a^k)^2 + \theta (t - t_a^k)^3] \]

or

\[ \varphi_k(t) = \epsilon_k + 2\pi k \zeta (t - t_a^k) \]  

(4.35)

It is worth noting that all the above expressions for the harmonic parts of the three models, can be written in the matrix notations as

for HNM: \[ B_1 x_1 = h_1 \]

for DSM: \[ B_2 x_2 = h_2 \]

for HSM: \[ B_3 x_3 = h_3 \]  

(4.36)

where \( B_1 \) is a \((2N+1)\)-by-\((2L+1)\) matrix, \( B_2 \) is a \((2N+1)\)-by-\((4L+2)\) matrix and \( B_3 \) is a \((2N+1)\)-by-\((4L+4)\) matrix. We recall that \( N \) is the integer closest to the local pitch period and \( L \) represents the number of harmonics included in the harmonic (or deterministic) part.

Using the same weighted least-squares criterion, and given that all these models can be written as a linear set of equation, the solution has the same expression for all of them, with the difference that the "basic" matrix to invert is not the same for each model. Specifically, the solution for each of these models is given by

HNM: \[ x_1 = (B_1^H W^H B_1)^{-1} B_1^H W^H Ws \]

DSM: \[ x_2 = (B_2^H W^H B_2)^{-1} B_2^H W^H Ws \]  

(4.37)

HSM: \[ x_3 = (B_3^T W^T B_3)^{-1} B_3^T W^T Ws \]

As was showed in the previous Chapter, the matrix to invert in (4.37) for the HNM, is Toeplitz which means that fast algorithms can be used to solve the respective linear set of equations. Also, it can be shown that the matrix to invert for the DSM, is a block Toeplitz matrix and like the HNM, fast algorithms can be used in order to estimate the complex amplitudes and slopes of the deterministic part. Unfortunately, the corresponding matrix of the HSM has not such an attractive structure. However, it is symmetric and positive definite. Therefore, one method that could be used is the Cholesky decomposition[PTVF94] for solving the linear set of equations of the HSM. Note that this method is about a factor of two faster than alternative methods for solving linear equations[PTVF94].

In the next sections we compare the three proposed models, the HNM, the DSM and the HSM, by considering the variance of the residual signals produced from each of these models, the frequency content of the residual signals, the robustness of the models to background additive noise and finally to their modelling error.
4.9. **COMPARISON OF THE THREE HARMONIC PLUS NOISE MODELS**

4.9.1 **Comparison of the residual signal variance**

In Section 4.3.1, it was shown that the variance of the residual signal, \( q(t) \), is the diagonal of the follow matrix:

\[
E(qq^h) = I - WB(B^hWB)^{-1}B^hW^h
\]  
(4.38)

Using the same weighting window \( w(t) \), a typical Hamming window, the variance of the residual signal from each of the above harmonic models has been computed. The three variances are depicted in the Fig.4.6. The variance of the HNM residual signal is represented by a solid line, the variance of the DSM residual by a dashed line and finally the variance of the HSM residual signal by dashdotted line. As is made clear by Fig.4.6 the variance of the HNM residual

![Figure 4.6: Variance of the least-squares residual signals from: HNM (solid line), DSM (dashed line) and HSM (dash dotted line). The weighting window was a typical hamming window.](image)

signal is not evenly distributed across the analysis frame (as it should ideally be) in contrast to the variance of the residual signal obtained from the two other models. The variance of the DSM residual error is comparable with the variance of the residual produced from HSM. Furthermore, it can be seen that the variance from HSM is closest to the ideal least-squares variance for most of the analysis time than the variance of the DSM residual signal. However at the boundaries of the analysis frame the variance of the HSM residual becomes more noisy.

4.9.2 **Comparison of the residual signal**

In this section, the three models are compared based on their residual signal. From the discussion presented in the previous sections, we recall that the HNM makes use of a stationary
model for the estimation of the harmonic amplitudes and phases. The amplitudes and the phases estimated by this model correspond at the centre of each analysis frame and there is, therefore, no information about the parameters variation within the analysis frame. This causes low frequencies to appear in the residual signal of the HNM, with the result that some of the information from the periodic structure of the original speech signal passes to the residual signal. To show the above behaviour of the HNM residual signal, a voiced fricative frame from an original speech signal was selected. The frame used in the present study is plotted in Fig.4.7(a); the length of the analysis frame is two times the local pitch period. The maximum voiced frequency was 3500Hz. Fig.4.7(b) shows the residual signal from the HNM. This figure, clearly shows the above indicated behaviour of the HNM residual signal. For the same frame

![Graphs showing residual signals](image)

Figure 4.7: (a) a fricative voiced of an original speech signal and the residual error signals from (b) HNM, (c) DSM and (d) HSM.

the residual signals from the DSM and HSM are computed and are presented in Fig.4.7(c) and (d) respectively. Based on the results shown in this figure we may make the observation that the residual errors from DSM and HSM are very similar and that the information about the periodicity of the original speech signal is not detectable in these residual signals.

The measure of the similarity of the harmonic (or deterministic) part with the original
4.9. COMPARISON OF THE THREE HARMONIC PLUS NOISE MODELS

speech signal which was used is given by

\[ E = 10 \log_{10} \frac{\sigma_q^2(t)}{\sigma_s^2(t)} \]  \hspace{1cm} (4.39)

where \( \sigma_q^2(t) \) denotes the variance of the residual signal \( q(t) \) and \( \sigma_s^2(t) \) denotes the variance of the original speech signal \( s(t) \). For example, the error produced by the three models for the original speech frame in Fig.4.7 was \(-15.8dB\) for HNM, \(-25.56dB\) for DSM and \(-25.58dB\) for HSM.

At first glance in Fig.4.7, it might be tempting to say that since the DSM and HSM residual signals appear to be noise signals and that, therefore, the periodic structure of the original speech signal would be found into the corresponding deterministic or harmonic parts (so the speech decomposition was efficient). This is true, but as was indicated earlier, the deterministic part of the DSM has cancelled the noise signal in the harmonic band; so this noise signal was passed on into the deterministic part. Fortunately, this is not the case for the HSM.

The above properties of the DSM and the HSM can be shown in the next figures. In Fig.4.8(a), both the magnitude in dB of the Fourier transform of the DSM residual signal (solid line) and of the original signal frame (dashed line) are plotted. The same representation of the HSM residual signal is presented by a solid line in Fig.4.8(b) (where the Fourier transform of the original signal has also been included (dashed line)). The comparison of part (a) and (b) of the Fig.4.8 shows that the DSM has cancelled the noise part of the harmonic band (in this example from 0 to 3500Hz) in contrast to the HSM which leaves the noise part untouched. Note that the modelling error for the two approaches is comparable; this means that the spectrum content of the residual signal in the harmonic band, produced from the HSM should not be included in the harmonic part. To make this property of the HSM clearer, the following test was done: a white noise was filtered by a low pass filter with cutoff frequency equal to the maximum voiced frequency (3500Hz) and it was added to the original speech signal; the DSM and HSM parameters are estimated using the resulting noise corrupted speech signal. The frequency content of the two residual signal (from the DSM and HSM) are shown in Fig.4.9. This figure clearly shows that the additive noise (about 25dB) has passed in the deterministic part in the case of DSM, as is not presented into the harmonic band of the DSM residual signal, in contrast to the HSM, where the additive noise remains into the stochastic part (the noise level into the harmonic band has been clearly augmented). In all the above examples the HSM made use of a first order polynomial for the phase. For the variance and the modelling error there is no difference between the two approaches. Provided that the variation in the pitch contour is slow compared with the time and given the definition of the phase polynomial by (4.15), the two approaches become very close. In contrast, if the pitch values change rapidly
Figure 4.8: Magnitude of the Fourier transform of the residual signal (solid line) from (a) DSM and (b) HSM. The magnitude of the Fourier transform of the original speech signal has been also included (dashed line).

Figure 4.9: The same as in Fig.4.8 with an additive noise.
within the analysis frame, the two approaches have different performances. Such an example (from a part of the original speech signal where the pitch varies rapidly) is shown in Fig. 4.10 where the same test of the additive noise was performed. Note that in this case the noise part in the harmonic band with first order polynomial is more flat than the noise when a third order polynomial is used. Also it is worth noting that in the low frequencies the HSM with first order phase polynomial leaves more noise at the stochastic part than the HSM with the third order phase polynomial. Note that from the above examples, the Fourier transform of the residual

![Figure 4.10: Comparison of the third order with the first order of the phase polynomial for the HSM.](image)

signal produced from HNM has not be included, because by the Fig. 4.7 it is evident that the HNM is not well suited for an efficient decomposition of the speech signal.

A last comparison between the three models considers the modelling error as it was defined by (4.39). Fig. 4.11(a) shows a segment of a speech signal and in (b) the modelling error in dB produced from the three models is plotted. The modelling error from HNM is presented by a solid line, the error of DSM by a dashed line and the modelling errors of the HSM using first and third order phase polynomial by a dashdotted and dotted line respectively. It is clear by this figure that the modelling error using HNM is greater than those produced by the two other models. As the full frequency band (from 0Hz up to half of the sampling frequency) of the residual signal is used in the definition of the modelling error, the error is large on voiced fricative regions (for example, between the 25 and 35 analysis frame). The modelling
errors from the HSM using third and first phase polynomial are very close as the pitch contour varies slowly. To give an acoustic notion to the modelling error, it is worth noting that if the modelling error is about $-25dB$, the synthetic speech signal produces from the models is indistinguishable from the original speech signal.

Preliminary results have shown that the third order phase polynomial leads to smaller modelling errors when the pitch values from one frame to another changes rapidly. However, as the HSM model has not been extensively tested for speech analysis and resynthesis, any conclusion for the phase order risks to be inaccurate. One of the first future research directions that we propose is to test the HSM which, from the above analysis, appears to be an efficient model for speech.

### 4.10 Conclusion

This chapter has presented two models which decompose the speech signal into two parts: a deterministic or harmonic part and a stochastic part. This type of decomposition appears to be very important for high-quality pitch and time-scale modification. The first proposed model, called DSM-Deterministic plus Stochastic Model, makes use of linearly varying complex
amplitudes. The second model called HSM-Harmonic plus Stochastic Model makes use of a third order polynomial of real coefficients for the amplitudes and there are two possibilities for the phase: first and third order polynomials.

The proposed models were compared with the Harmonic plus Noise Model, HNM, developed in the previous chapter. The comparison focused on the variance of the residual signal obtained from the three hybrid models, the frequency content of this residual signal and the error modelling. Overall, the HSM was shown to be an effective speaker model capable of achieving low modelling error and efficiently decomposing the speech signal into a harmonic and a stochastic part. The DSM achieves low modelling error but it passes all the energy of the harmonic band (from 0Hz to maximum voiced frequency) into the deterministic part, thus cancelling the stochastic part in this band. For this reason the DSM is unsuitable to speech modifications. In contrast, it is well suited for speech resynthesis, as the synthetic signal is indistinguishable from the original speech signal. Note also that the HSM has only recently proposed[Sty96b][Sty96c][Sty96a], so all the above results could be considered as preliminaries. However, the HSM appears to be a very promising model for speech synthesis and for this reason, in the last chapter of this dissertation, an extension of the present work on the HSM is proposed.

From the analysis presented in this chapter, the HNM has found to have the largest modelling error among the proposed hybrid models. This was expected, as the HNM does not include any information about the variation of the harmonic amplitudes and phases within the analysis frame; this is supposed to be linear when the synthetic sound is generated. Clearly, the HNM cannot be proposed for speech decomposition. However, as was mentioned in the previous chapter, the HNM is capable of producing speech resynthesis which is almost perceptually indistinguishable from the original speech signal. The prosodic modifications produced from the model are also of very high-quality. Taking into account the fact that the HNM parameters can be estimated using fast algorithms and given the high-quality of prosodic modification performed by the model, in spite of its relative simplicity, the HNM is finally the model that we will use for the speaker modification presented in the second part of the thesis.
CHAPTER 4. NEW DECOMPOSITION TECHNIQUES OF THE SPEECH SIGNAL
Part II

SPEAKER MODIFICATION
Chapter 5

Background for speaker modification

5.1 Introduction

The purpose of this chapter is to provide a background review of speaker modification (or voice conversion) and introduces notation and nomenclature for the second part of the thesis. Speaker modification is related to speaker identification and speaker recognition in that the above areas are concerned with the perceptual cues used by listeners in identifying speakers and their relation to acoustic parameters of the speech signal.

Text-independent speaker recognition aims at extracting the information relative to the speaker individuality from the speech waves without explicit reference to what is uttered. It proceeds by extracting from the speech samples, the information which is characteristic of the way a particular speaker utters individual phonemes, articulates strings of phonemes, and then produces words and sentences. The information which is pertinent for such a purpose is of course related to the physiological and the behavioural characteristics of the speaker. These characteristics exist both in the short-term spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics) of speech. Voice conversion consists exactly in the reverse operation: starting from the speech signal uttered by a speaker, (source speaker), it aims at transforming the characteristics of the speech signal, in such a way that a human listener could believe that the transformed speech is produced by another (target) speaker. In other words, the machine disguises the voice of the speaker to mislead the listener.

Using the HNM to modify the prosodic characteristics of the source speaker, the next
step for speaker modification\(^1\) consists of transforming the source spectral envelope to the target one. Hence, in the second part of the thesis we focus on the control of the spectral envelope and its robust transformation as well as the combination of prosodic and spectral transformations in order to modify the source speaker. The proposed speaker modification system will be discussed in the next Chapter. In the present Chapter some studies on the perceptual cues used by humans for speaker identification will be reviewed. This is followed by selecting the features for spectral transformation and finally, relatively new approaches for spectral transformations are reviewed.

This chapter is organized as follows. Section 5.2 gives a definition of speaker modification and discusses the purposes of this area of speech processing. Section 5.3 then examines the various features involved in voice individuality and some early studies of speaker identification are presented. This is followed in Section 5.4 by the description of the features used in this thesis in order to modify a speaker and their parameterization. Finally, in Section 5.5 an outline of some of the major spectral transformation techniques used in speaker modification systems are presented.

### 5.2 Speaker modification

#### 5.2.1 What is speaker modification?

The speech signal conveys a wide range of information of different natures. Among them, the meaning of the message being uttered is of prime importance, however, secondary information such as voice individuality also plays an important part in oral communication. Voice modification techniques attempt to transform the speech signals uttered by a given speaker so as to alter the characteristics of their voice. As the psychological correlates of voice individuality remain largely unknown, it is often convenient to specify the desired modifications of the voice characteristics by reference to an existing speaker (the so-called target speaker). This last problem - how to modify the speech of a speaker so that it sounds as if it was uttered by another one - is generally known as voice conversion [MS95].

#### 5.2.2 Purpose

In our daily life, voice individuality is useful because it enables us to differentiate between the various messages that we hear according to the speaker. If all voices sounded alike it

\(^{1}\)Throughout the second part of the thesis we will use both the terms of speaker modification and of voice conversion.
would for instance be almost impossible to follow a radio program involving different persons. Voice modification technology thus has many applications in all systems that make use of pre-recorded speech, such as vocal mailboxes or more elaborate text-to-speech synthesizers based on acoustical units concatenation. In such cases, voice modification would be a simple and efficient way to bring the desired variety to the spoken messages while avoiding the actual recording by different speakers [MS95].

Another reason why voice characteristics are useful is that they make it possible to identify the person who is speaking if we are familiar enough with him. Voice modification is thus an important aspect of ongoing projects on interpreted telephony. Such systems would make the communication between foreigners easier by first recognizing the speech sentences uttered by each speaker, and then translating and synthesizing them in a different language. In this application it is important for the naturalness of the conversation that the characteristics of each speaker's voice are to be maintained through the whole process. For the same reason, voice conversion techniques would also be needed in the context of speaking aids for the speech impaired.

It is finally interesting to note that the voice conversion problem is closely related to other familiar speech research topics that involve speaker individuality such as speaker adaptation or speaker recognition. The main difference being that in the case of voice conversion, the final output is a speech signal that is to be appreciated by a human listener.

5.3 ACOUSTICAL FEATURE PARAMETERS FOR SPEAKER IDENTIFICATION

In designing systems for speaker modification, it is helpful to understand the perceptual mechanisms used by human listeners in determining a speaker's identity. These mechanisms are not well understood, but roughly span from "high level" mechanisms which are related to semantic or linguistic aspects of speech to "low level" mechanisms which are related to the acoustic aspects of speech[Rey92] [KS95]. In [KS95] the high level and low level mechanisms are referred as "software" and "hardware" respectively. The high level mechanisms include such factors as word usage, idiosyncrasies in pronunciation and other non-speech characteristics that can be attributed to a particular speaker[Rey92].

These factors of voice individuality come more from the control commands to the speech organs, and are generally thought to arise from the individual's experience. The low level mechanisms are more directly related to the actual sound of a person's voice and include such attributes as soft or loud, clear or rough, slow or fast, etc. Perhaps the high level mechanism
may contain more important information for voice individuality than the low level. However, the high level mechanisms are not easily quantified and would be very difficult to extract them from the speech signal. On the other hand the low level mechanisms can be related to easily extracted acoustic parameters from the speech signal.

This section reviews several papers which have examined the speech characteristics humans use to recognize a speaker and how to correlate these perceptual characteristics to acoustic parameters which can be directly measured from the speech signal.

Existing studies on speaker recognition by humans indicate that voice individuality should be considered as the consequence of the combination of several factors. Among these factors, supra-segmental speech characteristics such as the speaking rate, the pitch contour or the duration of the pauses have been shown to contribute greatly to speaker individuality [Hol90] [JHH84] [Fur86] [Sam75]. In an attempt to show the contribution of the fundamental frequency to speaker identification, Compton [Com63] separated a voiced utterance into a low-pass signal (below 1000Hz) and a high-pass signal (above 1000Hz); he found that high-pass filtering the voiced utterance substantially reduced speaker identification performance, and low-pass filtering had no significant effect on performance.

LaRiviere [LaR75] attempted also to ascertain the contribution of fundamental frequency and formant frequencies to speaker identification. Using isolated vowel, he attempted to simulate a fundamental frequency only condition by filtering them with a lowpass filter at 200Hz. A formant frequency only condition was simulated by whispering the same vowels. He found that the contribution of voiced vowels is two times greater than the contribution of the whispered and filtered vowels. The contribution to speaker identification of the two last categories was found to be approximately equal. He also found that the back vowels /a/ and /æ/ yield better identification performance than front vowels /i/ and /u/, conclusion which is in contrast with the conclusion of Stevens et al. [SWCW68] that utterances containing front vowels (/i/) yield higher speaker identification scores than utterances containing back vowels (/a/). Similar studies of high-pass and low-pass filtering are reported in [Dod85]: R.W. Peters in 1954 showed that the octave band of frequencies from 1 to 2kHz is most useful to listeners for speaker recognition, while F.R. Clarke et al. in 1966 found that listeners have demonstrated only a modest doubling of speaker recognition error when speech signal is severely high-passed at 2kHz or low-passed at 1kHz.

Itoh and Saito [IS88] used the LPC analysis/synthesis system, they manipulated the spectrum envelope, the fundamental frequency, the speech power, and voiced/unvoiced ratio in order to synthesize various kinds of test sounds: by fixing the fundamental frequency at the averaged signal frequency of the test speakers, by fixing the fundamental frequency at the averaged frequency of each speaker and deleting the spectrum envelope, and finally by combining
the feature parameters of two speakers (spectrum envelope and pitch). The speech materials were short sentences and vowels both in Japanese. Their conclusion was that the spectrum envelope of speech signals has the greatest effect on speaker identification performance by hearing. The contribution ratios for the spectrum to the excitation source signal were 5:1 and 3:1 for short sentence and vowels respectively.

In many cases, it also appears that the perceived voice specificity is influenced by the linguistic style of speech [Hol90] [EBF71]. In the state of our knowledge the processing of such features of speech by an automatic system is delicate because high-level considerations are involved. In particular, the fact that both the meaning of the spoken message as well as the intention of the speaker have a strong influence on prosodic features clearly hinders their automatic processing in cases where the text of the speech utterance is not fixed a priori. Hopefully, it turns out that the average values of these features (average pitch frequency, overall speech dynamics) already carry a great deal of the speaker specific information [Sam75] [Fur86] [Hol90] [KS95].

There is also strong evidence that distinct speakers can be efficiently discriminated at the segmental level by comparison of their respective spectral envelopes [IS88] [Fur86]. Accordingly, most current speaker recognition techniques are based on the characterization of the statistical distribution of the spectral envelopes [Dod85] [RS91] [GS94]. It is generally admitted that the overall shape of the envelope as well as the formants characteristics are the major speaker identifying features of the spectral envelope [KS95] [Hol90] [Gol75]. However, some uncertainty remains about the respective contributions of these acoustics features to the individuality of the speaker. Recent studies suggest that some effective speaker specific features (onset times of resonant energy pulses) can also be extracted directly from the speech waveform in the time domain [QJR94].

5.4 Features selection

From the above analysis it is not evident to specify the single acoustic parameter that plays a decisive role in voice individuality. As Kuwabara and Sagisaka mentioned in [KS95], there is no single specific acoustic parameter alone which carries the entire individuality information, but the voice quality is an amalgam of many parameters and the degree or order of importance among them can differ from speaker to speaker. The main outcome of the many feature selection studies was that features which represent pitch and duration as a function of time (prosodic parameters) as well as the speech spectrum were the most effective for speaker identification. Spectral measurements from nasals and vowels were found to be particularly good for speaker discrimination [Sam75].
In this thesis we will consider that both prosodic and spectral parameters have to be changed in order to convert a speaker to another one. However, for an automatic voice conversion system a proper prosodic analysis should be done. Our purpose is not this automatic conversion system (which is not a clear signal processing problem) but, to provide robust and flexible methods to do it. In fact, the HNM discussed in the previous Chapter has been shown to allow a flexible and precise (time-varying modification factors) way to control the prosody of the synthetic speech signal. Thus, given a prosodic model for the target speaker, the HNM is an appropriate model to adapt the prosody of source speaker to that of the target one. Furthermore, due to its parametric scheme, the HNM gives access to spectral envelope manipulations.

Common spectrum representations of the speech spectrum are linear predictive coefficients (LPC) and their various transformations (reflection coefficients, cepstral coefficients, log-area ratio parameters, etc.) [RS78]. One drawback to the LPC representation is that it is based on an all pole filter model, which may omit significant speech spectral characteristics. The cepstrum coefficients, however, and more precisely the discrete cepstrum coefficients proposed in [CLM95] have already used by HNM for pitch modification purposes. Furthermore, with a small order (16-20 for sampling frequency at 16kHz) they provide a continuous frequency spectral envelope which fits very well the estimated harmonic amplitudes. Thus, the discrete cepstral coefficients will be used for the spectrum representation and they will serve as the feature sets used in the spectral modification system. In the next section some recent techniques for the control and the transformation of the spectrum parameters (not necessarily represented by cepstral parameters) are discussed.

5.5 Spectral transformation techniques

In order to convert the spectral envelopes of the source speaker to the spectral envelopes of the target one, a training (or learning step) is necessary. During this step, a conversion function is trained to convert the source spectral envelopes to the target ones. For this purpose, the two speakers utter the same sentences. This section reviews some recently proposed techniques and functions used for spectral transformation. The VQ-mapping codebook approach is described first. Next, the speech spectrum transformation by speaker interpolation is discussed. This is followed by a description of the Linear Multiple Regression, (LMR), an approach adapted for spectral transformation as well as the Dynamic Frequency Warping (DFW). Lastly, a neural networks approach is presented.
5.5. SPECTRAL TRANSFORMATION TECHNIQUES

5.5.0.1 Spectral transformation through vector quantization

One of the earliest approaches to the spectral conversion problem is the mapping codebook method of Abe et al. [ANSK88] [ANSK90] which was originally introduced by Shikano et al. for speaker adaptation [SLR86]. The basic idea of this technique is to make mapping codebooks which represent the correspondence between the two speakers. A conversion of acoustic features from one speaker to another is therefore reduced to the problem of mapping the codebooks of the two speakers [ANSK88]. The procedure for the mapping codebooks construction is given below [ANSK88]:

1. The source and target speakers pronounce a learning word set. Then all words are vector-quantified frame by frame.

2. Using a time warping technique (DTW) the correspondence between vectors of the same words for the two speakers is determined.

3. The vector correspondences between the speakers are accumulated as histograms. Using each histogram as a weighting function, the mapping codebook is defined as a linear combination of the target speaker’s vectors.

Typically, the acoustic space of a speaker is modelled by a vector quantization codebook of 32-256 vectors derived using the LBG algorithm [Gra84]. Let be noted by $H$ the weighting function (or better, the mapping matrix between source and target speaker) obtained at the learning step; this is a $m$-by-$m$ matrix, where $m$ denotes the codebook size. The sum of any line or column of the matrix $H$ is normalized to 1. An input source spectral vector, $x$, is classified to the closest template in the source speaker’s codebook using a distance measure of the form

$$d(\mu_s, x) = (\mu_s - x)^T W (\mu_s - x)$$

(5.1)

where $\mu_s$ is the $i$-th source template vector with $i = 1, \cdots, m$, and $W$ is a matrix used to allow different weightings to different directions in the feature space. When using features such as the log spectrum or cepstrum, the distance in (5.1) is a well established measure of spectral distance. The common Euclidean distance sets $W$ to be the identity matrix $I$, whereas the general Mahalanobis distance sets $W$ to the inverse covariance matrix of the speaker’s training vectors classified at the learning step in the class $i$.

Having classified the input vector $x$ in the class $l$ ($1 \leq l \leq m$), the converted vector, $\hat{y}$, is obtained by

$$\hat{y} = \sum_{k=1}^{m} h_{lk} \mu_{tk}$$

(5.2)
where \( \mu_{th} \) is the \( k \)-th template vector of the target speaker’s codebook and \( [h_{lk}] \) is the element of the \( l \)-line and \( k \)-column of the mapping matrix \( H \). Abe et al. [ANSK88] showed that in the female-to-female conversion task, the distortion decreased by 23\% compared to non conversion, by 45\% for the male-to-male conversion, and by 64\% for the male-to-female conversion. The codebook size was 256 and LPC parameters were used as spectrum parameters (order of LPC : 12). The main shortcoming of this method is the fact that the parameter space of the converted envelope is limited to a discrete set of envelopes. In practice, this restriction of the variability of the speech envelopes causes a severe drop in the quality of the converted speech signal. Several variations of this basic scheme have been investigated in order to overcome this limitation, including the use of fuzzy vector quantization, (FVQ) [KS95]. In FVQ, the input source speaker’s spectrum vector is not uniquely quantified at one template [NS89].

A voice conversion system based on the mapping codebook has been proposed by Abe et al. Mapping codebooks were also generated for power values and pitch frequencies using the training utterances. The size of these codebooks were set to 35-64. Speech conversion is then obtained by applying these mapping codebooks and synthesised using a LPC vocoder [ANSK88].

### 5.5.0.2 Speech spectrum transformation by speaker interpolation

Another technique for spectral transformation using spectral patterns between pre-stored multiple speakers proposed by Iwahashi and Sagisaka [IS94][IS95]. The spectral patterns are transformed smoothly by an interpolation ratio which is gradually changed. In this method the spectrum data of utterances spoken by a small number of speakers is pre-stored in the synthesis system. After spectrum sequences of the same utterance by multiple speakers are time-aligned by a Dynamic Time Warping, (DTW), technique, the interpolation is carried out between the time-aligned spectrum sequences by the following transformation:

\[
\hat{y}^i = \sum_{k=1}^{M} w_k \mathbf{x}_k^i
\]

under the constraint

\[
\sum_{k=1}^{M} w_k = 1
\]

Note that \( \mathbf{x}_k^i \) represents the spectral vector of the \( i \)-th frame in a time-aligned utterance of the \( k \)-th speaker, \( M \) is the number of pre-stored speakers, the \( w_k \) is the interpolation ration for \( k \)-th speaker and finally, \( \hat{y}^i \) represents the spectral vector of the \( i \)-th frame of the spectrum generated by interpolation. The optimal interpolation ratios \( w_1, w_2, \ldots, w_M \) are determined...
5.5. SPECTRAL TRANSFORMATION TECHNIQUES

by minimisation of the function

\[ F(w_1, w_2, \cdots, w_M) = \sum_i (\hat{y}^i - y^i)^2 \]  

(5.5)

where \( y^i \) represents the spectral vector of the \( i \)-th frame of the target speaker spectrum. This minimisation is done by solving normal equations. When Log Area Ratio (LAR) parameters are used as the interpolated parameters, a non-linear optimization process is carried out, whereas using cepstrum coefficients the above problem can be shown to be linear\cite{IS94}.

Using the above interpolation scheme, Iwahashi and Sagisaka \cite{IS94} showed that the distance between the spectrum of the target speaker and the spectrum generated by the interpolation is smaller by about 25\% than the distance between the spectrum of the target speaker and the spectrum of the pre-stored speaker closest to the target speaker. This distance reduction rate was found to be larger by interpolating cepstrum than LAR \cite{IS94}. In order to increase the flexibility of their system, Iwahashi and Sagisaka increased the number of adapted parameters. When cepstrum parameters are interpolated additive cepstrum parameters are adopted with the interpolation ratio. In this case, the adaptation is carried out by minimising the following function:

\[ F(w_1, w_2, \cdots, w_M, a) = \sum_i (\hat{y}^i - y^i)^2 \]  

(5.6)

with

\[ \hat{y}^i = \sum_{k=1}^M w_k x_k^i + a \]  

(5.7)

where \( a \) represents the additive cepstrum parameter. By using this adaptation method the distortion rates was found to increased to 36\%. The number of pre-stored speakers was four - two males and two females. Eight males and females were used as target speakers and the cepstrum and LAR order was 12.

5.5.0.3 Spectral transformation using Linear Multivariate Regression

A third method of spectral transformations proposed recently is that of the Linear Multivariate Regression, LMR, \cite{Val92} \cite{VMT92}. This method has been also used for speaker adaptation in speech recognition experiments. The first step of this method consists of modelling the acoustic space of the source speaker into non-overlapping classes using a standard unsupervised clustering technique \cite{LBG80}. Before the classification, the extracted spectral envelopes (cepstrum coefficients) from source and target test signals were time-aligned using a standard DTW technique. The LMR approach proposed a simple linear transformation function for each class.
CHAPTER 5. BACKGROUND FOR SPEAKER MODIFICATION

Let \( \{ \mathbf{x} \}_1^{N_k} \) be the set of the source spectral vectors and \( \{ \mathbf{y} \}_1^{N_k} \) the set of the target speaker belonging both to the \( k \)-th template (or class) of the source and target codebook respectively. Note that each of the spectrum vector has \( p \) dimension. The LMR approach estimates a linear regressor, that is a \( p \)-by-\( p \) matrix denoted by \( \mathbf{P}_k \) which minimises the mean square error between the normalized set of the source and target vectors denoted by \( \tilde{\mathbf{x}}_i \) and \( \tilde{\mathbf{y}}_i \) for each separately, and which are given by:

\[
\begin{align*}
\tilde{\mathbf{x}}_i &= \frac{\mathbf{x}_i - \mu_{x_k}}{\sigma_{x_i}} \\
\tilde{\mathbf{y}}_i &= \frac{\mathbf{y}_i - \mu_{y_k}}{\sigma_{y_i}}
\end{align*}
\]  

(5.8)

for \( i = 1, \cdots, N_k \), and where \( \mu_{x_k}, \mu_{y_k}, \sigma_{x_i}, \sigma_{y_i} \) denote respectively the within class sample mean and variance of \( \mathbf{x}_i \) and \( \mathbf{y}_i \). Thus, the matrix \( \mathbf{P}_k \) is obtained minimised the error

\[
\epsilon = \sum_{i=1}^{N_k} ||\mathbf{P}_k \tilde{\mathbf{x}}_i - \tilde{\mathbf{y}}_i||^2
\]  

(5.9)

where \( \mathbf{X} \) and \( \mathbf{Y} \) are the \( p \)-by-\( N_k \) matrices which contains the \( N_k \) spectrum vectors from the source and target speaker of the \( k \)-th class as they defined in (5.8). The solution is given by a system of normal equations :

\[
\mathbf{P}_k = \mathbf{Y} \left[ \mathbf{X}^T (\mathbf{X} \mathbf{X}^T)^{-1} \right].
\]  

(5.10)

The LMR method succeeds in moving the formants from their original frequencies to the target ones, but it seriously affects their amplitudes and their bandwidths. Unfortunately, using the LMR method was found that the timbre of the transformed speech does not sound "close" to the reference timbre of the target speaker, from a perceptual point of view [VMT92]. Moreover, the partition of the acoustic space of a speaker in distinct regions introduces some discontinuities into the spectral information.

5.5.0.4 Spectral transformations using a Dynamic Frequency Warping (DFW)

Another method proposed also by Valbret et al. [VMT92][Val92] for spectral transformation is the method of Dynamic Frequency Warping (DFW) originally proposed by Goncharoff et al. [GC88]. This method is closely related to the acoustic theory of speech production. Starting form the work of Fant [FKN91] that formant frequencies of one speaker are related to those of another one by non-uniform scaling, the DFW approach looks for a mapping path between the source and target speaker spectrum of the same class.

The procedure is the following : each reference and target log-magnitude spectra are computed and their spectral tilts are removed and stored for amplitude scaling; then a frequency
warping algorithm is then applied between each "normalized" source and target spectrum which leads to a warping curve [VMT92]. The number of warping function is equal to the number of pairs source-target spectral vectors within the class. Note that previous the spectrum vectors have been time-aligned. An average warping function is obtained for each class and then it is modelled using a third order polynomial.

Compared to the LMR technique, DFW needs much fewer parameters to approximate the mapping: 4 coefficients are enough to model the mapping in each class in contrast to LMR which requires storing a $p$-by-$p$ matrix, where $p$ is always the size of the cepstrum coefficients (typically $p$ is set to 20). However, DFW can change spectral shape only in the frequency domain and therefore while it can adjust resonance frequencies and their bandwidth shift, it has little affect on their amplitudes [KS95]. Also, it has been found that the converted source speech was closest to the target one using the LMR approach rather the DFW [Val92].

Using the LMR and DFW approaches to transform the spectral envelopes, Valbret et al. proposed a voice conversion scheme based on the PSOLA prosodic modification system [MC90]: the TD-PSOLA is used to alter the prosody characteristics of the source excitation signal; next, the source spectrum envelope is transformed by the LMR or DFW approach and finally, the converted speech is obtained using a classic LPC synthesiser by filtering the prosody modified excitation signal by a transformed spectrum envelope.

5.5.0.5 Spectral transformations using Neural Networks

Narendranath et al. [NMRY95] proposed a voice conversion scheme where the implicit transformation of formants is captured by a neural network. Thus the first step consists of a formant analysis phase, followed by the learning phase. Their observations were consistent with the conclusions conducted by Fant that the vocal tract shape transformation between two speakers is not linear. Based on the property that a multilayered feedforward neural network using nonlinear processing elements can capture any arbitrary input-output mapping only a few sample pairs of formant vectors can be used during the training phase.

A network with one input layer (3 units), two hidden layers (8 units each) and an output layer (3 units) was used by Narendranath et al. [NMRY95]. During the training phase the network is trained with a discrete set of points on the mapping function. The speech database uses steady voiced regions of continuous speech of the source and target speaker and the correspondences between source and target were identified manually. The first three formants were extracted for the test utterances using a method based on minimum phase group delay functions [MY91] and after they have been used as input and output for the neural network (input the source formants and output the target formants). Then weights are adjusted using
a backpropagation algorithm [NMRY95]. The algorithm is repeated until the weights converge.

The reduction rate of the distortion between transformed formants and target formants were found to be dependent on the vowel. For the first formant, a reduction rate of 67% was found for the vowels /i/ and /a/ whereas for the vowel /o/ 35%. The reduction rate for the second formant was 68% for the vowel /e/ and 25% for the vowels /o/ and /a/. Finally, for the third formant a reduction rate of 76% was achieved for /u/ and 55% for /a/.

Using the transformed formants and modifying the pitch contour of the source speaker by the ratio of the average pitch, $F_0$, of the source and target speaker a synthetic speech was generated by a formant synthesiser.

5.6 Conclusion

This chapter has presented a background review of the area of speaker modification or in other words of voice conversion. First, several studies which attempted to correlate the perceptual cues that distinguish a person’s voice to physical measurements of the speech waveform were discussed. These studies highlighted the difficulty in attempting to quantify the human perceptual process. It is clear that humans use several levels of knowledge in identifying speakers but it is not clear how these levels are used or which ones are the most important for speaker identification. Even with an imprecise characterisation, the studies did find that the amalgam of parameters such as the prosodic features, pitch and duration as a function of time, and the spectral shape play an important role in voice individuality.

Considering the prosodic modifications as the first step for voice conversion, this step has been already discussed in the first part of the thesis; the second step, spectral transformations, has now to be addressed. The discrete cepstral coefficients was selected as the features of the spectrum envelopes and will be used in the next Chapter for spectral transformations.

An overview of spectral transformation techniques was presented. Basically almost all these methods make use of methods initially proposed for speaker adaptation and speaker identification. In the codebook mapping proposed by Abe et al. the clustering is achieved through Vector Quantization (VQ) of the acoustic spaces of the source and target speaker. In this approach, the clustering procedure is applied to the spectral parameters of both the source and the target speakers. The two resulting VQ codebooks are used to obtain a mapping codebook whose entries represent the transformed spectral vectors corresponding to the centroids of the source speaker codebook. The main shortcoming of this method is the fact that the acoustic space of the converted signal is limited to a discrete set of envelopes. Most authors agree that the mapping codebook approach, although it provides an impressive perceptive voice conver-
sion effect, is plagued by its poor quality and its lack of robustness [MA95]. The spectral interpolation approach described in [IS94] and [IS95] solves these problems by interpolating between several speakers' spectra to determine the converted spectrum. The practical use of this method is however limited by the fact that it requires the pre-recording by a number of speakers of all the sentences that need to be converted.

Other recent publications suggest that a possible way to improve the quality of the converted speech consist in modifying only some specific aspects of the spectral envelope, such as the location of its formants [MA94] [MA95] [VMT92].

The LMR and DFW proposed in [Val92] have also presented. The first of the method, initially proposed for speaker identification, based on the idea of modelling the mapping in each class by a simple linear transformation. Unfortunately, this kind of transformation introduces discontinuities into the spectral information as the acoustic space of a speaker partitioned in distinct regions. The DFW approach succeeds in moving the formants frequencies but it has little effect on their amplitudes. Finally, a speech modification approach using artificial neural networks was presented. Notice that all the above methods used cepstrum coefficients for spectral envelope representation except the last one which makes use of formants.

The reported reduction rates of the spectral distortion between source transformed and target speaker can not be used as a criterion to choose the best approach as they have been applied to different databases and as also the explicit formula used for the computation of the spectral distance has not been given.

In the next Chapter an original approach for spectral transformation will be discussed.
Chapter 6

Statistical methods for speaker modification

6.1 Introduction

This chapter introduces the Gaussian Mixture Model as a robust representation of the acoustic space of a speaker, proposes a conversion function for spectral transformation and describes a voice conversion system. The discussion will focus on the conversion of the spectral envelope characteristics at the segmental level. More specifically, our aim is to represent by an appropriate model, trained from experimental data, the statistical relations between the spectral envelopes of two different speakers uttering the same text. To differentiate this last problem from the general voice conversion task, which would also necessitate a proper analysis and control of the prosodic characteristics, we will refer to the control of the spectral envelope as spectral conversion. An experimental evaluation of the conversion function is presented and a voice conversion system using the HNM to modify the source speaker prosody is discussed. The goal of the experiments was to evaluate the performance of the conversion function under a variety of parameterizations and conditions. The experiments examined some issues involved in training the conversion function, the performance effects of the number of mixture components and amount of training data used to model the acoustic space of speakers, and finally evaluated the performance of the voice conversion system using the HNM in order to convert the source speaker to the target one. The performance of the proposed function was judged from the spectral distortion of the transformed spectral envelopes to target and source spectral envelopes.
CHAPTER 6. STATISTICAL METHODS FOR SPEAKER MODIFICATION

The Chapter is organized as follows. The development and description of the GMM for speaker acoustic space modelling is given first. This is followed by a description of the maximum likelihood estimation procedure of the GMM parameters. Section 6.4 is devoted to the proposed conversion function for spectral transformations and to the optimization of its parameters, and Section 6.5 discusses the training step of the conversion function. Section 6.6 presents the voice conversion system which makes use of the HNM. Finally, in Section 6.7 the performance of the conversion function and the conversion system is showed and it is compared to the performance obtained using the classic method for voice conversion, the VQ-mapping codebook. The experimental results obtained for a conversion task between two male speakers are discussed.

6.2 The Gaussian Mixture Model

The Gaussian mixture density is a weighted sum of $m$ component densities and given by the equation

$$p(x|\Theta) = \sum_{i=1}^{m} \alpha_i p_i(x|\theta_i)$$  (6.1)

where $x = [x_1 \, x_2 \, x_3 \, \cdots \, x_p]^T$ is a $p$-by-1 dimensional random vector, $p_i(x|\theta_i)$, for $i = 1, \cdots, m$, are the component densities and $\alpha_i$ are the mixture weights. Each component density, $p_i(x|\theta_i)$, is a $p$-dimensional normal distribution of the form

$$p_i(x|\theta_i) = \frac{1}{(2\pi)^{p/2} |\Sigma_i|^{-1/2}} \exp \left[ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right].$$  (6.2)

with $\mu_i$ the $p$-by-1 mean vector and $\Sigma_i$ the $p$-by-$p$ covariance matrix. The mixture weights, $\alpha_i$, are normalized positive scalar weights ($\sum_{i=1}^{m} \alpha_i = 1$ and $\alpha_i \geq 0$). This ensures that the mixture is a true probability density function (pdf). The complete Gaussian mixture density is parameterized by the mixture weights, the mean vectors and the covariance matrices from all component densities which is represented by the notation,

$$\Theta = \{\alpha, \mu, \Sigma \}, \quad i = 1, \cdots, m$$  (6.3)

The Gaussian Mixture Model, GMM, is a classic parametric model used in many pattern recognition techniques [DH73] whose efficiency for text-independent speaker recognition has been illustrated by recent studies [RR90] [TSR92] [RR95].

As discussed in the previous chapter, it is desirable to model the acoustic space of a speaker in order to reduce the problem of the acoustic features conversion from one speaker to another by "mapping" properly the acoustic space of the two speakers. By modelling the speaker
dependent acoustic classes of a person’s voice, it is possible during conversion to minimize
the effects of textual differences between the training speech and test utterances. It is also
desirable that the acoustic class modelling be performed without explicit speech segmentation
(labelled data). The GMM is a direct realization of a probabilistic modelling of the underlying
acoustic classes of a speaker’s voice which does not require explicit speech segmentation.

In the GMM context, a speaker’s voice is characterized by \( m \) acoustic classes representing
some broad phonetic events, such as vowels, nasals or fricatives. The probabilistic modelling
of an acoustic class is important because, while an acoustic class will have some characteristic
spectral features, there will also be sources of variability in features from the same class due to
variation in pronunciation and co-articulation effects. Thus, the mean vector \( \mu_i \) represents the
average features for the acoustic class \( \omega_i \), and the covariance matrix \( \Sigma_i \) models the variability
of features within the acoustic class.

Since unlabelled data are used for training the GMM of a speaker, the acoustic classes are
considered to be "hidden" processes: all that is observable for speaker model training are the
feature vectors themselves with no labelling of which acoustic class they are from. One can find
some similarities between GMM and HMM (Hidden Markov Model). However, for the GMM
it is important to model the acoustic classes but not the temporal ordering of the acoustic
classes. In fact, the GMM can be viewed as a fully ergodic, uni-modal Gaussian state density
HMM with equal transition probabilities between states[Rey92].

6.3 GMM parameter estimation

This section is concerned with the problem of estimating \( \Theta \) in (6.1) using a sample of un-
labelled observed data. The first published investigation relating to the mixture density es-
timation problem for two univariate normal densities appears to be that of Pearson in 1894
([RW84]). The approach suggested by Pearson for solving the problem is known as the method
of moments. This method consists of equating some set of sample moments to their expected
values and thereby obtaining a system of (generally nonlinear) equations for the parameters
in the mixture density[RW84]. For example, in order to estimate the five independent param-
eters in a mixture of two univariate normal densities we have to solve a single ninth-degree polynomial[RW84]. The method of moments was usually the method of choice for the solution
of the above estimation problem until the widespread use of computers in the 1960’s. Then,
researchers began to turn from the method of moments to the method of maximum likeli-
hood; a maximum likelihood estimate associated with a sample of observations is a choice of parame-
ters which maximizes the probability density function of the sample, called in this context the
likelihood function. In the next section a maximum likelihood parameter estimation procedure, a special case of the Expectation- Maximization (EM) algorithm, is developed for the GMM.

6.3.1 Maximum likelihood estimation

For a sequence of $p$-dimensional observation vectors, $\{x_i\}_{i=1}^N$, assuming statistical independence between each vector and denoted by $X$ the set of the observations $x_i$: $X = \{x_1, x_2, \cdots, x_N\}$, the joint probability density of $X$ treated as a function of $\Theta$ is given by

$$P(X|\Theta) = \prod_{i=1}^N p(x_i|\Theta) \quad (6.4)$$

The (6.4) is known as the likelihood function for $\Theta$. When maximum likelihood estimates are of interest, it is usually convenient to deal with the logarithm of the likelihood function, called the log-likelihood function, rather than with the likelihood function itself. Thus, the log-likelihood function is given by

$$L(\Theta) = \sum_{i=1}^N \log p(x_i|\Theta) \quad (6.5)$$

Maximum likelihood parameter estimation is a general and powerful method for estimating the parameters of a stochastic process from a set of observed data [RWS4]. It is based on finding the model parameters, $\Theta$, which are the most likely to have produced the set of observed samples. This is achieved by finding the parameters which maximize the model’s likelihood function over a set of observations. A necessary condition for the maximum likelihood parameter estimates, is that they satisfy the likelihood equations

$$\nabla_{\Theta} L(\hat{\Theta}) = 0 \quad (6.6)$$

There is a unique strongly consistent solution of the likelihood equations (6.6), and this solution at least locally maximizes the log-likelihood function and is asymptotically normally [RWS4]. Consistent in the usual sense means converging with probability approaching 1 to the true parameters as the sample size approaches infinity; strongly consistent means having the same limit with probability 1. Unfortunately, attempting to solve (6.6) directly for the GMM parameters, $\Theta = \{\alpha_i, \mu_i, \Sigma_i\}, \quad i = 1, \cdots, m$ does not yield closed form solutions.

The maximum likelihood GMM parameter estimates can be found using an iterative parameter estimation procedure, which is a special case of Expectation-Maximization (EM) algorithm [DLR77]. The widespread use of the EM algorithm stems from the facts that it guarantees a non-decreasing likelihood function after each iteration [DLR77] and that it provides a general yet powerful framework capable of dealing with many complicated estimation
6.3. GMM PARAMETER ESTIMATION

problems [RW84]. The basic idea of the EM algorithm is, beginning with an initial model \( \Theta \), to estimate a new model \( \hat{\Theta} \), such that \( L(\hat{\Theta}) \geq L(\Theta) \). To accomplish it, an auxiliary function is used called \textit{Baum’s auxiliary function} [BPSW70] (see discussion below). The new model then becomes the current model and the process is repeated until some convergence threshold is reached. In the following, the maximum likelihood estimation equations for GMM parameters are derived using Baum’s auxiliary function [BPSW70].

Let the GMM speaker model consists of \( m \) acoustic classes and let \( X = \{x_1, x_2, \ldots, x_N\} \) be the sequence of observation vectors and \( \Omega = \{\omega_i, \ldots, \omega_N\} \) with \( i_t \in [1, m] \) be a particular sequence of acoustic classes (or states in HMM context) which produced \( X \). Both the observed data (or \textit{incomplete data}) \( X \) and the \textit{hidden data} \( \Omega \) are often called \textit{complete data}. Considering the independence between the observations and the acoustic classes, the joint pdf of \( X \) and \( \Omega \) is given by

\[
p(X, \Omega|\Theta) = \prod_{t=1}^{N} \alpha_{i_t} p_k(x_t|\theta_{i_t}) \quad (6.7)
\]

The likelihood function over \( X \) can then be written as,

\[
p(X|\Theta) = \sum_{\Omega} p(X, \Omega|\Theta) \quad (6.8)
\]

where the notation \( \sum_{\Omega} \) denotes the sum over all possible class sequences. As is discussed above, the objective is to find a new set of model parameters \( \hat{\Theta} \) which increases the likelihood function in (6.8) for a given current model \( \Theta \):

\[
p(X|\hat{\Theta}) \geq p(X|\Theta) \quad (6.9)
\]

The auxiliary function method of Baum et al. [BPSW70] provides means of \textit{iteratively increasing the likelihood function of the observed data by maximizing a function of complete data}. The auxiliary function is given by

\[
Q(\Theta|\hat{\Theta}) = \sum_{\Omega} p(X, \Omega|\Theta) \log p(X, \Omega|\hat{\Theta}) \quad (6.10)
\]

It can be shown [Lip82] that the strictly concave function exhibits the property where finding \( \hat{\Theta} \) that maximizes \( Q(\Theta|\hat{\Theta}) \), results in an increase in the observed data likelihood: \( p(X|\hat{\Theta}) \geq p(X|\Theta) \).

Substituting (6.7) using the new model \( \hat{\Theta} \) in (6.10) gives,

\[
Q(\Theta|\hat{\Theta}) = \sum_{\Omega} p(X, \Omega|\Theta) \sum_{t=1}^{N} \log \left[ \alpha_{i_t} \hat{p}_{i_t}(x_t|\hat{\theta}_{i_t}) \right] \quad (6.11)
\]
where \( \hat{p}_{it} \) is the new mixture weight and \( \hat{\theta}_{it} \) is a component density using the new mean and covariance model parameters. Using a delta function defined as

\[
\delta_t(i, \Omega) = \begin{cases} 
1 & i_t = i \\
0 & \text{otherwise}
\end{cases}
\]

we can write

\[
\log \left[ \alpha_k \hat{p}_{it}(x_t|\hat{\theta}_{it}) \right] = \sum_{i=1}^{m} \log \left[ \alpha_k \hat{p}_{it}(x_t|\hat{\theta}_{it}) \right] \delta_t(i, \Omega)
\]

thus, eliminating the dependence of the new model parameters on the hidden state variables \( i_t \). Substituting (6.13) in (6.11) we obtain,

\[
Q(\Theta|\hat{\Theta}) = \sum_{i=1}^{N} \sum_{\Omega} p(X, \Omega|\Theta) \sum_{i=1}^{m} \log \left[ \alpha_k \hat{p}_{it}(x_t|\hat{\theta}_{it}) \right] \delta_t(i, \Omega)
\]

\[
= \sum_{i=1}^{N} \sum_{\Omega} \log \left[ \alpha_k \hat{p}(x_t|\hat{\theta}_{it}) \right] \xi_t(i)
\]

where

\[
\xi_t(i) = \sum_{\Omega} \delta_t(i, \Omega)p(X, \Omega|\Theta)
\]

\[
= P(X = x_t, \omega_i = \omega_i|\theta_i)
\]

Using Baye’s formula [Pap91], \( \xi_t(i) \) can be written as

\[
\xi_t(i) = P(X = x_t|\Theta) P(\omega_i = \omega_i|X = x_t, \theta_i)
\]

where

\[
P(\omega_i = \omega_i|X = x_t, \theta_i) = \frac{P(X = x_t, i_t = i|\theta_i)}{P(X = x_t|\Theta)}
\]

\[
= \frac{\alpha_k p_{it}(x_t|\theta_i)}{\sum_{j=1}^{m} \alpha_j p_{it}(x_t|\theta_j)}
\]

\[
= \frac{\alpha_i \Sigma_i^{-1/2} \exp \left[ -\frac{1}{2} (x_t - \mu_i)^T \Sigma_i^{-1} (x_t - \mu_i) \right]}{\sum_{j=1}^{m} \alpha_j \Sigma_j^{-1/2} \exp \left[ -\frac{1}{2} (x_t - \mu_j)^T \Sigma_j^{-1} (x_t - \mu_j) \right]}
\]

where the latest expression in (6.17) is the conditional probability that a given observation vector \( x_t \) belongs to each one of the acoustic classes \( \omega_i \) of the GMM and will be denoted in following by \( P(\omega_i|x_t, \theta_i) \).

### 6.3.2 Parameter estimation equations

The parameters estimates that increase the likelihood function are now obtained by maximizing (6.14) with respect to each of the new model parameters \( \hat{\Theta} = \{ \hat{\alpha}_i, \hat{\rho}_i, \hat{\Sigma}_i \} \), \( i = 1, \cdots, m \).
6.3. GMM PARAMETER ESTIMATION

6.3.2.1 Mixture weights

The mixture weights are obtained by maximizing \( Q(\Theta|\hat{\Theta}) \) with respect to \( \hat{\alpha}_i \) under the constraint that \( \sum_{i=1}^{m} \hat{\alpha}_i = 1 \) (to ensure \( \hat{\alpha}_i \) are valid probabilities). The constrained optimization of \( Q(\Theta|\hat{\Theta}) \) will be solved by the classical method of Lagrange multipliers. Denoted by \( \lambda \) the Lagrange multiplier the optimization equation is given by

\[
\nabla_{\hat{\alpha}_i} \left[ Q(\Theta|\hat{\Theta}) - \lambda \sum_{i=1}^{m} \hat{\alpha}_i \right] = 0
\]

or

\[
\frac{\sum_{t=1}^{N} \xi_t(i)}{\hat{\alpha}_i} - \lambda = 0
\] (6.19)

Multiplying (6.19) by \( \hat{\alpha}_i \) and summing over \( j \) (after changing the variable of the sum) gives

\[
\lambda = \sum_{t=1}^{N} \sum_{j=1}^{m} \xi_t(j)
\] (6.20)

Substituting (6.20) in (6.19) yields

\[
\hat{\alpha}_i = \frac{\sum_{t=1}^{N} \xi_t(i)}{\sum_{t=1}^{N} \sum_{j=1}^{m} \xi_t(j)} = \frac{P(\mathbf{X}|\Theta) \sum_{t=1}^{N} P(\omega_i|x_t, \theta_i)}{P(\mathbf{X}|\Theta) \sum_{t=1}^{N} \sum_{j=1}^{m} P(\omega_j|x_i, \theta_j)}
\]

\[
= \frac{1}{N} \sum_{t=1}^{N} P(\omega_i|x_t, \theta_i)
\] (6.21)

where in (6.21) we have used the property : \( \sum_{j=1}^{m} P(\omega_j|x_i, \theta_j) = 1 \).

6.3.2.2 Component density means

Taking the gradient of (6.14) with respect to a density mean vector \( \hat{\mu}_i \) gives

\[
\nabla_{\hat{\mu}_i} Q(\Theta|\hat{\Theta}) = \sum_{t=1}^{N} \xi_t(i) \frac{\partial}{\partial \hat{\mu}_i} log p_i(x_t|\hat{\Theta})
\]

\[
= \sum_{t=1}^{N} \xi_t(i) \frac{\partial}{\partial \hat{\mu}_i} \left[ -\frac{1}{2}(x_t - \hat{\mu}_i)^T \Sigma_i^{-1}(x_t - \hat{\mu}_i) \right]
\] (6.22)
Using the identity $\frac{\partial}{\partial \alpha} [\alpha^T \Gamma \alpha] = 2\Gamma \alpha$, where $\alpha$ is a $p$-by-$1$ vector and $\Gamma$ is a $p$-by-$p$ matrix and setting (6.22) to zero we get

$$\hat{\mu}_t = \frac{\sum_{i=1}^{N} \xi_t(i)x_t}{\sum_{i=1}^{N} \xi_t(i)} \quad (6.23)$$

The final mean vector estimation formula is obtained by substituting $\xi_t(i)$ from (6.16) in (6.23) and cancelling like terms,

$$\hat{\mu}_t = \frac{\sum_{i=1}^{N} P(\omega_t|x_t, \theta_t)x_t}{\sum_{i=1}^{N} P(\omega_t|x_t, \theta_t)} \quad (6.24)$$

### 6.3.2.3 Component density covariance matrices

The gradient of (6.14) with respect to a covariance matrix $\Sigma_t$ yields

$$\nabla_{\Sigma_t} Q(\Theta | \hat{\Theta}) = \sum_{i=1}^{N} \xi_t(i) \frac{\partial}{\partial \Sigma_t} \log p_t(x_t | \hat{\theta})$$

$$= \sum_{i=1}^{N} \xi_t(i) \frac{\partial}{\partial \Sigma_t} \left[ -\frac{1}{2} (x_t - \hat{\mu}_t)^T \Sigma_t^{-1} (x_t - \hat{\mu}_t) - \frac{1}{2} \log |\Sigma_t| \right] \quad (6.25)$$

Using the identities

$$\frac{\partial}{\partial \Gamma} [\alpha^T \Gamma^{-1} \alpha] = -\alpha \alpha^T \Gamma^{-1} (\Gamma^{-1})^T$$

and

$$\frac{\partial}{\partial \Gamma} \log |\Gamma| = \Gamma^{-1}$$

where $\alpha$ is a $p$-by-$1$ vector and $\Gamma$ is a $p$-by-$p$ non-singular matrix, and setting (6.25) to zero we get

$$\hat{\Sigma}_t = \frac{\sum_{i=1}^{N} \xi_t(i)(x_t - \hat{\mu}_t)(x_t - \hat{\mu}_t)^T}{\sum_{i=1}^{N} \xi_t(i)} \quad (6.26)$$

Finally, using the equation for $\xi_t(i)$, the covariance matrix estimation formula is given by

$$\hat{\Sigma}_t = \frac{\sum_{i=1}^{N} P(\omega_t|x_t, \theta_t)(x_t - \hat{\mu}_t)(x_t - \hat{\mu}_t)^T}{\sum_{i=1}^{N} P(\omega_t|x_t, \theta_t)} \quad (6.27)$$
6.3.3 The EM algorithm

The Equations (6.17), (6.21), (6.24) and (6.27) appear to be rather formidable but their interpretation is actually quite simple. In the extreme case where \( P(\omega_i|\mathbf{x}_t, \theta_i) \) is one when \( \mathbf{x}_t \) is from class \( \omega_i \) and zero otherwise, \( \hat{\alpha}_i \), is the fraction of samples from \( \omega_i \), \( \hat{\mu}_i \) is the mean of those samples, and \( \Sigma_i \) is the corresponding sample covariance matrix [DH73]. More generally, \( P(\omega_i|\mathbf{x}_t, \theta_i) \) is between zero and one, and all of the samples play some role in the estimates [DH73]. In other words, GMM provides a soft classification between the several components of the mixture density. Our primary motivation for using the GMM is its capability for this soft classification.

These equations are the basis of the EM algorithm for iteratively estimating the parameters of a GMM. The algorithm consists of the following steps:

**Initialization** : Initialize the model parameters \( \Theta^{(0)} \).

**E-step** : Estimate the conditional probabilities using current model parameters \( \Theta^{(m)} \) by

\[
P(\omega_i|\mathbf{x}_t, \theta_i^m) = \frac{\alpha_i |\Sigma_i|^{-1/2} \exp \left[ -\frac{1}{2} (\mathbf{x}_t - \mu_i)^T \Sigma_i^{-1} (\mathbf{x}_t - \mu_i) \right]}{\sum_{j=1}^{m} \alpha_j |\Sigma_j|^{-1/2} \exp \left[ -\frac{1}{2} (\mathbf{x}_t - \mu_j)^T \Sigma_j^{-1} (\mathbf{x}_t - \mu_j) \right]}
\]

(6.28)

for \( i = 1, \ldots, m \).

**M-step** : Using the conditional probabilities estimated above, estimate the new model \( \Theta^{(m+1)} \) using the equations:

\[
\hat{\alpha}_i = \frac{1}{N} \sum_{t=1}^{N} P(\omega_i|\mathbf{x}_t, \theta_i^m)
\]

\[
\hat{\mu}_i = \frac{\sum_{t=1}^{N} P(\omega_i|\mathbf{x}_t, \theta_i^m) \mathbf{x}_t}{\sum_{t=1}^{N} P(\omega_i|\mathbf{x}_t, \theta_i^m)}
\]

(6.29)

\[
\hat{\Sigma}_i = \frac{\sum_{t=1}^{N} P(\omega_i|\mathbf{x}_t, \theta_i^m)(\mathbf{x}_t - \hat{\mu}_i)(\mathbf{x}_t - \hat{\mu}_i)^T}{\sum_{t=1}^{N} P(\omega_i|\mathbf{x}_t, \theta_i^m)}
\]

**Iterate** : Replace current model parameters with the new model parameters : \( \Theta^m \rightarrow \Theta^{m+1} \) and iterate between E and M steps until likelihood function stops increasing.

The EM algorithm iteratively increases the likelihood of the model parameters by successive maximizations of an intermediate quantity which, in the case of a GMM, is entirely defined by the conditional probabilities of (6.28).
An important implementation issue associated with the EM algorithm is its initialization. In the most general case, the EM algorithm is only guaranteed to converge towards a local maximizer of the likelihood function [DLR77]. In practice, the initialization of the EM algorithm affects its convergence rate but can also modify the final result [RW84]. For GMM speakers model with diagonal covariance matrices, it was found in [RR95], [Rey92] that the initialization of the EM algorithm only has a small influence. In this thesis, the GMM parameters are initialized by use of a standard full-search VQ procedure [RJ93]: the weight, mean vector and covariance matrix of the components are estimated independently using the clusters obtained by VQ of the source vectors \( \{ x_t \} \). The iterations of the EM algorithm are terminated when the relative increase of the log-likelihood function becomes too small (the threshold is set at about \( 10^{-3} \)).

Another concern in the implementation of the EM algorithm is the problem of small-variance components. It is easily verified that the likelihood functions has a diverging behaviour when any one of the covariance matrix approaches the null matrix [DH73]. This means that small events, such as the presence of a sufficient number of quasi-identical envelopes, can disturb the convergence of the whole model. The solutions used to counter this effect are analogous to that used in [RR95]. When using GMMs with diagonal covariance matrices, the diagonal variance components are constrained to be greater than minimal thresholds. The values of these thresholds are chosen 50 times smaller than the diagonal elements of the covariance matrix of the whole data. When working with full covariance matrices, a constant perturbation is systematically added to all the diagonal elements after each reestimation of the covariance matrices. The value of this perturbation is equal to the smallest of the thresholds used in the diagonal case.

### 6.4 Conversion functions

In this, as well as in the next section, we consider that the available data consists of two sets of paired spectral vectors \( x_t \) and \( y_t \) corresponding respectively to the spectral envelopes of the source and the target speakers. Each spectral vector \( x_t \) (or \( y_t \)) is a \( p \)-dimensional vector of discrete MFCC that represent the spectral envelope. The two sets of vectors \( \{ x_t, t = 1, \ldots, n \} \) and \( \{ y_t, t = 1, \ldots, n \} \) have the same length \( n \) and are supposed to be time-aligned using a classic Dynamic Time Alignment algorithm [RJ93]. What is desired is thus a function \( F() \) such that the transformed envelope \( F(x_t) \) best matches the target envelope \( y_t \), for all envelopes in the learning set \( (t = 1, \ldots, n) \). The mapping codebook approach of Abe et al.[ANSK88] reduces this problem to a lower dimensional problem by specifying the conversion function for a reduced set of codebook vectors obtained by applying a VQ procedure to the source vectors.
\{\mathbf{x}_i\}. We propose to use a finer description of the statistical distribution of the source vectors under the form of a continuous probability distribution provided by a Gaussian mixture model.

In what follows, we assume that a Gaussian mixture model \((\alpha_i, \mu_i, \Sigma_i \text{ for } i = 1, \ldots, m)\) was fitted to the source vectors \(\{\mathbf{x}_i, t = 1, \ldots, n\}\). Recall that the GMM also defines underlying classes that correspond to each Gaussian component. The fit between a source vector \(\mathbf{x}_i\) at each one of these classes can be evaluated in a probabilistic way by the computation of the conditional probabilities given by (6.28). Note, that after the convergence of the EM algorithm the estimated model \(\Theta\) is held constant. Thus, it is more convenient to write the conditional probabilities as: \(P(\omega_i|\mathbf{x}_i)\) rather \(P(\omega_i|\mathbf{x}_i, \theta_i)\).

We now turn to the problem of finding a conversion function \(\mathcal{F}\) that transforms each vector of the source data set \(\{\mathbf{x}_i\}\) into its counterpart in the target data set \(\{\mathbf{y}_i\}\). The following parametric form [SCM95][SCM96] is assumed for the conversion function \(\mathcal{F}\):

\[
\mathcal{F}(\mathbf{x}_i) = \sum_{i=1}^{m} P(\omega_i|\mathbf{x}_i) \left[ \nu_i + \Gamma_i \Sigma_i^{-1}(\mathbf{x}_i - \mu_i) \right]
\]  

(6.30)

The conversion function \(\mathcal{F}\) is entirely defined by the \(p\)-dimensional vectors \(\nu_i\) and the \(p\)-by-\(p\) matrices \(\Gamma_i\), for \(i = 1, \ldots, m\) (where \(m\) is the number of mixture components).

This form was chosen by analogy with the result obtained in the limit-case where the GMM is reduced to a single class. Indeed, if it is assumed that the source vectors \(\mathbf{x}_i\) follow a Gaussian distribution \(N(\mathbf{x}; \mu, \Sigma)\) and that the source and target vectors are jointly Gaussian, the minimum mean square error estimate of the target vector is given by [Kay93]

\[
E[\mathbf{y}|\mathbf{x} = \mathbf{x}_i] = \nu + \Gamma \Sigma^{-1}(\mathbf{x}_i - \mu),
\]

(6.31)

where \(E[\cdot]\) denotes expectation, and \(\nu\) and \(\Gamma\) are respectively the mean target vector

\[
\nu = E[\mathbf{y}],
\]

and the cross-covariance matrix of the source and target vectors

\[
\Gamma = E[(\mathbf{y} - \nu)(\mathbf{x} - \mu)^T],
\]

where always the superscript \(T\) denotes transposition [CC80].

In the jointly Gaussian case, the optimal conversion function (in the mnse sense) is thus a simple linear transformation given by (6.31). It was decided to extend this result to the case of the GMM by weighting terms that are analogous to the Gaussian conditional expectation (terms between brackets in (6.30)) by the conditional probabilities that the vector \(\mathbf{x}_i\) belongs to different classes \(\omega_i\). Although the conversion function of (6.30) is no longer supported by
a proper statistical model of the source and target vectors, it is useful to keep in mind the interpretation of the parameters $\nu$ and $\Gamma$ in the mono Gaussian case.

The parameters of the conversion function are computed by least squares optimization on the learning data so as to minimize the total squared conversion error

$$\epsilon = \sum_{i=1}^{n} ||y_i - \mathcal{F}(x_i)||^2$$

As the spectral parameters used in this paper are basically cepstral coefficients, $\epsilon$ can also be interpreted as the total quadratic log-spectral distortion between the converted and the target envelopes. A nice feature of (6.30) is the fact that if the source and the target envelopes were identical, the obtained conversion function would be transparent: if $y_i = x_i$, the minimum value of $\epsilon$ is obtained when $\Gamma_i = \Sigma_i$ and $\nu_i = \mu_i$ (for $i = 1, \ldots, n$) and is null.

We distinguish three particular types of conversion functions derived from (6.30):

**Full conversion** This first type simply corresponds to the general case of (6.30) where no constraints are applied either to the parameters of the GMM or to those of the conversion function.

**Diagonal conversion** The use of GMMs with diagonal covariance matrices is a common practice that notably reduces the computation load associated with this kind of model [TSR92] [RR95]. In the case of cepstral parameters, this modification is believe to be appropriate since the correlation between distinct cepstral coefficients is very weak [RJ93]. In our case, the computation load associated with the training of the conversion function is strongly reduced when both the covariance matrices of the GMM $\Sigma_i$ and the conversion matrices $\Gamma_i$ are constrained to be diagonal. This simplification is due to the fact that when $\Sigma_i$ and $\Gamma_i$ are diagonal, it is easily seen from (6.30) that the total conversion error can be separated along each coordinate of the vectors as

$$\epsilon = \sum_{i=1}^{n} \sum_{k=1}^{p} ||y_i^{(k)} - \mathcal{F}(x_i)^{(k)}||^2,$$

where the superscript $(k)$ denotes the $k$-th coordinate of a vector. The optimization problem is thus separated into $p$ independent scalar optimization problems. The term diagonal conversion refers to the case where the matrices $\Sigma_i$ and $\Gamma_i$ are diagonal.

**VQ-type conversion** If we omit the correction term that depends on the difference between the source vector $x_i$ and the mean of the GMM component $\mu_i$ in (6.30), the conversion function is reduced to

$$\mathcal{F}(x_i) = \sum_{i=1}^{n} P(\omega_i|x_i)\nu_i$$

(6.34)
6.4. CONVERSION FUNCTIONS

This last form of the conversion function is of the type used by Abe et al. [ANSK88] in the mapping codebook approach in the sense that the variability of the transformed spectral envelope is strongly restricted. However, the weighting of the conversion vectors \( \nu_i \) by the conditional probabilities provides a natural way of interpolating the converted spectral envelopes: the envelopes are restricted to the various interpolation paths between the discrete set of vectors \( \nu_i \) rather than just to the vectors \( \nu_i \) themselves. This conversion function will be referred to as VQ-type conversion and will be used for comparison purposes in the present discussion. Note that as a first consequence of the reduced variability of the converted envelopes, the VQ-type conversion is not transparent in the case where the source and target envelopes are identical.

6.4.1 Optimization of the conversion function

For the sake of clarity, we should simply denote by \( p_t(i) \) the conditional probability \( P(\omega_i|x_t) \) that \( x_t \) belongs to class \( \omega_i \).

6.4.1.1 Full conversion

Due to the linear nature of the conversion function given by (6.30), the optimization of its parameters is equivalent to the resolution of the following set of linear equations in the least-squares sense

\[
y_t = \sum_{i=1}^{m} p_t(i) \left[ \nu_i + \Gamma_i \Sigma_i^{-1} (x_t - \mu_i) \right]
\]

(6.35)

for all \( t = (1, \ldots, n) \). It is easily verified that these equations can be gathered into a single matricial formulation as

\[
y = P \cdot \nu + D_x \cdot \Gamma
\]

(6.36)

where \( y \) is a \( n \)-by-\( p \) matrix that contains the target spectral vectors ordered the following way

\[
y = [y_1; \cdots; y_n]^T,
\]

\( P \) is a \( n \)-by-\( m \) matrix that features the conditional probabilities \( p_t(i) \):

\[
P = \begin{bmatrix}
p_1(1) & p_1(2) & \cdots & p_1(m) \\
p_2(1) & p_2(2) & \cdots & p_2(m) \\
\vdots & \vdots & & \vdots \\
p_n(1) & p_n(2) & \cdots & p_n(m)
\end{bmatrix},
\]

(6.37)
\( \mathbf{D}_x \) is a \( n \)-by-\( p \) matrix that depends on the conditional probabilities, the source vectors and the parameters of the GMM defined by blocks as

\[
\mathbf{D}_x = \begin{bmatrix}
p_1(1)(x_1 - \mu_1)^T \Sigma_1^{-1}\Sigma_1^{-1} & p_1(2)(x_1 - \mu_2)^T \Sigma_2^{-1}\Sigma_2^{-1} & \cdots \\
p_2(1)(x_2 - \mu_1)^T \Sigma_1^{-1}\Sigma_1^{-1} & p_2(2)(x_2 - \mu_2)^T \Sigma_2^{-1}\Sigma_2^{-1} & \cdots \\
\vdots & \vdots & \ddots \\
p_n(1)(x_n - \mu_1)^T \Sigma_1^{-1}\Sigma_1^{-1} & p_n(2)(x_n - \mu_2)^T \Sigma_2^{-1}\Sigma_2^{-1} & \cdots \\
\end{bmatrix}
\]

(6.38)

and the two matrices

\[
\nu = \begin{bmatrix} \nu_1 \nu_2 \cdots \nu_m \end{bmatrix}^T
\]

and

\[
\Gamma = \begin{bmatrix} \Gamma_1 \Gamma_2 \cdots \Gamma_m \end{bmatrix}^T
\]

are the unknown parameters of the conversion function of dimension \( m \)-by-\( p \) and \( (m \ p) \)-by-\( p \) respectively.

The form of (6.36) is that of a standard least-squares problem whose solution is classically given by the normal equations [LH74]

\[
\begin{bmatrix} \mathbf{P}^T \\ \cdots \\ \mathbf{D}_x^T \end{bmatrix} \begin{bmatrix} \mathbf{P} & \mathbf{D}_x \end{bmatrix} \begin{bmatrix} \hat{\nu} \\ \hat{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^T \\ \cdots \\ \mathbf{D}_x^T \end{bmatrix} \cdot \mathbf{y}
\]

(6.39)

or

\[
\begin{bmatrix} \mathbf{P}^T \mathbf{P} & \mathbf{P}^T \mathbf{D}_x \\ \cdots & \cdots \\ \mathbf{D}_x^T \mathbf{P} & \mathbf{D}_x^T \mathbf{D}_x \end{bmatrix} \begin{bmatrix} \hat{\nu} \\ \hat{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^T \mathbf{y} \\ \cdots \\ \mathbf{D}_x^T \mathbf{y} \end{bmatrix}
\]

(6.40)

As usual, the matrix that needs to be inverted (left most matrix in (6.40)) is symmetric and positive definite so that the normal equations can advantageously be solved using the Cholesky decomposition [PTVF94]. However, when the number of component \( m \) becomes high the above matrix tends to be non positive definite because of numerical errors. The solution in this case is to use the QR decomposition method [PTVF94] which is more robust with numerical errors [PRLN92]. Note that the computation load as well as the storage requirements associated with the numerical resolution of (6.40) rapidly become problematic since the dimension of the leftmost matrix of (6.40) is \( (m + m \times p)^2 \). To illustrate this problem we recall that the
dimension of the spectral parameters is \( p = 20 \) so that a mixture of \( m = 128 \) components already necessitates the inversion of a \( 2688 \times 2688 \) matrix!

### 6.4.1.2 Diagonal conversion

As was already noted previously the optimization of the conversion function is notably simplified in the case where both the covariance matrices of the GMM \( \Sigma_i \) and the conversion matrices \( \Gamma_i \) are diagonal (diagonal conversion).

More precisely, it is possible in this case to split the optimization problem into \( p \) independent scalar minimization problems by considering each coordinate \( k \) \((k = 1, \ldots, p)\) of the vectors separately. The \( k \)-th coordinate of the overdetemined set of least squares equations of (6.35) can be written as

\[
y^{(k)} = \sum_{i=1}^{m} p_i(i) \left[ \gamma_i^{(k)} \left( \sigma_i^{(k)} \right)^{-1} (x_i^{(k)} - \mu_i^{(k)}) + \nu_i^{(k)} \right]
\]

where the superscript \((k)\) denotes the \( k \)-th coordinate for vectors, and \( \sigma_i^{(k)} \) and \( \gamma_i^{(k)} \) are the \( k \)-th diagonal elements of matrices \( \Sigma_i \) and \( \Gamma_i \). Proceeding as before yields a matricial formulation of the optimal value of the parameters analogous to (6.39)

\[
\begin{bmatrix}
    \mathbf{P}^T \mathbf{P} & \mathbf{P}^T \mathbf{D}^{(k)}_x \\
    \vdots & \vdots \\
    \mathbf{D}^{(k)}_x \mathbf{P} & \mathbf{D}^{(k)}_x \mathbf{D}^{(k)}_x \\
\end{bmatrix}
\begin{bmatrix}
    \hat{\nu}^{(k)} \\
    \vdots \\
    \hat{\gamma}^{(k)} \\
\end{bmatrix} =
\begin{bmatrix}
    \mathbf{P}^T \mathbf{y}^{(k)} \\
    \vdots \\
    \mathbf{D}^{(k)}_x \mathbf{y}^{(k)} \\
\end{bmatrix}
\]

where \( \mathbf{D}^{(k)}_x \) is a \( n \)-by-\( m \) matrix defined by

\[
\mathbf{D}^{(k)}_x =
\begin{bmatrix}
    p_1(1)\sigma_1^{(k)-1}(x_1^{(k)} - \mu_1^{(k)}) & p_1(2)\sigma_2^{(k)-1}(x_1^{(k)} - \mu_2^{(k)}) & \ldots \\
    p_2(1)\sigma_1^{(k)-1}(x_2^{(k)} - \mu_1^{(k)}) & p_2(2)\sigma_2^{(k)-1}(x_2^{(k)} - \mu_2^{(k)}) & \ldots \\
    \vdots & \vdots & \ddots \\
    p_n(1)\sigma_1^{(k)-1}(x_n^{(k)} - \mu_1^{(k)}) & p_n(2)\sigma_2^{(k)-1}(x_n^{(k)} - \mu_2^{(k)}) & \ldots \\
\end{bmatrix}
\]

and \( \mathbf{y}^{(k)} \) is 1-by-\( n \) vector defined by

\[
\mathbf{y}^{(k)} = \begin{bmatrix} y_1^{(k)} , \ldots , y_n^{(k)} \end{bmatrix}^T
\]
CHAPTER 6. STATISTICAL METHODS FOR SPEAKER MODIFICATION

Note that the matrix $P$ stays as defined in (6.37). As we only consider the $k$-th coordinate, the unknown parameters of the conversion function are reduced in (6.42) to the two 1-by-$m$ vectors

$$\nu^{(k)} = \left[ \nu_1^{(k)}, \ldots, \nu_m^{(k)} \right]^T$$

and

$$\gamma^{(k)} = \left[ \gamma_1^{(k)}, \ldots, \gamma_m^{(k)} \right]^T$$

It is recalled that (6.42) only yields the values of one coordinate of the conversion parameters (vectors $\nu_i$, and diagonal elements of matrices $\Gamma_i$). (6.42) should thus be applied for each coordinate $k$, with $k = 1, \ldots, p$ (where $p$ is the dimension of the parameter space). Note that all the matrices featured in the block-defined matrix in the left-hand part of (6.42) will need to be re-computed from one coordinate to the other except $P^TP$ which only involve the scalar terms $p_i(i)$ (conditional probability associated with vector $x_i$ and class $\omega_i$).

It is important to underline that the block defined matrix in the left-hand part of (6.42) only has $m \times m$ elements. The optimization of a diagonal conversion function is thus reduced to $p$ successive inversions of $m$-by-$m$ matrices which makes this type of conversion function very attractive compared to the general case (inversion of a $(m + mp)$-by-$(m + mp)$ matrix) when the number of component $m$ becomes large.

6.4.1.3 VQ-type conversion

The optimization of the conversion function in the case of VQ-type conversion is easily obtained as a special case of (6.42) by omitting the diagonal matrix elements ($\gamma^{(k)}$). The $k$-th coordinate of the unknown conversion vectors $\nu_i$ are given by

$$\nu^{(k)} = (P^TP)^{-1}P^Ty^{(k)} \quad (6.45)$$

6.5 Training of the conversion function

This section is concerned with the learning of the spectral conversion function from experimental data. We consider that the available data consists of two sets of paired spectral vectors $x_i$ and $y_i$ corresponding respectively to the spectral envelopes of the source and the target speakers. Each spectral vector $x_i$ (or $y_i$) is a $p$-dimensional vector of discrete Mel-Frequency Cepstrum Coefficients, MFCC, that represent the spectral envelope. The computation of these spectral vectors have been developed in Chapter 3. Some specific practical issues on spectral vectors computation when they are used to train the conversion function are discussed below.
6.5. Training of the Conversion Function

6.5.1 Practical issues on spectral vectors computation

Preliminary voice conversion tests conducted with the HNM system led us to conclude that the conversion of the noise part is a rather delicate task. In practice, the spectral envelopes associated with the noise part exhibit strong variations and the corresponding GMM components are characterized by their high variance and their large overlapping. In these conditions, the obtained conversion function is not very efficient except for the general features of the spectrum such as its average decrease with frequency. Moreover, the contribution of the noise part to the individuality of the speaker was found to be by far less important than that of the harmonic part. In the present discussion, the conversion methodology presented in the previous section is applied to the transformation of the harmonic part of the signal. As a consequence, only the voiced frames are used for training the conversion function. As a further simplification the maximum voiced frequency was fixed at a constant value of 4kHz (the sampling frequency is 16kHz).

The conversion of the noise part is simply achieved by use of two different correction filters, one for voiced frames and one for unvoiced frames (see discussion below). These correction filters, implemented as 6th order all-pole filters, model the difference between the average noise spectra of the source and target speaker. The distinction between voiced and unvoiced frames appears to be necessary because the characteristics of the noise part (especially its level) are very different in the two cases.

The aim of the spectral conversion function is thus to transform the harmonic part of speech which is supposed to extend between 0 and 4 kHz (for voiced frames). The spectral envelope corresponding to the voiced part of speech is computed from the amplitudes of the harmonics by the discrete regularized cepstrum method using a warped frequency scale (see Chapter 3). The main steps of the envelope estimation procedure are the following:

1. The amplitudes of the harmonics $a_k$ ($k = 1, \ldots, L$) determined by the HNM analysis are expressed in the log domain.

2. The frequencies of the harmonics ($k f_1$, where $f_1$ is the fundamental frequency) are converted to a Bark frequency scale using the analytical formulas reported in [ZT80]. The obtained values $w_k$ are normalized in order to ensure that the upper limit of the band (4 kHz) corresponds to a value of 1/2 on the normalized warped frequency axis.

3. The real cepstrum parameters $c_i$ ($i = 1, \ldots, p$) that represent the envelope $S_c(w)$ are obtained by minimizing a least squares criterion in the log-spectral domain [CLM95].

The obtained cepstral parameters are similar to the usual MFCC except for the fact that
they are obtained from the minimization of a discrete set of frequency points. Such parameters are generally referred to as discrete MFCCs [GR90] [GR91] and are known to provide a better envelope fit (at the specified frequency points) than the usual "continuous" spectral modelization methods [CLM95].

When the order $p$ of the cepstrum coefficients is greater than 15, the model of the cepstral envelope obtained by the above method provides a high quality representation of the signal. In most cases, the synthetic signals obtained by use of the envelope representation or by a direct synthesis from the HNM parameters are indistinguishable. For lower values of the cepstrum order however some smoothing of the envelope occurs. In order to maintain an accurate description of the characteristics of the spectral envelope, an order of $p = 20$ was systematically used throughout our voice conversion experimentation.

In the present study, the first cepstrum coefficient $c_0$ was omitted in order to get rid of the energy normalization problem. In practice, it is not convenient to include $c_0$ in the training parameters because the classification achieved by the GMM would depend almost exclusively upon this parameter. As a consequence, the value of $c_0$ is not modified and the spectral conversion is normalized in the sense that it conserves the average level of the spectrum (in the log-spectral domain). The spectral parameters are thus $p$ dimensional vectors which contain the discrete MFCC coefficients $c_1, c_2, \ldots, c_p$.

### 6.5.2 The main steps for the transformation function learning

As was said above, only the voiced frames are used to estimate the spectral conversion function since we focus on the transformation of the harmonic part of speech. At the end of the source and the target training utterances analysis, two sets of spectral vectors are obtained which are denoted by $\{x_i\}_t^N$ and $\{y_i\}_t^M$ for the source and target speaker respectively. Note that the number of source vectors, $N$, differs (in general) from that of the target ones, $M$.

The first step for the spectral conversion consists of aligning the two sets of the spectral vectors. This is accomplished by using a Dynamic Time Alignment, (DTW), algorithm [RJ93]. For this purpose the analysis is performed asynchronously with a constant frame rate of $10 \text{msec}$. Note that the asynchronous mode is used only for the data needed to train the conversion function. For the subsequent voice conversions, the HNM analysis is carried pitch-synchronously because it allows higher quality prosodic modifications. The phonetic boundaries have been used as anchor points for the dynamic time warping procedure. After the DTW procedure, the two sets of vectors $\{x_i, \ t = 1, \ldots, n\}$ and $\{y_i, \ t = 1, \ldots, n\}$ have the same length $n$ and are supposed to be time-aligned. What is desired is thus a function $F()$ such that the transformed envelope $F(x_i)$ best matches the target envelope $y_i$, for all envelopes in the learning set ($t = 1, \ldots, n$).
The second step of the spectral conversion consists of estimating the GMM parameters associated to the source cepstral vectors. As explained in the previous section, the GMM parameters are initialized by use of a standard full-search VQ procedure [RJ93]. Then, the EM algorithm is used to estimate the GMM parameters. While each iteration of the EM algorithm is guaranteed to increase the likelihood function, it is useful to determine the number of iterations that are required for the likelihood function to have effectively converged. Fig.6.1 shows the typical behaviour of the log likelihood function for several iterations of the EM algorithm. The model has 128 Gaussian components and was trained using 3.5 minutes of speech (21000 vectors). The covariance matrices, for this example, were constrained to be diagonals. Note that after an initial large increase in the likelihood function, each successive iteration only increases the likelihood function slightly. After 18 iterations it has effectively converged.

The last step of the conversion function training procedure is to estimate the vectors, \( \mu_i \) and the matrices \( \Gamma_i \) of the conversion function, for \( i = 1, \ldots , m \), where \( m \) is the number of components in the GMM. This is accomplished by solving the equation (6.40) in the case of full conversion function, or the equation (6.42) in the case of diagonal conversion function or finally, the equation (6.45) in the case of VQ-type conversion function.

In order to minimize the influence of local errors in the time alignment path between the two speakers, it is proposed to learn the conversion function incrementally. In a first iteration, the data from the source speaker and the target speaker are aligned by a standard DTW

![Figure 6.1: Typical behaviour of the log-likelihood function for several iterations of EM algorithm for GMM training.](image)
algorithm. This alignment is used to obtain the initial parameters of the conversion function. In subsequent iterations, the DTW procedure is applied to the converted data and the target data in order to refine the alignment path.

In Fig.6.2 the system for the conversion function training is depicted.

![Diagram](image)

Figure 6.2: Block diagram of the learning step. $x_t$ represents the source spectral vectors and $y_t$ the target spectral vectors.

### 6.5.3 Corrective filters for the conversion of the noise part

In the previous section we have seen the main steps for the training of the conversion function in order to convert the harmonic part. In this section the problem of the noise part conversion is addressed.

Earlier it was indicated that if the noise part is included in the training data for the conversion function learning, the efficiency of the conversion function is reduced as then, the GMM components are characterized by their high variance and their large overlapping. Note also that the contribution of the noise part to the individuality of a speaker is less important than that of the harmonic part. For all these reasons, it was decided to exclude the noise part from the training data used for the learning of the conversion function. For the conversion of the noise part, two corrective filters, one for the voiced frames and the other for the unvoiced frames were used. The procedure to estimate these filters is given below.

First, the training data (speech samples) are separated into voiced and unvoiced frames resulting in a voiced and an unvoiced signal. This distinction was found to be necessary as the characteristics of the noise part differ for voiced and unvoiced frames. Then, the power spectral density for each of these signal is estimated using an averaged periodogram estimator [Kay88]:

$$
\hat{P}_{av\text{per}}(f) = \frac{1}{K} \sum_{m=0}^{K-1} \hat{P}_{\text{per}}^{(m)}(f)
$$

(6.46)
where \( \hat{P}_{\text{per}}^{(m)}(f) \) is the periodogram for the \( m \)-th set of samples of the voiced or unvoiced signal \( s(t) \), denoted by \( s_m(t) \). The \( \hat{P}_{\text{per}}^{(m)}(f) \) is defined by [Kay88]:

\[
\hat{P}_{\text{per}}^{(m)}(f) = \frac{1}{T} \sum_{t=0}^{T-1} s_m(t) e^{-j2\pi ft} \tag{6.47}
\]

This procedure is repeated for both the source and the target speaker. Let \( \hat{P}_{\text{per}}^{(0)}(f) \) and \( \hat{P}_{\text{per}}^{(1)}(f) \) denote respectively the power spectral density estimates of the voiced and unvoiced signals of the source speaker, and \( \hat{P}_{\text{per}}^{(0)}(f) \) and \( \hat{P}_{\text{per}}^{(1)}(f) \) denote the corresponding estimates for the target speaker, then the subtraction of these estimates gives the average difference between the source and target speaker for voiced and unvoiced frames

\[
\hat{D}_v(f) = \hat{P}_{\text{per}}^{(0)}(f) - \hat{P}_{\text{per}}^{(1)}(f) \\
\hat{D}_u(f) = \hat{P}_{\text{per}}^{(0)}(f) - \hat{P}_{\text{per}}^{(1)}(f) \tag{6.48}
\]

Lastly, two autocorrelation sequences, the first one for the voiced frames and the second for the unvoiced frames, are estimated by taking the inverse Fourier transform of the \( \hat{D}_v(f) \) and \( \hat{D}_u(f) \). From the two estimated autocorrelation sequences the coefficients of the corrective filters (reflection coefficients) are obtained using a classic \( LPC \) algorithm [Kay88].

The overall procedure for the estimation of the corrective filters for the voiced and unvoiced frames is depicted in Fig.6.3

![Figure 6.3: Block diagram of the corrective filters estimation for the conversion of the noise part.](image)

### 6.6 The voice conversion system

Given the spectral conversion function which transforms the source spectral vectors to target ones, it is possible to convert any new source spectral vector. A source test signal is analyzed in a pitch-synchronous way from the HNM and then, for each (voiced) spectral vector, the conditional probability is computed using (6.17). Next, the conversion function (equation (6.30)) is applied to these spectral vectors to obtain the transformed ones. Note that as the
first cepstral coefficient \( c(0) \) was omitted from the training parameters it is not transformed at this stage.

In order to complete the voice conversion, prosodic modifications also have to be done. The HNM is an efficient and simple tool for modifying the fundamental frequency and the articulation rhythm of speech. In the present voice conversion experimentation, only the average pitch and speaking rate of the target speaker are matched. However, in cases where the same sentence uttered by the two speakers is available, the HNM can also be used to impose the pitch and time contours of the target speaker to the converted speech signal. The block diagram of the conversion step is depicted in the Fig.6.4.

![Block diagram of the conversion step](image)

Figure 6.4: Block diagram of the conversion step. \( t_a^i \): analysis time-instants, \( t_s^i \): synthesis time-instants.

In Fig.6.4, the conversion of the noise part has also been included; the noise part is only roughly modified by use of two separate time invariant filters, one for voiced frames and the other for unvoiced frames. This simple transformation scheme was found to be sufficient since the exact frequency content of the noise part does not seem to contribute significantly to speaker individuality. An example of these kinds of time invariant filters will be showed in the next section.

### 6.7 Results and discussion

To evaluate the performance of the proposed conversion functions, two databases were used. The first database was created at ENST- Télécom Paris, which was only used for statistical tests of the proposed conversion functions. As there is no theoretical way to estimate the number of mixture components needed to model the acoustic space of a speaker adequately and also to estimate the parameters of the conversion functions efficiently, it was informative to relate the performance of the conversion functions with respect to the number of components and the amount of training data. Choosing too few mixture components can produce a speaker model
which does not accurately model the distinguishing characteristics of a speaker’s distribution. In this case the components are characterized by large overlapping; this leads to a converted speech signal which suffers from a muffling effect. Choosing too many components can reduce the conversion performance because there are too large a number of parameters to estimate relative to the available training data. Note also that too many components can also result in excessive computational complexity in the training step.

The following experiments examine the performance of the conversion functions (only the full-type and the diagonal-type conversion function will be examined). The database was separated into an amount of training data (4.1 minutes) and another of test data (0.3 minutes). Next the amount of training data was also separated into different amounts of training data: 0.3 min., 0.8 min., 1.7 min., 2.5 min., 3.3 min., and the total amount of the training data (4.1 min.). The full-type and the diagonal-type conversion functions were trained using 2, 4, 8, 16, 32, 64, and 128 mixture components. The obtained conversion functions were used to convert the test data from the source speaker and then, the spectral distortion to target test data was calculated. The spectral distortion that we used is given by

\[ d(\hat{c}^s, c^t) = \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{p} (\hat{c}_{in}^s - c_{in}^t)^2 \]  

(6.49)

where, \( \hat{c}^s \) denotes the transformed cepstral coefficients of the source speaker and \( c^t \) denotes the cepstral coefficients for the target speaker. Fig.6.5 shows the performance of the conversion function versus the number of GMM components and the amount of training data. The spectral distortion was normalized so that 1 represents the initial average distortion between the source and the target envelopes. Note that for the full-type conversion function with 128 components, using 0.3 and 0.8 minutes of training data, the spectral distortion after the conversion was particularly high; the same observation was done, using the full-type conversion function with 64 mixture components and 0.3 minutes of training data. To enable a better view of the spectral distortions obtained from the other cases, the above spectral distortions was set equal to 1 (the maximum distance between source and target). The results presented in Fig.6.5 clearly shows that with increased training data, the performance of the transformation functions increases. With a sufficient amount of training data (for example, 3.3 minutes) the performance increases with the number of mixture components but at a decreasing rate. This suggests that at least 3.3 minutes of training data are necessary to maintain the conversion performance both for the full-type and the diagonal-type conversion function. In this case, the suggested number of mixture components is 64 for the full-type conversion function and 128 for the diagonal-type function.

The second database was furnished by the CNET, and it was used for the conversion functions training. The database consisted of the diphones of the French language uttered
Figure 6.5: Normalized spectral distortion to target as a function of the number of components and the amount of training data in minutes. The solid line indicates using the Full-type conversion function and the dashed line using the Diagonal-type conversion function. The test data was 2000 spectral vectors (about 0.3 minutes). The first database was used.

in context by two different male speakers. Using all the database, the two corrective filters for the noise part transformation were first estimated. To do this, the power spectral density of the voiced and unvoiced frames, both for source and target speaker was estimated. The average periodogram was used to estimate the power spectral estimates. The inverse Fourier transform of the difference between the source and the target power spectral density yields an autocorrelation sequence; using a classic autocorrelation approach[Kay88], a \( p \)-th order corrective AR filter was estimated; one for the voiced frames and the other for the unvoiced frames. A typical value of \( p \) was 5. In Fig.6.6 the estimated power spectral density (psd), for the source (solid line) and the target speaker (dashed line) is presented, both for voiced and unvoiced frames. In the same figure the corrective filters that were obtained are showed.

Having determined the two corrective filters, the next step was the training of the conversion function for the harmonic part. The time alignment between the source and target signals was performed on each diphone separately using a standard DTW technique and discarding the unvoiced portions of the signal. The remaining time-aligned data, consisting of approximately \( N = 27000 \) spectral vectors which corresponds to more than 4.6 minutes of speech, was split into a certain amount of training data (3.6 minutes) and a certain amount of test data (about 1 min.). To investigate the conversion performance of the proposed conversion functions as well as to compare them with the VQ-mapping conversion function proposed in[ANSK88],
6.7. RESULTS AND DISCUSSION

![Graphs showing PSE on voiced and unvoiced frames and corrective filters for noise part transformation.](image)

Figure 6.6: Corrective filters for noise part transformation. The solid line is the psd of the source speaker while the dashed line is the psd of the target speaker.

The following experiment was conducted. The full-type, the diagonal-type and the VQ-type conversion functions were trained with 2, 4, 8, 16, 32, 64, and 128 mixture components using the above amount of training data. For the sake of comparison, the VQ-mapping matrix was also computed for the same number of components. The procedure used for the computation of the VQ-mapping matrix was presented in the previous chapter. Fig. 6.7 presents the relative spectral distortion to source (left) and target (right) test data. The distortion was normalized by the initial distortion between the two speakers. The spectral distortion was calculated using the equation (6.49) where the dimension $p$ of the spectral vectors was 20. Note that a second iteration using 128 mixture components was also performed. The duplication of the 128 components in Fig. 6.7 indicates the second iteration. Based on the results shown in Fig. 6.7 we may make the following observations:

- As the number of components increases, the distortion between the converted and source vectors gradually increases in the case of Full-type and Diagonal-type conversion functions. At the same time the distortion between the converted and target vectors decreases. This is in total contrast with what is observed for the VQ-type and the VQ-mapping approach. For these two types of conversion, the distortion between the converted vector both to source and target speaker decreases!

- For the same number of components the performance of the Full-type conversion function
Figure 6.7: Normalized spectral distortion to source (left) and to target (right) as a function of the number of components. Using Full-type conversion function (solid line), Diagonal-type conversion function (dotted line), VQ-type (dashed line) and VQ-mapping (dashdot line).

- The incremental learning (using 128 components) shows the improvement in the spectral distortion achieved by the proposed conversion functions.

The results presented in Fig.6.7 clearly show the superiority of the proposed (Full-type and Diagonal-type) conversion function over the classic VQ-mapping approach.

In Fig.6.8, an example of 1 sec. frame to frame distortion is presented. The frames were constructed by concatenating the diphones of the test data. The spectral distortion is presented in dB. In Fig.6.8, it is also clear that by increasing the number of components, the spectral distortion decreases. However, one important point to note is that there are some frames where the spectral distortion does not decrease even if the number of components increases. This is especially the case of the Diagonal-type and the VQ-type conversion functions. In contrast, the Full-type conversion function aims to decrease the distortion on these frames when the number of components increases. Note that these frames correspond to the passage of one diphone to another.

To show how close the transformed spectral envelopes are to the target spectral envelopes, an example of spectral transformation in the frequency domain using the three types of the conversion functions is given in Fig.6.9. This allows us to understand the figure 6.8 better.
6.7. RESULTS AND DISCUSSION

Figure 6.8: Short-term log-spectral distance between source and target data (solid line) as well as between transformed and target data for 100 short-term frames (≈ 1sec) using 32 and 128 components and the three types of the conversion function: Full-type (dashed line), Diagonal-type (dashdot line) and VQ-type (dotted line).

In Fig.6.9(a) an example of spectral transformation using the Full-type conversion function is presented where the spectral distortion to source and to target is −2.6dB and −15dB respectively. For the same frame, spectral transformations using the Diagonal-type and the VQ-type conversion functions are plotted in Fig.6.9(b) and Fig.6.9(c) respectively. The spectral distortion achieved by the Diagonal-type was −5.2dB to source and −10dB to target. The corresponding performance of the VQ-type was: −3.7dB to source and −8.3dB to target. Note that the spectral envelopes are given up to 4000Hz, as the maximum voiced frequency for the spectral conversion experiments was fixed to 4000Hz. The important point to note is that, in contrast to DFW approach, not only the frequencies of the formants was changed but also their amplitudes.

Once the transformation functions are trained, they are stored for future transformation purposes. To test the voice conversion system, phrases uttered by the two speakers were used. The speech signals were analyzed in pitch synchronous mode using the HNM, and the extracted spectral features are then transformed by the spectral conversion functions. The average pitch and the articulation rhythm of the target speaker were used to specify the prosodic modifications. However, when the same phrase from the source and target speaker was available, the pitch and time contour of the target speaker were imposed. Using the converted spectral envelopes and the synthesis time instants obtained for the above prosodic
Figure 6.9: Example of spectral conversion using (a) Full-type, (b) Diagonal-type and (c) VQ-type conversion function: Solid line for the source envelope, dashed line for the target envelope and dashdot line for the transformed envelope.

Figure 6.10: Fig. 6.9(b).
specifications, the harmonic part of the output speech signal is obtained in a manner similar to that described in Chapter 3. The noise part was modified using the average corrective filters. Informal listening tests indicate that the proposed system produces speech signals which are free of artefacts (bubbles and other oddities) associated with the VQ technique. The obtained conversion effect is impressive and the general quality of the transformed speech signal was satisfying. However, a muffling effect was perceptible when the number of mixture components was too small. Future developments of the conversion system would include their evaluation with other sets of speech parameters (such as formant parameters) which may be best suited for voice conversion.

6.8 Conclusion

This chapter has presented a new method for the statistical learning of the correspondence between spectral parameters measured from two different speakers uttering the same text. In order to increase the robustness of the conversion, the source speaker space was described by a continuous probability density corresponding to a parametric Gaussian Mixture Model (GMM). It was shown that this "soft clustering" naturally overcomes the interpolation problem that appears with the mapping codebook approach. Moreover, the transformation function itself is also "continuous" in the sense that it does not rely on an underlying discrete set of target
envelopes. The proposed conversion function makes use of the complete description of each component of the GMM thus considering these components as complete clusters rather than as single vectors as is the case in VQ approaches. The parameters of the conversion function are determined by minimization of the total quadratic spectral distortion between the converted envelopes and the target envelopes. Compared to existing VQ based techniques, this method significantly reduces the spectral distortion between the target envelopes and the converted envelopes[BS96]. A last improvement was brought by what we call "incremental learning". It was based on the simple observation that a noticeable part of the residual mismatch between the transformed envelopes and the corresponding target envelopes can be attributed to local errors in the time alignment path. Some errors in the DTW procedure are unavoidable since intrinsic spectral differences between the two speakers are mixed with spectral differences due the temporal misalignment. The time alignment path can thus be improved by re-applying the DTW procedure between the converted envelopes and the target envelopes. Depending on the contents of the speech data base used for training the spectral conversion function, this refinement of the time alignment can provide a significant improvement in the conversion accuracy.

This spectral conversion method was tested on speech signals analyzed by the HNM (Harmonic plus Noise Model) [LSM93b] [SLM95]. The results obtained on a large speech database demonstrate effective high-quality transformations of the voice characteristics. The quality of the converted speech signal remains acceptable even when there are very few components in the GMM.
Chapter 7

Summary of results and future research directions

7.1 Summary of results

This thesis has introduced the HNM, a new analysis/ modification/synthesis model based on a harmonic plus noise representation of the speech signal. Using this twofold representation of the speech signal, different modification methods can be applied to each part, yielding more natural resyntheses. The harmonic part accounts for the periodic (voiced) structure of the speech, whereas the noise part accounts for the friction noise, the period-to-period fluctuations produced by the turbulences of the glottal airflow, etc. The speech spectrum (for voiced frames) is divided into two bands delimited by a maximum voiced frequency which is a time-varying parameter. In the lower band the signal is supposed to be harmonic, which means that only the fundamental frequency has to be determined. To estimate this parameter (as well as make voiced/unvoiced decisions and detect for each voiced frame the maximum voiced frequency), a method adapted to HNM, has been developed. The harmonic amplitudes and phases are estimated using fast algorithms like the algorithm of Levinson. Because of the signal’s harmonicity, the synthesis step is also simplified, as it is not necessary to track sinusoids from frame to frame: the $k$-th harmonic in frame $i$ corresponds thus to the $k$-th harmonic in frame $i + 1$. The noise part includes the unvoiced frames and the upper band (from the maximum voiced frequency up to half of sampling frequency) of the voiced frames. In voiced frames, the noise part is modelled in both time domain and frequency domain. The frequency contents of the noise part is modelled by an AR model and its time-behaviour is
imposed by a parametric time-domain envelope which modulates the noise part. In unvoiced frames, only the frequency contents of the noise part is modelled (also by an AR model).

The analysis and synthesis are performed in a pitch synchronous way which allows to develop a flexible method for pitch and time-scale modification. Time-varying modifications as well as joint pitch and time-scale modifications are quite easy to carry out using HNM. In the case of pitch modification, the re-estimation of the harmonic amplitudes and phases (at frequencies not necessarily corresponding to a pitch-harmonic in the original signal) is done using continuous frequency phase and amplitude envelopes. To obtain the phase envelopes, an unwrapped technique of the phase was developed; and for the amplitude envelopes, the method of discrete cepstrum developed in[CLM95] was adapted. An experimental evaluation using HNM for speech analysis, synthesis and modification was also presented.

The evaluation focused on re-synthesis quality, with and without speech modification. Without modification the synthetic speech signal is of very high quality, almost indistinguishable from the original. The sounds which are not very well reproduced were the plosive sounds in the case where these sounds were detected as unvoiced. For pitch modification with moderate modification factors (0.5-2) the results are satisfactory. Time-scale modifications are of high-quality, even for very large modification factors (for example 8). The results are better than those obtained with the most frequently used method for speech modification at the present, the TD-PSOLA. Note that HNM (with minor changes) has already tested by the CNET and a subjective comparison test was performed between the PSOLA and HNM using 12 sentences taken from a French prosodic corpus[BF94]; the results have shown that the HNM is preferred over the PSOLA system and that the naturalness of the unvoiced sounds (fricatives and stops) was noticeably better than with the PSOLA system[BF94].

Furthermore, another study of speech signal decomposition based on a deterministic plus stochastic approach (DSM - Deterministic plus Stochastic Model) was also presented. In this model, the deterministic part is assumed to contain only harmonically related sinusoids with linearly varying complex amplitudes. The stochastic part is obtained by modelling the error (or residual signal) between the original speech signal and the deterministic part. To model the stochastic part, an AR model was used for the frequency aspect and a piecewise linear envelope function for the time-behaviour of the residual signal. The analysis and the synthesis are performed in a pitch-synchronous way. To obtain the synthetic signal an overlap-add technique was used. The synthetic speech obtained by the DSM is indistinguishable for the original signal. However, using linearly varying complex amplitudes, the deterministic part has entirely cancelled the low frequencies which means that the noisy part of these frequencies was transferred to the deterministic part. When speech modifications are carried out, and especially in the case when some frames are repeated, the noisy part (which was passed into
7.1. **SUMMARY OF RESULTS**

the deterministic part) of the repeated frames will be correlated, resulting in a synthetic signal with a slightly metallic quality. Note that, omitting the complex slope of the amplitudes, the DSM is essentially the HNM.

Finally, a new model (a third one), called HSM - Harmonic plus Stochastic Model, for the speech signal has been proposed, based also on a harmonic plus stochastic representation of speech. This model makes use of a third order polynomial of real coefficients for the amplitudes and there are two possibilities for the phase: a first order polynomial or a third order polynomial. The HSM was found to be better than the HNM and the DSM approaches: the modelisation error is smaller than this obtained from the two above models and the noisy part at the low frequencies (on the voiced frames) remains in the stochastic part (and it is not transferred to the deterministic part as was the case for the DSM). Note that the HSM has not been tested (up to the present) for speech synthesis but it is intended to be used successfully for speech resynthesis and modification.

For speaker modification, this thesis has introduced also the Gaussian Mixture Model, GMM, as a robust representation of the acoustic space of a speaker and proposes an original linear conversion function for spectral transformations. The proposed function takes advantage of the statistical modelling provided by GMM of an acoustic space, to transform the spectral envelopes (expressed in discrete cepstrum coefficients) of the source speaker to the envelopes of the target speaker. The probabilistic framework of the GMM allows for the application of well established and powerful techniques for the GMM weight, mean vectors and covariance matrices estimation. In particular, the Expectation Maximization, EM, algorithm provides an iterative maximum likelihood estimation technique to estimate the above parameters. The EM algorithm is initialised by use of a standard full-search Vector Quantization, VQ, procedure. In contrast with VQ, the GMM provides continuous ‘smooth’ classification indexes which avoid unnatural discontinuities.

Considering two sets of time aligned spectral vectors (using a DTW technique) from the source and target speaker, and given the GMM parameters (estimated by EM), the conversion function is optimised by minimisation of the average quadratic spectral distortion between the converted envelopes and the target envelopes. Three types of conversion functions have been considered: full-type (conversion and covariance matrices are full matrices), diagonal-type (conversion and covariance matrices are restricted to be diagonal matrices) and VQ-type (the conversion matrices are omitted) conversion function. Since the optimization of the conversion function is performed on time aligned data, a noticeable part of the conversion error can be attributed to local errors in the time alignment path. By re-applying the DTW procedure between the converted envelopes and the target envelopes a refined time alignment path is obtained. This iterative procedure called as ’incremental learning’ was found to produce only
a marginal improvement in the conversion accuracy. This is believed to be due to the fact that the DTW is performed only on very short segments of speech.

A performance analysis of the three types of conversion function was completed for two different evaluation criteria: performance effects of the number of GMM components and the amount of training data and training issues. The converted spectral envelopes, approach the target spectral envelopes as the number of mixture component increases. However, and especially in the case of the full-type conversion function the increase in the number of components without a comparable increase in the training data amount, produces inaccuracies in parameter estimation. Using 128 components, the full-type conversion function reduces the spectral distortion in the first iteration by 75% and in the second iteration by 80%. For the same number of components, the diagonal-type function reduces the spectral distortion by 63% in the first iteration and by 67% in the second iteration whereas the VQ-type by 58% and 63% respectively. These results highlight the ability of the proposed conversion function to reduce the spectral distortion between source and target spectral envelopes. It was also demonstrated that the above function outperforms the classic VQ codebook mapping and the LMR approaches for spectral transformation.

Using the HNM for speech synthesis and prosodic modifications, as well as the transformation function to carried out spectral transformations, a voice conversion system was proposed in order to convert the voice of the source speaker to the voice of the target speaker. In the present voice conversion experimentation, only the average pitch and speaking rate of the target speaker were matched. However, in cases where the same sentence uttered by the two speakers was available, the pitch and time contour of the target speaker was imposed to the converted speech signal.

The implementation of such a voice conversion system consists of three major steps. The creation (or selection) of an appropriate speech database, which should characterise the two speakers, is the first and most important step. Note that almost no systematic research has been carried out on the design of such optimal database for conversion. In the present discussion two databases of male speakers have been used: one given by the CNET Lannion and the other one created at ENST Paris. The first one seems to be more complete than the second and it was finally selected for voice conversion purposes. The second database was larger than the database of the CNET Lannion and was used only for statistical tests of our conversion function (for example, consistency of the conversion function when the number of GMM components increases). The extraction of the feature vectors for source and target speaker using the HNM and the training of the conversion function is the second step of the system. The last step consists of converting the source voice: a test speech signal of the source speaker is analysed using HNM and the extracted spectral features are transformed by the
transformation function; taking into account the prosodic specifications, the synthesis time-instants are computed and then, the converted synthetic signal is obtained at the output of the synthesis system. The size of the spectral features (discrete cepstral coefficients) was fixed to 20 (sampling frequency at 16kHz).

Overall, the elapsed time for analysis/modification/synthesis and conversion without the training of the conversion function is less than 3 seconds for 1 second of speech on a SPARC-10 system. The training of the conversion function is the longest step of the conversion system: for example, for full-conversion function with 32 mixture components, about 5 hours are required for an amount of speech corpora of 3.5 minutes for each speaker (so, in all, 7 minutes). For more mixture components the elapsed time for the training increased exponentially than linearly. However, once the training is done, the transformation function is stored for future conversion purposes.

### 7.2 Future research directions

The research discussed in this thesis is by no means an exhaustive treatment of the general area of speech analysis/modification/synthesis, and speaker conversion. This section briefly discusses some of the future research directions that could extend or augment the work in this thesis. Some possible applications are also discussed.

- **Systematic test of the HSM for speech synthesis and modification**
  Preliminary results of the application of the Harmonic plus Stochastic Model for the speech signal are very satisfactory and show that the HSM is a very promising model for speech indeed. However, as it has only recently been proposed, it has not been tested extensively. One of the future areas of research that we propose is to test the HSM systematically for speech synthesis and modification.

- **Interaction between pitch modified harmonic structure and spectral envelope**
  In this thesis, the main emphasis was on speech analysis and synthesis, prosodic modifications and spectral transformations. Although the quality of the pitch modified speech signal was good, the quality was reduced when a modification factor out of the range [0.3, 2.5] was used. Pitch modifications with such factors (which are needed for voice conversion of type female ↔ male) should be combined with frequency scaling in order to preserve the quality of the modified synthetic signal. To do this, however, requires some rules of interactions between the modified harmonic structure and the frequencies of the formants.
CHAPTER 7. SUMMARY OF RESULTS AND FUTURE RESEARCH DIRECTIONS

- **New feature sets**
  While the proposed transformation function using discrete cepstral coefficients was able to reduce the spectral distortion between source and target speaker by a score of 70% – 80%, the quality of the converted speech signal, although satisfactory, suffers from a muffling effect when the number of mixture components is very small (8 or 16). Careful examination of different feature sets, such as formant parameters, may lead to improving the quality of the converted signal, particularly for more mixture components.

- **Definition of prosodic models for a speaker**
  In the present discussion the time and pitch contour of the source speaker has been modified by constant factors in order to match the average pitch and speaking rate of the target speaker. Because the HNM allows for time varying modification factors, using a prosodic model for the target speaker the prosody of the converted speech signal could be adjusted to be very close to the prosody of the target speaker. This could provide a way to increase the quality and the naturalness of the converted speech signal.

- **Application of the HNM to wideband speech coding**
  This work has shown the HNM to be a good model for speech which produces high quality speech re-synthesis. Because the harmonic amplitudes and phases can be estimated by fast algorithms and only the fundamental frequency has to be estimated, this model could be used for wideband (and narrowband) speech coding. It is worth noting that only a small number, typically 16 for sampling frequency at 16kHz, of discrete cepstral coefficients is needed in order to fit the harmonic amplitudes very well. Note that, if the alternative approach for the noise part proposed in the chapter 3 is used, harmonic amplitudes can be used both for the harmonic and the noise part. For the harmonic phases (only for the harmonic part) a model similar to this one proposed in [MQ91] could be used. Hence, by efficiently coding the discrete cepstral coefficients, the phases, the fundamental frequency and the number of harmonics included in the harmonic part, a (low rate) wideband speech coding scheme could be obtained.

- **Text-to-speech application**
  The speech and speaker modification techniques developed in this thesis could be used to improve the quality of the text-to-speech synthesis systems based on acoustical unit concatenation. This aspect has been discussed in the first Chapter of the thesis as one of the main applications of this work. As the quality of the modified speech signal using HNM is better than this one obtained from TD-PSOLA, which is the method the most frequently used in these systems, and after a careful examination of different difficulties posed for such an application (as for example the boundaries of the acoustical units), the HNM could be used directly for text-to-speech synthesis.
Appendix A

Another expression for the harmonic part

This appendix makes use of a sum of sinusoids terms as to represent the harmonic part of the HNM. In this case, it will be shown that the system used for estimating the harmonic amplitudes and phases, could be split into two sub-systems. We recall that the harmonic part of the HNM presented in Chapter 3, was expressed as a sum of exponential functions. Moreover, the matrices of these two sub-systems are characterized by a special structure: they are dominated by their main diagonal allowing the use of iterative approaches like Gauss-Seidel or SOR to solve the above sub-systems.

For convenience, we repeat here the basic equations for the harmonic part, \( h(t) \),

\[
h(t) = \sum_{k=1}^{\kappa(t)} a_k(t) \cos \phi_k(t) \quad (A.1)
\]

and the stationary conditions in the neighborhood of the analysis time-instants, \( t_a^i \),

\[
a_k(t) = a_k(t_a^i) \quad P(t) = P(t_a^i) \quad (A.2)
\]

for small \( |t - t_a^i| \). In Chapter 3, the instantaneous phase of \( k \)-th harmonic \( \phi_k(t) \) was developed in the neighborhood of \( t_a^i \) as

\[
\phi_k(t) = \phi_k(t_a^i) + k \, 2\pi f_0(t_a^i)(t - t_a^i) \quad (A.3)
\]

for small \( |t - t_a^i| \). \( f_0(t_a^i) = 1/P(t_a^i) \) is the fundamental frequency in Hz, at the time-instant \( t_a^i \) (the pitch period, \( P(t_a^i) \), is expressed in sec). It was also supposed that the maximum
voiced frequency $F(t)$ was constant around the time-instant $t_a^i$ and equal to $F(t_a^i)$. Using these so-called stationary conditions, the harmonic part around the time-instant $t_a^i$ can be rewritten as

$$h(t) = \sum_{k=1}^{K(t)} a_k(t)cos\phi_k(t)$$

$$= \sum_{k=1}^{L} a_k(t_a^i)cos(\phi_k(t_a^i) + k 2\pi f_0(t_a^i)(t - t_a^i))$$

(A.4)

where $L$ is the number of harmonics included in the harmonic part: $L = F(t_a^i)/f_0(t_a^i)$. The harmonic part can be split into a cosine and a sine part

$$h(t) = \sum_{k=1}^{L} \{c_kcos(k 2\pi f_0(t - t_a^i)) + s_k sin(k 2\pi f_0(t - t_a^i))\} + c_0$$

(A.5)

where $c_k$ and $s_k$ are related with the amplitudes $a_k$ and phases $\phi_k$ by

$$a_k = \sqrt{c_k^2 + s_k^2}$$

$$\phi_k = atan^{-1}\left(\frac{s_k}{c_k}\right)$$

(A.6)

Note that for convenience the dependence on the time-instant $t_a^i$ in (A.5) and in (A.6) has been dropped. The contribution of the $dc$ component, $c_0$, has been also included in (A.5). The unknown parameters, $c_k$, $c_0$ and $s_k$ are evaluated by use of a weighted least-squares method aiming at minimizing the following criterion with respect to $c_k$ and $s_k$

$$\epsilon = \sum_{t=t_a^i-N}^{t_a^i+N} w^2(t) (s(t) - \hat{h}(t))^2$$

(A.7)

where $s(t)$ is the original signal, $\hat{h}(t)$ is the harmonic part defined as in (A.5), $w(t)$ is the weighting function and $N$ is the integer closest to local pitch period $P(t_a^i) : N = \langle P(t_a^i) \rangle$, where $\langle \cdot \rangle$ represents the "round to nearest integer" operator. Note that the analysis frame is centered around the analysis instant $t_a^i$ and the length of the analysis window is $M = 2N + 1$. The weighting function $w^2(t)$ is needed in order to provide a better match with the signal portion located at the center of the analysis frame.

Using matricial formulation, we may rewrite the harmonic part simply as

$$\hat{h} = B x$$

(A.8)

The solution to least-squares problem is then given by the normal equations [LH74]

$$\left(B^T W^T W B\right) x = B^T W^T W s$$

(A.9)

where $B$ is a $M$-by-$(2L + 1)$ matrix with the last column corresponding to $dc$ component.

$$B = [bc_1 bs_1 bc_2 bs_2 \cdots bc_L bs_L 1]$$

(A.10)
where the vectors $\mathbf{b}_i$ and $\mathbf{b}_s$ for $i = 1, \ldots, L$ are of dimension $M \times 1$ and given by
\begin{align*}
\mathbf{b}_i^T &= \begin{bmatrix}
\cos(i \, 2\pi \, f_0 \, (-N)) \\
\cos(i \, 2\pi \, f_0 \, (-N + 1)) \\
\vdots \\
\cos(i \, 2\pi \, f_0 \, (N))
\end{bmatrix} \\
\mathbf{b}_s^T &= \begin{bmatrix}
\sin(i \, 2\pi \, f_0 \, (-N)) \\
\sin(i \, 2\pi \, f_0 \, (-N + 1)) \\
\vdots \\
\sin(i \, 2\pi \, f_0 \, (N))
\end{bmatrix}
\end{align*}
(A.11)
while $\mathbf{1}$ denotes the $M$-by-$1$ vector $\mathbf{1}^T = [1 \, 1 \, \ldots \, 1]$. $\mathbf{W}$ is a $N$-by-$N$ diagonal matrix with diagonal elements the weight vector $\mathbf{w}^T = [w(-N) \, w(-N + 1) \, \ldots \, w(N)]$ which is a typical Hamming window.
Note that $\mathbf{s}$ is a $M$-by-$1$ vector which contains the original data $\mathbf{s}^T = [s(-N) \, s(-N + 1) \, \ldots \, s(N)]$ and $\mathbf{x}$ is a vector of length $(2L + 1)$ which contains the sought parameters and is defined by
\begin{equation}
\mathbf{x}^T = [c_1 \, c_2 \, c_3 \, \ldots \, c_L \, s_L \, c_0]
\end{equation}
(A.12)
where the symbol "$T$" denotes the transpose operator.
Using some simple trigonometrical algebra and the fact that the weight vector $\mathbf{w}$ is a symmetric window, the $(2L + 1)$-by-$(2L + 1)$ matrix $\mathbf{A} = \mathbf{B}^T \mathbf{W}^T \mathbf{W} \mathbf{B}$ is a definite positive matrix with a special structure. More precisely the matrix $\mathbf{A} = [a_{ij}]$ is a symmetric matrix with
\begin{equation}
a_{kl} = \begin{cases}
\neq 0 & k = 2i \\
& l = 2j \\
& i, j = 1, 2, \ldots, L \\
\neq 0 & k = 2i - 1 \\
& l = 2j - 1 \\
& i, j = 1, 2, \ldots, L, L + 1 \\
= 0 & \text{otherwise}
\end{cases}
\end{equation}
(A.13)
Due to this special structure of the matrix $\mathbf{A}$, it is possible to split the $(2L + 1)$-by-$(2L + 1)$ optimization problem into two optimization problems of dimension $(L + 1)$-by-$(L + 1)$ and $L$-by-$L$ respectively. For example with $L = 2$ harmonics the matrix $\mathbf{A}$ has the follow structure
\begin{equation}
\mathbf{A} = \begin{bmatrix}
0 & a_{13} & 0 & a_{15} \\
a_{13} & 0 & a_{24} & 0 \\
0 & a_{24} & 0 & a_{35} \\
a_{15} & 0 & a_{35} & 0 \\
a_{51} & 0 & a_{53} & 0
\end{bmatrix}
\end{equation}
(A.14)
Noting as $\mathbf{b} = \mathbf{B}^T \mathbf{W}^T \mathbf{W} \mathbf{s}$ the right part in (A.9) the two split problems are
\begin{equation}
\begin{bmatrix}
a_{11} & a_{13} & a_{15} \\
a_{31} & a_{33} & a_{35} \\
a_{51} & a_{53} & a_{55}
\end{bmatrix}
\begin{bmatrix}
c_1 \\
c_2 \\
c_0
\end{bmatrix}
= \begin{bmatrix}
b_1 \\
b_3 \\
b_5
\end{bmatrix}
\text{ and }
\begin{bmatrix}
a_{22} & a_{24} \\
a_{42} & a_{44}
\end{bmatrix}
\begin{bmatrix}
s_1 \\
s_2
\end{bmatrix}
= \begin{bmatrix}
b_2 \\
b_4
\end{bmatrix}
\end{equation}
(A.15)
In order to reduce the computational effort for the construction of the matrix $\mathbf{A}$ an auxiliary vector $\mathbf{r}$ of length $(2L + 1)$ can be used which is defined as
\begin{equation}
r(i) = \sum_{n=1}^{N} w^2(n) \cos(i \, 2\pi \, f_0(n)) \quad \text{for} \ i = 0, \ldots, 2L
\end{equation}
(A.16)
and the elements of the matrix $A = [a_{ij}]$ can be computed as follows

\[
\begin{align*}
    a_{2i,2j} &= r(|i - j|) - r(i + j) \quad i, j = 1, \ldots, L \\
    a_{2i-1,2j-1} &= r(|i - j|) - r(i + j) \quad i, j = 1, \ldots, L \\
    a_{2i-1,2L+1} &= 2r(i) + w^2(0) \quad i = 1, \ldots, L \\
    a_{2L+1,2L+1} &= 2r(0) + w^2(0)
\end{align*}
\]  

(A.17)

The matrices that need to be inverted into the two split optimization problems are symmetric and positive definites so that the respective normal equations can be solved advantageously using the Cholesky decomposition. However, another property of the above matrices is that they are dominated by the diagonal and most of theirs elements can be put to zero since the harmonics very often are well separated and so have a negligible correlation. This fact makes it possible to use iterative methods to solve the above equations. For example, the Gauss-Seidel method or a more efficient method like the SOR (Successive Over-Relaxation) [PTVF94] algorithm, are well suited to solve these kind of problems.
Appendix B

Sound examples

This appendix presents a few sound examples that show the capabilities of the HNM for high-quality speech resynthesis and prosodic modifications. Also, some examples of speaker conversion are given.

An audio tape with the examples can be ordered from: ENST-Paris, 46 Rue Barrault, Signal Department, 75013 Paris.

Sound Example 1:
Subject: Speech analysis / modification / synthesis
Speaker: MDnas  Sex: female
Utterance: Mon père m’a donné la permission de jouer au tennis cet après-midi.

Index 1: Analysis and Synthesis

- Original
- Harmonic part
- Noise part
- Harmonic + Noise part = Synthetic signal
- Original
- Synthetic
Index 2: Prosodic modifications

- Pitch modification by 0.6
- Pitch modification by 0.8
- Pitch modification by 1.4
- Pitch modification by 1.6
- Time-scale modification by 0.7
- Time-scale modification by 1.6
- Time-scale modification by 2.0
- Time-varying pitch and time-scale modification

**Sound Example 2:**
Subject: Speech analysis / modification / synthesis
Speaker: SLnas Sex: male
Utterance: (a) Souvent je m’accoude au muret de ce pont. and (b) Un colonel commandait le régiment

Index 3: Analysis and Synthesis

- Original
- Harmonic part
- Noise part
- Harmonic + Noise part = Synthetic signal
- Original
- Synthetic

Index 4: Prosodic modifications

- Pitch modification by 0.6
- Pitch modification by 0.8
- Pitch modification by 1.4
- Pitch modification by 1.6
- Time-scale modification by 0.7
- Time-scale modification by 1.6
- Time-scale modification by 2.0
- Time-varying pitch and time-scale modification

**Sound Example 3:** Subject: Voice transformation
Source speaker: ob1 Sex: male
Target speaker: rg1 Sex: male
Conversion function: Diagonal-type
Number of components: 128

Utterance: *Je ne veux pas que vous les changez pour le moment*

Index 5 and 6: Analysis and Synthesis
- Source (original)
- Target (original)
- Prosodic modifications of the source
- Source transformed (Prosodic modif.+ spectral transf.)
- Source transformed (Prosodic modif.+ spectral transf.)

**Sound Example 4:** Subject: Voice transformation
Source speaker: ob2 Sex: male
Target speaker: rg2 Sex: male
Conversion function: Full-type
Number of components: 64

Utterance: *Il a été condamné pour un vol de voiture.*

Index 7 and 8:
• Source (original)

• Target (original)

• Prosodic modifications of the source

• Source transformed (Prosodic modif.+ spectral transf.)

• Source transformed (Prosodic modif.+ spectral transf.)

**Sound Example 5:**
Subject: Voice transformation
Source speaker: ob3  Sex: male
Target speaker: rg3  Sex: male
Conversion function: Diagonal-type
Number of components: 128

Utterance: *La jeune fille se peigne devant sa glace.*

Index 9 and 10:

• Source (original)

• Target (original)

• Prosodic modifications of the source

• Source transformed (Prosodic modif.+ spectral transf.)

• Source transformed (Prosodic modif.+ spectral transf.)

**Sound Example 6:**
Subject: Time-scale modifications with large modification factors
Speaker: FCnas  Sex: male
Utterance: *Magnifique.*

Index 11: Large time-scale factors

• Original

• Time-scale modification by 2
- Time-scale modification by 4
- Time-scale modification by 6
- Time-scale modification by 8
Appendix C

Résumé détaillé en français

Cet annexe contient un résumé détaillé en français du travail effectué dans cette thèse.

C.1 Introduction

Bien que la parole soit le moyen le plus utilisé pour la communication humaine, celle-ci est presque absente de la communication homme-machine. La raison de cette absence serait les techniques de traitement de parole qui, liées à la communication homme-machine, n’établissent pas encore une communication de haute qualité. L’établissement d’une telle communication devient de plus en plus nécessaire vu l’accroissement des applications d’ordinateurs dans la vie quotidienne. Les domaines du traitement de la parole concernant à la communication homme-machine sont: la reconnaissance de parole et la synthèse de parole. Notre attention se focalise sur le domaine de la synthèse de parole. Il y a deux types de systèmes de synthèse: les systèmes de simple concaténation de messages pré-enregistrés et les systèmes de synthèse de parole “complets” (full text-to-speech systems). Malheureusement, la bonne qualité de synthèse obtenue par la première catégorie, s’accompagne d’un manque de flexibilité de ces systèmes, qui sont incapables produire n’importe quel message (vocabulaire limité). D’autre part, bien que le deuxième type soit très flexible (vocabulaire illimité) la qualité de synthèse n’est pas à la hauteur. Or, les systèmes de synthèse de parole complets ont un potentiel plus intéressant. L’amélioration de la qualité de la synthèse obtenue par ces systèmes nécessite le perfectionnement des trois points suivants:

- l’analyse linguistique
• les règles de la prosodie

• les modèles de la synthèse (synthesizers)

Ce dernier point ne concerne que les techniques de traitement du signal.

On peut distinguer deux classes de systèmes de synthèse de parole complets: la synthèse par formants (formant synthesis) et la synthèse par concatenation d’unités d’acoustiques (diphones ou des unités de longueur supérieure). La deuxième catégorie fournit une qualité de synthèse supérieure à la première. Pourtant, afin de préserver la qualité de voix synthétique, les systèmes de synthèse basés sur la concatenation d’unités d’acoustiques exigent des passages non brutaux d’une unité d’acoustique à l’autre; cela implique la modification de certaines caractéristiques de ces unités, telles que la durée et la hauteur (modifications prosodiques).

Une autre propriété désirée des systèmes de synthèse consiste en leur capacité de changer l’identité de la voix synthétique. Dans le cas des systèmes de synthèse par concatenation d’unités d’acoustiques, on ne peut changer l’identité de la voix synthétique que par le réenregistrement de la base de données par un nouveau locuteur. Mais l’enregistrement d’une nouvelle voix est très coûteux, et cette solution est en pratique irréaliste si on veux souvent changer la voix du système. Une autre possibilité consiste à transformer la voix enregistrée en une voix désirée (voice conversion or speaker modification). Le problème de modification du locuteur est très proche du problème de l’identification du locuteur. Les recherches dans ce dernier domaine du traitement de parole montrent que les paramètres qui contribuent le plus à l’identification du locuteur sont les paramètres prosodiques et l’enveloppe spectrale. Ainsi, afin de modifier le locuteur il faut modifier à la fois les paramètres prosodiques et l’enveloppe spectrale.

On peut envisager maintenant la puissance et la flexibilité d’un système de synthèse de parole permettant de synthétiser une voix de haute qualité et en même temps de changer l’identité du locuteur. Or, ces systèmes se caractérisent d’un vocabulaire et d’une gamme de voix illimitée.

Cette thèse met l’accent sur la modification de haute qualité de la parole (modification prosodiques) et du locuteur (modification du locuteur ou conversion de voix). Notre objectif est de fournir, en utilisant les techniques de traitement du signal, un synthétiseur qui permettra d’une façon flexible et robuste de synthétiser et modifier la parole et le locuteur tout en gardant la qualité de la voix synthétique. Plus précisément, nous cherchons un modèle paramétrique qui serait capable de:

• extraire du signal de parole les paramètres appropriés pour notre objectif;
C.2. LE MODÈLE HNM POUR LA PAROLE

- utiliser ces paramètres pour faire une reproduction du signal de haute qualité (c'est-à-dire, sans qu'il soit possible de distinguer le signal original du signal reproduit par le synthétiseur);
- modifier les paramètres prosodiques (durée et hauteur) pour la modification de parole;
- contrôler l'enveloppe spectrale pour les modifications spectrales;
- combiner les modifications prosodiques avec les modifications spectrales afin de pouvoir modifier le locuteur.

Cette annexe se divise, comme le document de la thèse, en deux parties: la première concerne la proposition d'un synthétiseur de parole et les modifications prosodiques de parole en utilisant ce synthétiseur. La deuxième partie traite de la modification et du contrôle de l'enveloppe spectrale. Dans cette même partie nous présentons un système combinant les deux types de modification, prosodiques et spectrales, pour la modification du locuteur.

C.2 Le modèle HNM pour la parole

Cette section présente le modèle HNM (Harmonic plus Noise Model), un nouveau système d'analyse/modification/synthèse de parole. Le modèle HNM suppose que le signal de parole peut être décomposé en une partie harmonique $h[n]$ et en une partie bruitée $r[n]$. La partie harmonique modélise la composante quasi-périodique du signal de parole; la partie bruitée modélise la composante non-périodique du signal, c'est-à-dire le bruit de friction et les variations de l'excitation glottale d'une période à l'autre. Le spectre est divisé seulement en deux bandes et uniquement pour les trames voisées. Les deux bandes se délimitent par la fréquence maximale de voissement qui est en général un paramètre variant dans le temps. Dans la bande inférieure, le signal est représenté par une somme d'harmoniques avec des amplitudes qui varient lentement (processus passe-bas). Nous avons distingué trois versions du modèle HNM:

- le modèle HNM$_1$, où les amplitudes d'harmoniques sont constantes dans la fenêtre d'analyse et la phase est supposée être linéaire.
- le modèle HNM$_2$ qui utilise des amplitudes complexes variant d'une façon linéaire dans la fenêtre d'analyse, et

---

1 Ces trois modèles sont notés dans le texte anglais HNM, DSM et HSM. Ici, nous allons simplifier leur notation en HNM$_1$, HNM$_2$ et HNM$_3$ respectivement.
le modèle HNM3 qui utilise un polynôme de troisième ordre pour les amplitudes d’harmoniques et un polynôme de premier ordre pour les phases (phase linéaire).

Les expressions mathématiques pour la partie harmonique des trois modèles HNM sont

- **HNM1:**
  \[
  h_1[n] = \sum_{k=-L(n_a^i)}^{L(n_a^i)} a_k(n_a^i) e^{j2\pi k f_o (n_a^i)(n-n_a^i)}
  \]  
  (C.1)

- **HNM2:** Somme de fonctions exponentielles avec une pente complexe
  \[
  h_2[n] = \sum_{k=-L(n_a^i)}^{L(n_a^i)} A_k(n) e^{j2\pi k f_o (n_a^i)(n-n_a^i)}
  \]  
  (C.2)
  où
  \[
  A_k(n) = a_k(n_a^i) + (n - n_a^i) b_k(n_a^i)
  \]  
  (C.3)
  avec \(a_k(n_a^i), b_k(n_a^i)\) sont des nombres complexes (amplitude et pente respectivement).

- **HNM3:** Somme de sinusoïdes avec des amplitudes réelles variant dans le temps
  \[
  h_3[n] = \sum_{k=0}^{L(n_a^i)} a_k(n) \cos(\varphi_k(n))
  \]  
  (C.4)
  où
  \[
  a_k(n) = \alpha_k + \beta_k (n - n_a^i) + \gamma_k (n - n_a^i)^2 + \delta_k (n - n_a^i)^3
  \]
  \[
  \varphi_k(n) = \epsilon_k + 2\pi k \zeta (n - n_a^i)
  \]  
  (C.5)

Les paramètres des trois modèles s’estiment à des instants spécifiques notés par \(n_a^i\). Les \(f_o(n)\) et \(L(n)\) représentent respectivement la fréquence fondamentale et le nombre d’harmoniques inclus dans la partie harmonique. Ces deux paramètres sont gardés constants dans chaque fenêtre d’analyse et sont égaux à leur valeur correspondant au centre, \(n_a^i\), de la fenêtre d’analyse.

La partie apériodique est obtenue en soustrayant dans le domaine temporel la partie harmonique du signal original

\[
r[n] = s[n] - h[n]
\]  
(C.6)

où \(h[n]\) est la partie harmonique d’un des modèles ci-dessus et \(s[n]\) représente la version discrète du signal original continu \(s(t)\) échantillonné à une fréquence \(F_s\).
C.2.1 Estimation des paramètres des modèles

Étant donné que nous avons considéré que le signal est harmonique l’estimation de la fréquence fondamentale et de la fréquence maximale du voisement constitue la première étape de l’analyse. Dans un premier temps, une estimation initiale de la fréquence fondamentale est obtenue en utilisant un détecteur basé sur la fonction normalisée de l’autocorrélation. Puis, un algorithme adéquat à la détection des pics, utilisant l’estimation initiale du pitch (fréquence fondamentale), sépare les pics “harmoniques” (voisés) des pics aléatoires. La fréquence du dernier pic voisé détermine la fréquence maximale du voisement. Une fréquence fondamentale raffinée est également obtenue en choisissant comme fréquence fondamentale celle dont les fréquences harmoniques se rapprochent mieux des fréquences des pics voisés. À la fin de cette première étape, les instants d’analyse, $n^*_a$, sont imposés synchrones avec la fréquence fondamentale (pitch-synchronous rate).

Lors de la deuxième étape de l’analyse les amplitudes et les phases des harmoniques s’estiment en minimisant le critère

$$\epsilon = \sum_{n = n^*_a - N}^{n^*_a + N} w^2[n](s[n] - \hat{h}[n])^2$$  \hspace{1cm} (C.7)

par rapport: aux $A_k(n^*_a)$ pour le HNM$_1$, aux $a_k(n^*_a)$ et $b_k(n^*_a)$ pour le HNM$_2$, et aux coefficients inconnus des polynômes des amplitudes et des phases pour le HNM$_3$. Dans l’équation (C.7), $\hat{h}[n]$ représente un des modèles harmoniques HNM$_1$, HNM$_2$ et HNM$_3$, $w[n]$ la fenêtre de pondération et $N$ le nombre entier de la période fondamentale locale, $T(n^*_a)$. A ce point de l’analyse nous trouvons plus convivial de mettre sous forme matricielle les trois modèles harmoniques:

$$\begin{align*}
\text{HNM}_1 : \quad & P_1 x_1 = h_1 \\
\text{HNM}_2 : \quad & P_2 x_2 = h_2 \\
\text{HNM}_3 : \quad & P_3 x_3 = h_3
\end{align*}$$  \hspace{1cm} (C.8)

où $P_1$ est une matrice $(2N + 1)\times(2L + 1)$, $P_2$ est une matrice $(2N + 1)\times(4L + 2)$ et $P_3$ est une matrice $(2N + 1)\times(4L + 4)$. La matrice $P_1$ est définie comme

$$P_1 = \begin{bmatrix}
\mathbf{b}_{-L} : \mathbf{b}_{-L+1} : \mathbf{b}_{-L+2} : \cdots : \mathbf{b}_L
\end{bmatrix}$$  \hspace{1cm} (C.9)

où $\mathbf{b}_k$ est un vecteur $(2N + 1)\times1$ correspondant à la $k$-ème harmonique définie comme :

$$\mathbf{b}_k = \begin{bmatrix}
e^{j2\pi k f_0(t_a^* - N)} & e^{j2\pi k f_0(t_a^* - N + 1)} & e^{j2\pi k f_0(t_a^* - N + 2)} & \cdots & e^{j2\pi k f_0(t_a^* + N)}
\end{bmatrix}^T$$  \hspace{1cm} (C.10)

et où “T” représente l’opération de transposition. La matrice $P_2$ est donnée par

$$P_2 = [B_1|B_2|B_3|B_4]$$  \hspace{1cm} (C.11)
avec
\[
\begin{align*}
(B_1)_i & = E^{(i-N)(l+1)} \\
(B_2)_i & = E^{-(i-N/2)(l+1)} \\
(B_3)_i & = (i - N/2) (B_1)_i \\
(B_4)_i & = (i - N) (B_2)_i
\end{align*}
\]

(où \( E = \exp(j2\pi f_0(n^i_a)) \) avec \( i \) de 0 à 2N et \( l \) de 0 à L).

La matrice \( P_3 \) est définie par
\[
P_3 = \begin{bmatrix}
B_1 : \text{diag}(n)B_1 : \text{diag}(n)^2B_1 : \text{diag}(n)^3B_1
\end{bmatrix}
\]

(où \( \mathbf{n} = [-N, -N + 1, \ldots, N]^T \) représente un vecteur temporel \((2N + 1) \times 1\), \( \text{diag}(\mathbf{n}) \) représente une matrice diagonale \((2N + 1) \times (2N + 1)\) avec \( \mathbf{n} \) les entrées diagonales de celle-ci, et \( B_1 \) une matrice \((2N + 1) \times (L + 1)\) donnée par
\[
B_1 = \begin{bmatrix}
b_1 : b_2 : \cdots : b_L : 1
\end{bmatrix}
\]

(où \( \mathbf{b}_k \) \((k = 1, \ldots, L)\) est un vecteur \((2N + 1) \times 1\) défini par
\[
\mathbf{b}_k = [\cos(\varphi_k[-N]) \cos(\varphi_k[-N + 1]) \cdots \cos(\varphi_k[N])]^T \quad \text{for } k = 1 \cdots L
\]

et finalement, \( \mathbf{1} \) est le \((N + 1) \times 1\) vecteur unitaire : \( \mathbf{1} = [1 1 \cdots 1]^T \). Dans (C.15), \( \varphi_k[n] \) représente le polynôme du premier ordre de la phase.

Le vecteur \( \mathbf{x} \) pour chaque modèle harmonique est défini par:

- **HNMM** \((1)\) (un vecteur \((2L + 1) \times 1\)):
\[
\mathbf{x}_1 = \begin{bmatrix}
a_{-L}^* \ a_{-L+1}^* \ a_{-L+2}^* \ \cdots \ a_L^*
\end{bmatrix}
\]

- **HNMM** \((2)\) (un vecteur \((4L + 2) \times 1\)):
\[
\mathbf{x}_2 = [a_1 \ a_2 \cdots a_L \ a_{-L}^* \ a_{-L+1}^* \ b_1 \ b_2 \cdots b_L \ a_1^* \ b_1^* \ b_{-L}^* \ b_{-L+1}^*]^T
\]

- **HNMM** \((3)\) (un vecteur \((4L + 4) \times 1\)):
\[
\mathbf{x}_3 = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_L \ \alpha_0 \ \beta_1 \ \beta_2 \ \cdots \ \beta_L \ \beta_0 \ \gamma_1 \ \gamma_2 \ \cdots \ \gamma_L \ \gamma_0 \ \delta_1 \ \delta_2 \ \cdots \ \delta_L \ \delta_0]^T
\]

Pour les deux premiers modèles le critère (4.4) conduit à un problème de moindres carrés pour lequel la solution est donnée par

- **HNMM** \((1)\):
\[
\mathbf{x}_1 = \left( P_1^* \mathbf{W}^T \mathbf{P}_1 \mathbf{W} \right)^{-1} P_1^* \mathbf{W}^T \mathbf{W} \mathbf{s}
\]
C.2. LE MODÈLE HNM POUR LA PAROLE

- HNM\(_2\):

\[
x_2 = \left( P_2^h W^T P_2 W \right)^{-1} P_2^h W^T W s
\]

où \( s \) dénote le vecteur \((2N+1) \times 1\) contenant les échantillons du signal original de la trame courante:

\[
s = [s[-N] \, s[-N+1] \cdots s[N]]^T
\]

et \( W \) dénote une matrice diagonale \((2N+1) \times (2N+1)\), ayant comme diagonale principale le vecteur:

\[
w = [w[-N] \, w[-N+1] \cdots w[N]]^T
\]

qui est une fenêtre de Hamming. Nous avons montré que la matrice à inverser pour le HNM\(_1\) dans (C.19) est Toeplitz alors que pour le modèle HNM\(_2\), elle est Toeplitz par bloc. Ainsi, on peut utiliser des algorithmes rapides pour résoudre les systèmes d'équations correspondant aux HNM\(_1\) et HNM\(_2\).

Malheureusement une telle solution directe n'est pas possible pour le modèle HNM\(_3\) parce que les paramètres à optimiser ne sont pas séparables. Afin de résoudre ce système la solution suivante a été adoptée (relaxation d'une itération): nous avons initialisé les coefficients du polynôme pour les amplitudes en mettant tous les coefficients à zéro sauf les \( \alpha_k \). On obtient donc un polynôme d'ordre zéro pour les amplitudes et ainsi le modèle HNM\(_3\) se réduit au modèle HNM\(_1\). Alors, les coefficients du polynôme de phase sont données par

\[
\begin{align*}
\zeta &= f_0(n^i_u) \\
\epsilon_k &= \phi^i_k \text{ for } k = 1, \cdots, L
\end{align*}
\]

où \( f_0(n^i_u) \) est la fréquence fondamentale et \( \phi^i_k \) est la phase de la \( k \)-ième harmonique au centre de la fenêtre d’analyse, \( n^i_u \). Il faut noter que les phases \( \phi^i_k \) peuvent être estimées facilement en utilisant le modèle HNM\(_1\). Ayant estimé tous les coefficients du polynôme des phases, les coefficients du polynôme des amplitudes sont estimés par

\[
\begin{align*}
x_3 &= \left( P_3^T W^T P_3 W \right)^{-1} P_3^T W^T W s \quad \text{(C.24)} \\
x_3 &= \left( P_3^T W^T P_3 W \right)^{-1} P_3^T W^T W s \quad \text{(C.25)}
\end{align*}
\]

Malheureusement la matrice à inverser pour le modèle HNM\(_3\) n’a pas une structure particulière (comme HNM\(_1\) et HNM\(_3\)).

La partie harmonique pour chaque modèle est obtenue par

\[
\hat{h}_i = W^{-1} P_1 x_i
\]

où l’indice “i” dénote le modèle utilisé.

Avant de continuer avec l’estimation des paramètres pour la partie bruitée, comparezons le signal résiduel, \( r_i[n] = s[n] - \hat{h}_i[n] \) obtenu pour chaque modèle.
C.2.2 Comparaison des variances du signal résiduel

Nous avons montré que si le signal d’entrée, \( s[n] \), est un bruit blanc avec une variance 1, la variance du signal résiduel est donnée par la diagonale de la matrice suivante

\[
E(\mathbf{rr}^h) = \mathbf{I} - \mathbf{W} \mathbf{P}^h \mathbf{W}^h \mathbf{P}^{-1} \mathbf{W}^h
\]  

(C.27)

Les variances de trois modèles sont représentés Fig.C.1. Nous pouvons constater sur cette figure que les variances des modèles HNM\(_2\) et HNM\(_3\) sont proches de la variance idéale (variance constante) au contraire de la variance du signal résiduel obtenue par le modèle HNM\(_1\).

![Figure C.1: La variance du signal résiduel obtenue par: HNM\(_1\) (trait continu), HNM\(_2\) (trait interrompu) et HNM\(_3\) (trait pointillé). La fenêtre de pondération était une fenêtre de Hamming.](image)

C.2.3 Comparaison des signaux résiduels dans le domaine temporel

Le modèle HNM\(_1\) utilise un modèle stationnaire pour l’estimation des amplitudes et des phases des harmoniques. On note l’apparition de basses fréquences dans le signal résiduel obtenu par le HNM\(_1\). Afin de montrer ce comportement du HNM\(_1\) nous avons sélectionné une trame fricative voisée d’un signal de parole échantillonné à 16kHz. La trame utilisée est affichée Fig.C.2(a); la longueur de la trame est égale aux deux périodes fondamentales. Des pics harmoniques sont détectés jusqu’à 4000Hz. Les Fig.C.2 (b)(c) et (d) montrent les signaux résiduels obtenus par les modèles HNM\(_1\), HNM\(_2\) et HNM\(_3\) respectivement. Nous pouvons constater sur cette figure que dans les signaux résiduels obtenus par HNM\(_2\) et HNM\(_3\) nous ne trouvons que la partie apériodique du signal tandis que dans le signal résiduel obtenu par HNM\(_1\) nous observons l’existence de basses fréquences (qui normalement appartiennent à la partie harmonique). Pour mieux étudier les propriétés des signaux résiduels de HNM\(_2\) et
C.2. LE MODÈLE HNM POUR LA PAROLE

HNM₃ nous montrons dans la Fig.C.3 l’amplitude en dB de la transformée de Fourier des deux signaux résiduels (par HNM₂ et HNM₃). Dans la même figure nous avons inclus le spectre du signal original (trait discontinu). Nous rappelons que la bande “harmonique” se situe entre 0Hz et 4000Hz. Nous voyons sur cette figure que le spectre du signal résiduel obtenu par le HNM₂ (Fig.C.3(a)) est (bizarrement) plat dans la bande harmonique. Afin de vérifier que les deux modèles séparent bien la partie périodique de la partie aperiodique, nous avons effectué le test suivant: un bruit blanc filtré par un filtre passe-bas avec une fréquence de coupure égale à la fréquence maximale du voeux (4000Hz) a été ajouté avec une puissance 25dB au signal original de parole; les paramètres de HNM₂ et HNM₃ sont estimés à partir du signal bruité (parole + bruit filtré). Le contenu fréquentiel des deux signaux résiduels obtenus par les modèles HNM₂ et HNM₃ sont représentés Fig.C.4(a) et Fig.C.4(b). L’amplitude du signal original sans bruit additif est aussi représenté sur cette figure (trait discontinu). Cette figure montre clairement que presque tout le bruit additif passe dans la partie harmonique dans le cas du HNM₂ ce qui n’est pas le cas pour le HNM₃.

C.2.4 Comparaison de l’erreur de modélisation

Un dernier critère de comparaison des trois modèles sera l’erreur de modélisation. La mesure de similarité entre la partie harmonique (bande harmonique : 0Hz jusqu’à la fréquence maximale de voeux) et le signal original (jusqu’à 8000Hz) que nous avons utilisé est la suivante

$$E = 10\log_{10} \frac{\sigma_r^2(t)}{\sigma_s^2(t)}$$  \hspace{1cm} (C.28)
Figure C.3: Le spectre en dB des amplitudes de la transformée de Fourier des signaux résiduels pour HNM2 (en haut) et HNM3 (en bas). Le spectre en dB des amplitudes de la transformée de Fourier du signal original est aussi inclu (trait discontinu).

Figure C.4: Le spectre en dB des amplitudes de la transformée de Fourier des signaux résiduels pour HNM2 (en haut) et HNM3 (en bas) avec un bruit additif. Le spectre en dB des amplitudes de la transformée de Fourier du signal original est aussi inclu (trait discontinu).
où $\sigma_r^2(t)$ dénote la variance du signal résiduel $r(t)$ et $\sigma_s^2(t)$ dénote la variance du signal original $s(t)$. Par exemple, l’erreur de modélisation produite par les trois modèles pour la trame utilisée dans la section précédente était de: $-15.8dB$ pour le HNM$_1$, $-25.6dB$ pour le HNM$_2$ et $-25.58dB$ pour HNM$_3$.

La Fig.C.5(a) montre un signal de parole ("wasi") échantillonné à 16kHz et la Fig.C.5(b) montre l’erreur de modélisation obtenue par les trois modèles (HNM$_1$ trait solide, HNM$_2$ trait discontinu, HNM$_3$ trait en pointillé) en fonction du nombre des trames. Il est évident d’après cette figure que l’erreur de modélisation produite par le HNM$_1$ est plus grande que celles des deux autres modèles. Notez que comme on a utilisé toute la bande fréquentielle pour le calcul de l’erreur, celle-ci est plus importante dans les régions fricatives voisées (entre la 25$^{\text{ème}}$ et la 35$^{\text{ème}}$ trame). Il faut noter aussi que si l’erreur de modélisation est inférieure à $-25dB$, le signal synthétique produit par les modèles sera indiscernable du signal original. Bien que 

![Image](image_url)

Figure C.5: Erreur de modélisation en dB utilisant les trois modèles.

l’erreur obtenue par le modèle HNM$_1$ soit la plus grande, l’analyse et la synthèse (comme nous allons voir) sont les plus simples et pourraient être implantées en temps réel, tout en conservant une bonne qualité au signal synthétisé. Pour toutes ces raisons nous avons décidé de choisir le modèle HNM$_1$ pour la modification prosodique du signal de parole et pour la modification (conversion) du locuteur. Dans la suite nous désignerons le HNM$_1$ simplement par HNM.

C.2.5 Le modèle pour la partie bruitée

La partie bruitée $n[n]$, rend compte soit des trames non-voisées (où la partie harmonique n’existe pas) soit du bruit de friction et des fluctuations d’une période à l’autre pour les
trames voisées. Dand le cas des trames voisées la partie bruitée du modèle HNM occupe seulement la partie haute du spectre, au-delà de la fréquence maximale du voisement (signal passe-haut). L’énergie de la partie bruitée dans les trames voisées est synchrone à la fréquence fondamentale. Le modèle HNM suit cette observation: le contenu fréquentiel de la partie bruitée est représenté par un modèle AR variant dans le temps; sa structure temporelle est représentée par une fonction linéaire par morceau. La partie bruitée $n[n]$ peut donc être obtenue en filtrant un bruit blanc gaussien, $u[n]$ par un filtre tout pôle $h[n, \tau]$ normalisé et variant dans le temps et ensuite multiplié les résultats par une enveloppe d’énergie $e[n]$

$$n[n] = e[n] \cdot h[n, \tau] \ast u[n]$$

(C.29)

Le filtre $h[n, \tau]$ est évalué à chaque instant d’analyse $n^a_i$. Finalement le signal synthétisé $s[n]$ est obtenu par la superposition de la partie harmonique $h[n]$ et de la partie bruitée $n[n]$: 

$$s[n] = h[n] + n[n]$$

(C.30)

### C.2.6 Estimation des paramètres de la partie bruitée du HNM

Après l’estimation des paramètres de la partie harmonique, l’étape suivante est l’estimation des paramètres de la partie bruitée. Pour chaque trame, la densité spectrale du signal original se modélise par un filtre d’ordre $p$ tout pôle. Nous avons utilisé la densité spectrale du signal original et non celle du signal résiduel car dans la section précédente nous avons vu que le signal résiduel du modèle HNM contient des basses fréquences, ce qui perturbe la modélisation de la partie bruitée. Le modèle HNM suppose que dans la bande harmonique (0Hz - fréquence maximale de voisement) il n’y ait que la partie périodique et que la partie bruitée se trouve au-delà de la fréquence maximale de voisement. Nous modélisons lors de l’analyse toute la bande spectrale du signal original. Le signal synthétisé sera filtré passe haut pour en extraire la partie bruitée.

Pour la structure temporelle de la partie bruitée, nous avons trouvé qu’une enveloppe triangulaire (voir Fig.C.6) donne des résultats satisfaisants. L’avantage de ce type d’enveloppe est double: d’abord nous n’avons pas besoin de l’estimer analytiquement, ensuite on change facilement la longueur d’une enveloppe paramétrique pour des modifications prosodiques.

### C.2.7 Synthèse sans modification

La synthèse comme l’analyse est synchrone à la fréquence fondamentale (pitch-synchronous synthesis). Ceci signifie que les instants de synthèse $n^s_i$ coïncident avec les instants d’analyse $n^a_i$ i.e.: $n^s_i = n^a_i \ \forall i$. 
Figure C.6: L’enveloppe utilisée pour la structure temporelle de la partie bruitée. Les $t^i_a$ et $t^{i+1}_a$ représentent deux instants d’analyse successifs. Des valeurs typiques pour $l_1$ et $l_2$ sont : $l_1 = 0.15(n^{i+1}_a - n^i_a)$ et $l_1 = 0.85(n^{i+1}_a - n^i_a)$.

La partie harmonique est synthétisée en utilisant directement l’équation C.1 (dans le cas du HNM$_3$). Les amplitudes et les phases des harmoniques sont interpolées linéairement entre les trames successives. Les phases doivent être déroulées (unwrapping) dans le domaine temporel avant l’interpolation linéaire, en effectuant une prédiction de la phase utilisant la phase de la trame précédente et la fréquence fondamentale moyenne des deux instants de synthèse successifs. La Fig.C.7 montre le diagramme de la synthèse pour la partie harmonique. Il faut noter

Figure C.7: Diagramme de la synthèse de la partie harmonique pour $n \in [n^i_a, n^{i+1}_a]$.

que l’hypothèse d’harmonicité simplifie énormément l’association des fréquences entre deux instants de synthèse ; la $k$-ième harmonique de la trame $i$ correspond à la $k$-ième harmonique de la trame $i+1$. 
Comme la fréquence maximale de voisement est un paramètre variable dans le temps, la partie bruitée \( n[n] \) est synthétisée en utilisant la méthode de recouvrement et d’addition (Overlap-Add, OLA). Pour un instant de synthèse donné deux périodes fondamentales sont synthétisées en filtrant un bruit blanc gaussien de variance 1 par un filtre treillis normalisé (normalized lattice filter). Le signal obtenu à la sortie du filtre est multiplié par la variance du signal original estimée à l’instant \( n_s^i \). Si la trame est voisée les basses fréquences sont déjà synthétisées par la partie harmonique. Afin de conserver seulement les hautes fréquences, le bruit coloré est filtré par un filtre passe-haut.

Finalement, la partie bruitée de la trame courante (entre les instants \( n_s^i \) et \( n_s^{i+1} \)) est obtenue en ajoutant les parties bruitées synthétisées aux instants \( n_s^i \) et \( n_s^{i+1} \). Pour obtenir la structure temporel caractéristique de la partie bruitée des trames voisée, l’enveloppe temporelle (voir Fig.C.6) est appliquée directement sur le signal synthétisé. Le diagramme de la synthèse de la partie bruitée est donné Fig.C.8.

![Diagramme de la synthèse de la partie bruitée](image)

\( n(t, t_s^{i-1}) \)

\( \hat{n}(t) \)

\( V(t_s^i) \)

\( H(t_s^i) \): Normalized lattice filter

\( \alpha(t) \)

**Figure C.8**: Diagramme de la synthèse de la partie bruitée pour \( n \in [n_s^i, n_s^{i+1}] \).

### C.2.8 Synthèse avec modifications prosodiques

Dans cette section nous abordons les modifications prosodiques: les modifications de l’échelle temporelle (time-scale modification) et fréquentielle (pitch modification). Pour ces types de modifications la première tâche est de trouver, à partir des instants d’analyse et des facteurs de modification, les nouveaux instants de synthèse. Dès que nous disposons de ces derniers, la
C.2. LE MODÈLE HNM POUR LA PAROLE

175

synthèses s’effectue comme précédemment. En cas de modification de la fréquence fondamentale, nous devons recalcule les phases et les amplitudes des nouvelles harmoniques.

Nous allons voir dans un premier temps comment calculer les instants de synthèse et dans un deuxième temps comment recalcule les phases et les amplitudes des nouvelles harmoniques en cas de modification de la fréquence fondamentale.

C.2.8.1 Modification de l’échelle temporelle

L’objectif de la modification de l’échelle temporelle est de changer le rythme d’articulation sans affecter le contenu spectral. Connaissant les instants d’analyse, les facteurs de modification de l’échelle temporelle \( \beta(t) > 0 \) et la courbe mélodique \( P(t) \), en supposant que l’instant de synthèse \( t_s^{i+1} \) est connu, l’instant de synthèse suivant \( t_s^i \) est donné par la formule suivante:

\[
t_s^i + \frac{1}{t_s^{i+1} - t_s^i} \int_{t_s^i}^{t_s^{i+1}} P(t) \, dt \quad \text{avec} \quad t_s^{i+1} = D(t_s^i)
\]  

(C.31)

où les \( t_s^i \) sont des instants virtuels qui sont liés aux instants de synthèse par

\[
\begin{align*}
t_s^i &= D(t_s^i) \\
t_s^i &= D^{-1}(t_s^i)
\end{align*}
\]

(C.32)

et où \( D(t) \) est la fonction de déroulement de l’échelle temporelle donnée par

\[
D(t) = \int_0^t \beta(\tau) \, d\tau
\]

(C.33)

En supposant que le facteur \( \beta(t) \) est constant entre deux instants d’analyse \( t_s^i \) et \( t_s^{i+1} \) et égal à \( \beta_s \), la fonction \( D(t) \) peut se calculer par:

\[
D(t) = D(t_s^i) + \beta_s(t - t_s^i), \quad t_s^i \leq t < t_s^{i+1}
\]

(C.34)

avec \( D(t_s^i) = 0 \).

C.2.8.2 Modification de la fréquence fondamentale

L’objectif de la modification de la fréquence fondamentale est de changer la hauteur de la voix du locuteur tout en conservant le même rythme d’articulation et la même enveloppe spatiale. Connaissant les facteurs de modification de la fréquence fondamentale \( \alpha(t) \) et supposant que nous avons déjà trouvé l’instant de synthèse \( t_s^i \), l’instant de synthèse \( t_s^{i+1} \) est donné par

\[
t_s^{i+1} - t_s^i = \frac{1}{t_s^{i+1} - t_s^i} \int_{t_s^i}^{t_s^{i+1}} \frac{P(t)}{\alpha(t)} \, dt
\]

(C.35)

Nous allons utiliser aussi bien la notation \( n_s^i \) que \( t_s^i \) pour les instants de synthèse (pour l’analyse: \( n_s^i \) et \( t_s^i \)).
C.2.8.3 Modification de la fréquence fondamentale et de l’échelle temporelle

Afin de modifier en même temps l’échelle temporelle et la fréquence fondamentale il suffit de combiner les équations (C.31) et (C.35):

\[ t_{s}^{i+1} - t_{s}^{i} = \frac{1}{t_{v}^{i+1} - t_{v}^{i}} \int_{t_{v}^{i}}^{t_{v}^{i+1}} \frac{P(t)}{\alpha(t)} dt \]  \hfill (C.36)

C.2.8.4 Élimination des instants d’analyse répétés

En appliquant la technique ci-dessus pour le calcul des instants de synthèse, des instants d’analyse sont répétés plusieurs fois. Comme nous utiliserons le même ensemble de paramètres (amplitudes et phases) pour les trames répétées, ils n’évolueront pas d’une trame à l’autre. Ceci conduit à une qualité médiocre du signal synthétique. La solution consiste alors à éliminer les instants d’analyse redondant. Un exemple de modification de la fréquence fondamentale d’un facteur de 1.5 accompagné d’une élimination des instants d’analyse répétées, est donné Fig.C.9.

Figure C.9: Exemple de calcul des instants de synthèse pour une modification de la fréquence fondamentale d’un facteur 1.5 avec une élimination des instants d’analyse redondants.

C.2.8.5 Calcul des amplitudes et des phases aux nouvelles harmoniques

Dans le cas de la modification de la fréquence fondamentale le problème de réestimation des amplitudes et des phases aux nouvelles harmoniques se pose. Ce problème se pose pendant la synthèse où nous ne disposons que des amplitudes et des phases des harmoniques originales.
C.2. LE MODÈLE HNM POUR LA PAROLE

Une solution simple est l’interpolation linéaire entre les amplitudes des harmoniques originales. Nous avons utilisé une méthode proposé dans [CLM95] qui permet d’obtenir à partir de points fréquentiels discrets, une enveloppe spectrale continue qui passe par les amplitudes des harmoniques originales. L’enveloppe spectrale est représentée par les coefficients cepstraux discrets, $c_i$. Si on note $a_k$ l’amplitude de la kème harmonique $f_k$, les coefficients du cepstre discret $c = [c_0, \ldots, c_p]$ où $p$ est l’ordre du cepstre, sont obtenus en minimisant l’erreur suivante:

$$
\epsilon_r = \sum_{k=1}^{L} \left[ |\log a_k - \log |S(f_k; c)||^2 + \lambda \mathcal{R}[S(f; c)] \right].
$$

où $\mathcal{R}[S(f; c)]$ pénalise les variations rapides de l’enveloppe spectrale et $\lambda$ est le paramètre de régularisation. Il s’agit d’une méthode de moindres carrés régularisée. La fonction $\mathcal{R}[S(f; c)]$ est donnée par:

$$
\mathcal{R}[S(f; c)] = \int_{-1/2}^{1/2} \left[ \frac{d}{df} \log |S(f; c)| \right]^2 df.
$$

Pour mieux modéliser les basses fréquences, nous avons utilisé l’échelle Bark au lieu d’une échelle fréquentielle linéaire. La formule de conversion de Hz en Bark est donnée dans [ZT80]

$$
\begin{align*}
    f(Bark) &= 13 \arctan \left( \frac{f(Hz)}{1000} \right) \quad \text{if } f(Hz) \geq 605 \\
    f(Bark) &= 8.7 + 14.2 \log 10 \left( \frac{f(Hz)}{1000} \right) \quad \text{if } f(Hz) < 605
\end{align*}
$$

La Fig. C.10 montre un exemple d’une enveloppe spectrale estimée en utilisant la méthode proposée dans [CLM95] utilisant soit une échelle de fréquence linéaire (C.10(a)) soit une échelle Bark (C.10(b)).

Pour le calcul des phases, la solution de l’interpolation linéaire ne peut être appliquée, puisque les phases sont calculées modulo $2\pi$. On doit donc dérouler fréquemment la phase (phase unwrapping) tout en conservant sa continuité dans le domaine temporel. Les algorithmes proposés jusqu’à ce jour ne peuvent être utiles pour deux raisons: ils ne peuvent pas dérouler la phase quand on ne dispose que de peu de points fréquentiels (ce qui est notre cas) et parce qu’ils ne garantissent pas la continuité de la phase dans le domaine temporel. Nous avons donc développé un nouvel algorithme de déroulement de la phase qui garantit la continuité de la phase aussi bien dans le domaine fréquentiel que dans le domaine temporel.

Dans la première trame voisée la phase est déroulée dans le domaine fréquentiel en ajoutant de multiples entiers de $2\pi$ afin de conserver la pente fréquentielle $d\phi_k$ de la phase aussi douce que possible. $d\phi_k$ est défini comme

$$
\frac{d\phi_k}{\phi_{k+1} - \phi_k}
$$
où $k$ correspond à la $k$-ième harmonique. Le principe du déroulement de la phase dans le domaine fréquentiel est donné Fig.C.11(a). Pour les trames voisées suivantes, les phases sont déroulées en utilisant la pente de la phase de la trame précédente $d\phi_k$ et non la pente courante $d\phi_k^1$ (voir Fig.C.11(b)). Ainsi la continuité de la phase dans les domaines fréquentiel et temporel est garantie. Un exemple de déroulement de la phase pour 10 trames successive est donné

Figure C.11: Estimation de l'enveloppe spectrale. a) pour la première trame d'un segment voisé b) pour les trames suivantes.

Fig.C.12.

Ayant estimé les enveloppes continues fréquentielles des amplitudes et des phases, une
C.2. LE MODÈLE HNM POUR LA PAROLE

Figure C.12: Exemple de l’estimation de l’enveloppe de la phase et de sa déroulement.

simple interpolation entre les harmoniques originales nous fournit les amplitudes et les phases aux nouvelles harmoniques. Enfin, pour obtenir le signal synthétique modifié nous utilisons la même approche que pour la synthèse sans modification. La Fig.C.13 montre le diagramme de la synthèse avec des modifications prosodiques.

C.2.9 Exemple d’application

Dans cette section, les résultats de l’application du modèle HNM sur des signaux de parole sont présentés. La Fig.C.14 montre les résultats de l’algorithme du pitch (Fig.C.14(b) pour la fréquence fondamentale et Fig.C.14(c) pour la fréquence maximale du voisement) appliqué sur un signal de parole prononcé par un locuteur masculin (mot : “wazi waza”). En utilisant les valeurs estimées du pitch, les instants d’analyse sont synchrones à la fréquence fondamentale. L’estimation pour les deux parties, harmonique et bruitée, est faite comme expliquée plus haut: minimisation d’une erreur de moindres carrés pour la partie harmonique, et estimation des coefficients LPC et de la variance du signal original pour la partie bruitée. La synthèse est aussi synchrone à la fréquence fondamentale. La partie harmonique est obtenue en interpolant linéairement les amplitudes et les phases entre deux instants de synthèses. La partie bruitée est obtenue en filtrant un bruit blanc par le filtre de prédiction linéaire dont la sortie est multipliée par la variance du signal original. Si la trame synthétisée est voisée, nous filtrons le signal bruité par un filtre passe haut dont la fréquence de coupure est égale à la fréquence maximale locale du voisement. On applique ensuite l’enveloppe temporelle sur le signal bruité pour lui
Figure C.13: Diagramme de la synthèse avec des modifications prosodiques

Figure C.14: Estimation de la fréquence fondamentale et de la fréquence maximale de voisement sur un signal de parole prononcé par un locuteur masculin. (a) le signal original, (b) l’estimation du contour de la fréquence fondamentale et (c) l’estimation de la fréquence maximale de voisement.
C.2. LE MODÈLE HNM POUR LA PAROLE

donner la structure caractéristique de la partie bruitée. La Fig.C.15 montre un segment de la partie harmonique (Fig.C.15(a)) et de la partie bruitée synthétique (Fig.C.15(b)). La Fig.C.16 montre le même segment que celui de la Fig.C.15 pour le signal original (Fig.C.16(a)) et pour le signal synthétique (Fig.C.16(b)). Le signal synthétique est de haute qualité et difficilement discernable du signal original. Le modèle HNM a été testé sur une grande base de données avec des locuteurs masculins et féminins. Les résultats sont très satisfaisants. Le modèle HNM permet aussi de faire des modifications prosodiques de haute qualités avec des facteurs de modifications variant dans le temps. La Fig.C.17 présente un exemple de modification de l’échelle temporelle avec des facteurs de modification variant dans le temps. Les contours mélodiques des signaux originaux (trait continu) et synthétisé après la modification (trait solide) avec les facteurs donnés Fig.C.17(a). La Fig.C.18 présente la même chose mais pour une modification de la fréquence fondamentale. Ces deux figures mettent en évidence les propriétés des modifications prosodiques: dans le cas de modification de l’échelle temporelle le contour mélodique s’étend en conservant le même aspect. De même, pour la modification de la fréquence fondamentale, le contour mélodique se modifie tout en ayant la même durée. Deux exemples de modification de l’échelle temporelle et de la fréquence fondamentale, avec des facteurs constants, sont donnés Fig.C.19 et Fig.C.20. Même pour des facteurs de modification importants, le signal modifié ne présente aucun effet métallique (comme c’est le cas avec le TD-PSOLA).

Figure C.15: (a) Partie harmonique (b) Partie bruitée.
Figure C.16: (a) Un segment du signal original et (b) le segment correspondant du signal synthétique (partie harmonique + partie bruitée).

Figure C.17: (a) Exemple de facteurs de modification de l’échelle temporelle, variants dans le temps. (b) Contour mélodique original (trait interrompu) superposé avec le contour mélodique du signal synthétique après la modification de l’échelle temporelle (trait solide).
C.2. LE MODÈLE HNM POUR LA PAROLE

Figure C.18: (a) Exemple de facteurs de modification de la fréquence fondamentale (b) Contour mélodique original (trait interrompu) superposé avec le contour mélodique du signal synthétique après la modification de la fréquence fondamentale.

Figure C.19: (a) Signal original. (b) Signal synthétique après une modification de l’échelle temporelle avec un facteur de 1.3.
Figure C.20: (a) Signal original. (b) Signal synthétique après une modification de la fréquence fondamentale avec un facteur de 1.3

C.3 Modification du locuteur

Le changement de locuteur consiste à modifier un signal de parole prononcé par un locuteur d’une façon telle qu’il semble, à l’écoute, être prononcé par un locuteur différent. Le terme de modification (ou conversion) indique que l’objectif consiste à transformer les signaux de parole enregistrés par un locuteur, dit de référence, de telle façon qu’ils deviennent aussi proches que possible sur le plan acoustique de ceux enregistrés par un autre locuteur dit cible (ayant prononcé le même texte).

La modification du locuteur est fortement liée à la reconnaissance du locuteur (c’est à dire l’identification d’une personne à partir de sa voix), elle tire profit des progrès réalisés dans ce domaine au cours de la dernière décennie. Il a en effet été possible de mettre en évidence les principaux éléments qui permettent de distinguer les voix de locuteurs différents [Dod85][Fur86] [RS91]. Parmi les caractéristiques des signaux de parole liées à l’identité du locuteur on distingue en général:

1. Les caractéristiques spectrales

Pour des raisons notamment physiologiques les sons caractéristiques du langage ne sont pas produit de manière identique par tous les locuteurs. Ceci se traduit par des différences notables et systématiques au niveau de la forme du spectre du signal vocal. Ces différences spectrales jouent un rôle très important pour la reconnaissance du locuteur. À tel point
que la quasi-totalité des techniques actuelles de reconnaissance du locuteur se fondent sur la caractérisation de ces différences spectrales [Dod85][RS91].

2. Les caractéristiques prosodiques
D’autres indices forts de reconnaissance du locuteur sont liés aux variations dynamiques des paramètres prosodiques qui caractérisent l’élocution: fréquence fondamentale, vitesse d’élocution, énergie. Parmi ces derniers, l’élément qui semble prépondérant en ce qui concerne la reconnaissance du locuteur est la valeur moyenne de la fréquence fondamentale [RS91][Fur86].

La modification du locuteur est une application qui devient envisageable du fait de l’émergence de systèmes d’analyse/synthèse de signaux de parole de bonne qualité. En effet, la modification du locuteur implique des modifications fines des caractéristiques acoustiques du signal de parole. Pour cela, il est nécessaire de pouvoir extraire du signal des paramètres permettant d’avoir accès aux caractéristiques spectrales et prosodiques (étape d’analyse). Inversement, il faut être capable d’obtenir un signal de parole de bonne qualité à partir des paramètres modifiés (synthèse).

C.3.1 Applications

La modification du locuteur est un complément indispensable pour les systèmes de synthèse de parole fonctionnant par concaténation d’unités acoustiques préenregistrées (synthèse par diphones ou par demi- syllabes). En effet, grâce à la modification du locuteur, il est possible de synthétiser toute une palette de timbres de voix sans avoir à réenregistrer une base de données d’unités acoustiques pour chaque nouveau locuteur. L’utilisation de timbres de voix variés est importante afin de personnaliser les différents services utilisant la synthèse de parole. Les différents timbres de voix peuvent aussi être utiles pour distinguer les messages vocaux correspondant à différents niveaux de dialogue (commandes, information, messages d’erreur, etc.).

Pour les mêmes raisons, la modification du locuteur est amenée à jouer un rôle important dans le domaine des serveurs vocaux utilisant des messages préenregistrées. Dans les deux cas cités, la conversion de voix permet de diminuer le volume de données à stocker ainsi que d’éviter la multiplication des séances d’enregistrements. On peut aussi envisager des utilisations analogues dans le cadre d’applications multimédia intégrant images et parole préenregistrée.

La modification du locuteur peut trouver des applications dans des contextes tel que celui du doublage sonore de films. À plus long-terme, l’utilisation de ce type de techniques est envisagée dans le cadre de systèmes de traduction automatiques destinés à permettre les conversations.
téléphoniques entre locuteurs ne parlant pas la même langue (applications de téléphonie interprétée).

Enfin, il existe une forte interaction entre les domaines de la modification du locuteur et celui de la reconnaissance du locuteur. Ainsi les systèmes de modification du locuteur pourraient, par exemple, être utilisés pour évaluer la robustesse des systèmes de reconnaissance du locuteur. Par ailleurs, d’un point de vue méthodologique, les techniques employées (notamment pour tout ce qui concerne l’apprentissage de la transformation spectrale) sont très proches de celles utilisées en reconnaissance robuste de la parole (adaptation au locuteur ou à un environnement difficile).

C.3.2 État de l’art

La modification du locuteur (conversion de voix) est un domaine relativement nouveau: les premiers essais datent de moins de 10 ans, et ce sujet reste pour l’instant largement marginal par rapport aux autres thèmes liés au traitement de la parole (codage, reconnaissance, débruitage, etc.). L’essentiel des recherches dans ce domaine est actuellement concentré au Japon (laboratoires NTT, ATR, SONY) avec pour application principale la téléphonie interprétée. En Europe et aux États-Unis, on trouve principalement des travaux de recherche émanant du milieu académique.

D’une manière générale, la plupart des systèmes proposés jusqu’à maintenant utilisent des modèles de type-source-filtre [Val92][ANSK90] [SN91]. Tous travaillent également en alignant préalablement les données spectrales issues de la référence et celles provenant de la cible par des techniques d’alignement temporel automatique à la base de programmation linéaire (techniques dites de DTW).

Par ailleurs, l’essentiel de l’effort de recherche de la communauté parole porte sur la détermination d’une fonction de transformation spectrale qui est l’aspect le plus délicat. Les modifications prosodiques sont considérées comme un problème secondaire, en général les auteurs proposent de modifier uniquement la fréquence fondamentale de voisement [Val92][MA95] (en utilisant la technique non-paramétrique TD-PSOLA [ML95] qui permet d’obtenir très simplement des résultats de bonne qualité).

En ce qui concerne l’apprentissage de la modification spectrale on peut considérer que les premières solutions efficaces sont apportées par [SN91] et [ANSK90]. Dans ces deux derniers travaux, la transformation spectrale consiste à choisir une enveloppe spectrale issue d’un “dictionnaire” (ensemble discret) obtenu par quantification vectorielle (on parle de discrétisation de l’espace des enveloppes spectrales de la cible). Les techniques d’optimisation proposées pour estimer la fonction de choix de l’enveloppe spectrale cible diffèrent dans les deux articles.
C.3. MODIFICATION DU LOCUTEUR


[IS95] utilise aussi un espace d’enveloppes spectrales cibles continu obtenu par interpolation (somme pondérée) entre les données spectrales mesurées pour plusieurs locuteurs. L’optimisation porte sur les coefficients de la somme pondérée. Cette méthode est un peu particulière dans la mesure où elle nécessite l’enregistrement du vocabulaire complet par plusieurs locuteurs.

C.3.3 Description du système proposé

C.3.3.1 Procédure d’apprentissage

Par rapport aux systèmes mentionnés précédemment, la technique de modification du locuteur proposée présente principalement quatre aspects originaux:

1. **Utilisation de l’analyse/synthèse/modification HNM.**
   L’utilisation du modèle HNM dans le cadre de la modification du locuteur permet d’améliorer la qualité des signaux de parole convertis. De plus, par rapport à la technique utilisée dans [Val92], le modèle HNM permet de contrôler précisément à la fois la prosodie et l’enveloppe spectrale du signal synthétique, ce qui rend la modification du locuteur plus efficace.

2. ** Pré-classification probabiliste des données de référence**
   Lors de la phase d’apprentissage de la transformation spectrale, nous utilisons une clas-
sification “probabiliste” des données spectrales du locuteur de référence. Cette classification obtenue par un modèle de mélange de densités gaussiennes fournit des probabilités d’appartenance aux différentes classes qui varient continûment avec l’enveloppe spectrale. Contrairement à la méthode de quantification vectorielle utilisée habituellement, ce choix permet d’éviter la création de frontières peu significatives dans l’espace des enveloppes spectrales de référence. Cette classification probabiliste (combinée avec la fonction de transformation spectrale utilisée) permet de garantir une certaine homogénéité de la transformation spectrale.

3. Itération de l’alignement temporel
L’efficacité des procédures d’alignement temporel est limitée par le fait qu’elles se basent à la fois sur des différences temporelles (décalages entre les signaux de parole référence et cible) et des différences spectrales (différences intrinsèques entre les données spectrales issues de la référence et de la cible). Le principe retenu ici consiste à déterminer le chemin d’alignement optimal entre la référence et la cible en alignant les données transformées et les données cibles. L’utilisation des données transformées (qui sont donc spectralement proches des données cibles) permet de raffiner l’alignement temporel obtenu initialement avec les données de référence. La procédure complète (calcul de l’alignement temporel, optimisation de la fonction de transformation et transformation des données de référence) est donc itérée un certain nombre de fois.

4. Transformation spectrale
Un dernier aspect original du système proposé concerne la transformation spectrale. C’est à dire, d’une part la forme paramétrique adoptée pour la fonction de transformation (avec en particulier l’usage qui est fait des résultats de la pré-classification probabiliste), et d’autre part la procédure permettant d’optimiser la fonction de transformation. Ces deux points seront développés au paragraphe suivant.

C.3.3.2 Classification probabiliste des données de référence

La première étape de l’apprentissage de la fonction de transformation spectrale ne fait intervenir que les données de référence; il s’agit de la modélisation de celles-ci sous la forme d’un mélange de densités gaussiennes, GMM (Gaussian Mixture Model). Plus précisément, nous supposons que les vecteurs des paramètres cepstraux de référence sont des réalisations de variables aléatoires indépendantes entre elles et suivant une même loi de mélange caractérisée par la densité de probabilité suivante:

\[ p(x|\Theta) = \sum_{i=1}^{m} \alpha_i N(\mu_i, \Sigma_i) \]  \hspace{2cm} (C.41)
où $N(\mu_i, \Sigma_i)$ désigne la densité gaussienne de dimension $p$ (taille des vecteurs de paramètres cepstraux) caractérisée par son vecteur moyen $\mu_i$ et sa matrice de covariance $\Sigma_i$. $m$ est le nombre de composante du mélange. Les quantités $\alpha_i$ sont des paramètres scalaires (proportions du mélange) qui vérifient $\sum_{i=1}^{m} \alpha_i = 1$. Chaque densité gaussienne est associée à une classe notée $\omega_i$ dont la fréquence statistique est données par $\alpha_i$ [DH73].

L’apprentissage des paramètres du modèle à partir des données de référence se fait grâce à l’algorithme EM [DLJR77] qui permet d’atteindre de manière itérative l’estimation des paramètres au sens du maximum de vraisemblance. Dans le cas d’un mélange de densité gaussienne, chaque itération de l’algorithme EM implique deux étapes successives:

1. **Expectation (E-step)**
   Calcul d’une quantité intermédiaire qui se ramène à celui des probabilités conditionnelles d’appartenance à chaque classe $\omega_i$ pour chaque observation $x_i$:

   $$
P(\omega_i|x_i) = \frac{\alpha_i |\Sigma_i|^{-1/2} \exp \left[ -\frac{1}{2} (x_i - \mu_i)^T \Sigma_i^{-1} (x_i - \mu_i) \right]}{\sum_{j=1}^{m} \alpha_j |\Sigma_j|^{-1/2} \exp \left[ -\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \right]}$$ (C.42)

2. **Maximisation (M-step)**
   Réactualisation du modèle par maximisation de la fonction intermédiaire qui donne:

   Pour les proportions

   $$\hat{\alpha}_i = \frac{1}{N} \sum_{t=1}^{N} P(\omega_i|x_t)$$ (C.43)

   Pour les vecteurs moyens

   $$\hat{\mu}_i = \frac{\sum_{t=1}^{N} P(\omega_i|x_t) x_t}{\sum_{t=1}^{N} P(\omega_i|x_t)}$$ (C.44)

   Pour les matrices de covariance

   $$\hat{\Sigma}_i = \frac{\sum_{t=1}^{N} P(\omega_i|x_t) (x_t - \hat{\mu}_i) (x_t - \hat{\mu}_i)^T}{\sum_{t=1}^{N} P(\omega_i|x_t)}$$ (C.45)

Dans les équations ci-dessus, $N$ désigne le nombre de vecteurs de référence.

Dans ce type d’algorithme l’initialisation revêt un caractère particulièrement important compte-tenu du nombre important de paramètres inconnus. Les valeurs initiales des paramètres de chaque classe sont déterminées à partir de la classification des vecteurs de référence obtenue par quantification vectorielle (avec un dictionnaire de $m$ vecteurs).
C.3.3.3 Fonction de transformation

Dans la suite de la procédure d’apprentissage, les aspects du modèle des vecteurs de référence qui sont utilisés sont ses paramètres \((\alpha_i, \mu_i, \Sigma_i)\) pour \(i = 1, \ldots, m\) ainsi que les probabilités conditionnelles d’appartenance à chaque classe \((P(\omega_i|x_t)\) pour \(i = 1, \ldots, m\) et \(t = 1, \ldots, n\)).

Nous proposons la forme paramétrique suivante pour la fonction de transformation spectrale:

\[
\mathcal{F}(x_t) = \sum_{i=1}^{m} P(\omega_i|x_t) \left[ \nu_i + \Gamma_i \Sigma_i^{-1}(x_t - \mu_i) \right]
\] (C.46)

Les paramètres \(\mu_i\) (vecteurs) et \(\Gamma_i\) (matrices) de cette fonction sont optimisés de façon à minimiser la distance quadratique moyenne entre les vecteurs de référence transformés et les vecteurs cibles définie par

\[
\epsilon = \sum_{t=1}^{n} \|y_t - \mathcal{F}(x_t)\|^2
\] (C.47)

Nous pouvons distinguer trois types de fonctions de transformations\(^3\):

1. **Full conversion**
   - Aucune contrainte ne s’applique pas aux paramètres du GMM ou de la fonction de transformation.

2. **Diagonal conversion**
   - Les matrices \(\Sigma_i\) et \(\Gamma_i\) sont diagonales. Il faut noter que dans ce cas il y a une simplification (séparation pour chaque coordonnée) dans l’expression de la distance quadratique moyenne entre les vecteurs de référence transformés et les vecteurs cibles

   \[
   \epsilon = \sum_{t=1}^{n} \sum_{k=1}^{p} \|y_t^{(k)} - \mathcal{F}(x_t)^{(k)}\|^2;
   \] (C.48)

3. **VQ-type conversion**
   - Les matrices \(\Sigma_i\) et \(\Gamma_i\) sont zéros. Alors la fonction de transformations devient

   \[
   \mathcal{F}(x_t) = \sum_{i=1}^{m} P(\omega_i|x_t)\nu_i
   \] (C.49)

C.3.3.4 Optimisation de la fonction de transformation

Nous allons noter \(p_t(i)\) la probabilité conditionnelle, \(P(\omega_i|x_t)\), que \(x_t\) appartienne à la classe \(\omega_i\).

\(^3\)Nous utilisons ici les termes anglais.
1. **Full conversion**

Comme la fonction de transformation est linéaire, l’optimisation par rapport à ses paramètres est équivalente à la résolution d’un système d’équations linéaires au sens des moindres carrés:

\[
y_t = \sum_{i=1}^{m} p_t(i) \left[ \nu_i + \Gamma_i \Sigma_i^{-1}(\mathbf{x}_i - \mu_i) \right]
\]  

pour tout \( t = (1, \ldots, n) \). Il est facile de vérifier que ces équations peuvent s’écrire sous une forme matricielle:

\[
\mathbf{y} = \mathbf{P} \cdot \nu + \mathbf{D}_x \cdot \Gamma
\]

\[
\mathbf{y} = \left[ \begin{array}{c}
\mathbf{P} \\
\mathbf{D}_x
\end{array} \right] \cdot \left[ \begin{array}{c}
\nu \\
\Gamma
\end{array} \right],
\]

où \( \mathbf{y} \) est une matrice \( n \times p \) qui contient les vecteurs spectraux du locuteur cible:

\[
\mathbf{y} = \left[ y_1, \ldots, y_n \right]^T,
\]

\( \mathbf{P} \) est une matrice \( n \times m \) des probabilités conditionnelles \( p_t(i) \):

\[
\mathbf{P} = \left[ \begin{array}{cccc}
p_t(1) & p_t(2) & \cdots & p_t(m) \\
p_2(1) & p_2(2) & \cdots & p_2(m) \\
\vdots & \vdots & \ddots & \vdots \\
p_n(1) & p_n(2) & \cdots & p_n(m)
\end{array} \right],
\]

\( \mathbf{D}_x \) est une matrice \( n \times pm \) qui dépend des probabilités conditionnelles, des vecteurs du locuteur de référence et des paramètres du mélange de densités gaussiennes. Cette matrice est définie par blocs

\[
\mathbf{D}_x = \left[ \begin{array}{cccc}
p_t(1)(x_1 - \mu_1)^T \Sigma_1^{-1} & p_t(2)(x_1 - \mu_2)^T \Sigma_2^{-1} & \cdots \\
p_2(1)(x_2 - \mu_1)^T \Sigma_1^{-1} & p_2(2)(x_2 - \mu_2)^T \Sigma_2^{-1} & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
p_n(1)(x_n - \mu_1)^T \Sigma_1^{-1} & p_n(2)(x_n - \mu_2)^T \Sigma_2^{-1} & \cdots 
\end{array} \right]
\]

et les deux matrices

\[
\nu = \left[ \begin{array}{c}
\nu_1 \\
\nu_2 \\
\vdots \\
\nu_m
\end{array} \right]^T,
\]

\[
\nu = \left[ \begin{array}{c}
p_t(1) \\
p_t(2) \\
\vdots \\
p_t(m)
\end{array} \right]^T
\]

\[
\nu = \left[ \begin{array}{c}
p_2(1) \\
p_2(2) \\
\vdots \\
p_2(m)
\end{array} \right]^T
\]

\[
\nu = \left[ \begin{array}{c}
p_n(1) \\
p_n(2) \\
\vdots \\
p_n(m)
\end{array} \right]^T
\]
et
\[ \mathbf{\Gamma} = \left[ \mathbf{\Gamma}_1 : \mathbf{\Gamma}_2 : \cdots : \mathbf{\Gamma}_m \right]^T \]  \hspace{1cm} \text{(C.56)}

sont les paramètres de la fonction de transformation qu’il faut calculer. La forme de la
(C.51) est celle d’un problème classique de moindres-carrés pour lequel la solution est
donnée par les équations normales suivantes:
\[
\left( \begin{bmatrix} \mathbf{P}^T \\ \vdots \\ \mathbf{D}_x^T \end{bmatrix} \cdot \begin{bmatrix} \mathbf{P} : \mathbf{D}_x \\ \mathbf{\hat{f}} \end{bmatrix} \right) \cdot \begin{bmatrix} \mathbf{\hat{\nu}} \\ \mathbf{\hat{\mu}} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^T \\ \vdots \\ \mathbf{D}_x^T \end{bmatrix} \cdot \mathbf{y} \]  \hspace{1cm} \text{(C.57)}

ou
\[
\begin{bmatrix} \mathbf{P}^T \mathbf{P} : \mathbf{P}^T \mathbf{D}_x \\ \vdots : \vdots \\ \mathbf{D}_x^T \mathbf{P} : \mathbf{D}_x^T \mathbf{D}_x \end{bmatrix} \cdot \begin{bmatrix} \mathbf{\hat{\nu}} \\ \mathbf{\hat{\mu}} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^T \mathbf{y} \\ \vdots \\ \mathbf{D}_x^T \mathbf{y} \end{bmatrix} \]  \hspace{1cm} \text{(C.58)}

Comme la matrice qu’il faut inverser est symétrique et définie positive, ces équations
peuvent se résoudre facilement en utilisant la méthode de décomposition de Cholesky.
Il faut par contre noter le problème qui peut se poser du fait de la dimension de cette
matrice: \((m + m \times p)^2\). Ainsi, si la dimension \(p\) des vecteurs vaut 20 et le nombre \(m\)
de composantes du mélange de densités gaussiennes vaut 128 nous avons une matrice de
2688 \times 2688 éléments!.

2. Diagonal conversion

Dans le cas de la conversion diagonale les valeurs optimales sont obtenues pour chaque
cordonné séparée par la formule suivante (pour la \(k\)ème cordonnée)
\[
\begin{bmatrix} \mathbf{P}^T \mathbf{P} : \mathbf{P}^T \mathbf{D}_x^{(k)} \\ \vdots : \vdots \\ \mathbf{D}_x^{(k)T} \mathbf{P} : \mathbf{D}_x^{(k)T} \mathbf{D}_x^{(k)} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{\hat{\nu}}^{(k)} \\ \mathbf{\hat{\mu}}^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^T \mathbf{y}^{(k)} \\ \vdots \\ \mathbf{D}_x^{(k)T} \mathbf{y}^{(k)} \end{bmatrix} \]  \hspace{1cm} \text{(C.59)}

où maintenant \(\mathbf{D}_x^{(k)}\) est une matrice \(n \times m\) définie par
\[
\mathbf{D}_x^{(k)} = \begin{bmatrix} p_1(1)\sigma_1^{(k)-1}(x_1^{(k)} - \mu_1^{(k)}) & p_1(2)\sigma_2^{(k)-1}(x_1^{(k)} - \mu_2^{(k)}) & \cdots \\ p_2(1)\sigma_1^{(k)-1}(x_2^{(k)} - \mu_1^{(k)}) & p_2(2)\sigma_2^{(k)-1}(x_2^{(k)} - \mu_2^{(k)}) & \cdots \\ \vdots & \vdots & \vdots \\ p_n(1)\sigma_1^{(k)-1}(x_n^{(k)} - \mu_1^{(k)}) & p_n(2)\sigma_2^{(k)-1}(x_n^{(k)} - \mu_2^{(k)}) & \cdots \end{bmatrix} \]  \hspace{1cm} \text{(C.60)}

\[
\begin{bmatrix} p_1(m)\sigma_m^{(k)-1}(x_1^{(k)} - \mu_m^{(k)}) \\ p_2(m)\sigma_m^{(k)-1}(x_2^{(k)} - \mu_m^{(k)}) \\ \vdots \\ p_n(m)\sigma_m^{(k)-1}(x_n^{(k)} - \mu_m^{(k)}) \end{bmatrix} \]
3. VQ-type conversion

La solution pour ce type de conversion est donnée par

\[ \nu^{(k)} = (P^T P)^{-1} P^T y^{(k)} \]  (C.61)

pour la \( k \)ème cordonnée.

La transformation spectrale est appliquée à des paramètres représentant l’enveloppe spectrale du signal de manière plus concise: nous avons utilisé le cepstre discret régularisé [CLM95] que nous avons déjà utilisé pour recalculer les amplitudes d’harmoniques pour les nouveaux harmoniques dans le cas de la modification de la fréquence fondamentale. L’ordre du cepstre utilisé était 20 et le cepstre était estimé sur l’échelle Bark normalisée. Conformément à ce qui a été dit au paragraphe précédent, l’apprentissage se fait sur des données alignées temporellement. L’algorithme de calcul de l’alignement temporel utilisée est la DTW avec points d’ancrage développée au CNET Lannion. Afin de permettre l’alignement temporel, le calcul des paramètres spectraux s’effectue, durant la phase d’apprentissage, à des intervalles de temps réguliers (10ms- mode d’analyse dit asynchrone). Notez aussi que le premier coefficient du cepstre a été exclu de l’alignement temporel. La Fig.C.21 montre le diagramme de l’apprentissage de la fonction de transformation \( F() \).

![Diagramme de l'apprentissage](image_url)

Figure C.21: Diagramme pour l’apprentissage de la fonction de transformation.

3.3.3.5 Transformation de la partie bruitée

Le procédure développée aux paragraphes précédents ne s’applique que pour la partie harmonique. Pour transformer la partie bruitée deux filtres correctifs sont estimés un pour les trames voisées et un pour les non-voisées. Les coefficients de ces filtres sont estimés par la transformée de Fourier inverse sur la différence de périodogramme moyenne de la référence et
Figure C.22: Diagramme pour l’estimation de filtres correctifs.

de la cible. Cette procédure est présentée Fig.C.22. La Fig.C.23 montre un exemple des filtres correctifs.

Figure C.23: Filtres correctifs pour la transformation de la partie bruitée. En trait solide est la densité spectrale de puissance (dsp) de la référence et en trait interrompu la dsp de la cible.

C.3.3.6 Schéma de principe du système de modification du locuteur

Le système de modification du locuteur se compose de trois parties principales:

- L’analyse/synthèse HNM qui permet d’effectuer des modifications de signaux de parole de haute qualité. Pour la modification du locuteur le HNM fonctionne en mode pitch-synchrone.


C.3. MODIFICATION DU LOCUTEUR

- La transformation spectrale qui a pour rôle de modifier le timbre du signal de parole. Elle est appliquée à des paramètres représentant l’enveloppe spectrale (cepsre discret). Les filtres correctifs, pour la transformation de la partie bruissée, sont directement appliqués à la synthèse (à la sortie des filtres estimés pendant l’analyse).

- Les modifications prosodiques permettent de compléter la modification du locuteur. Actuellement, la fréquence fondamentale et la vitesse d’élocution du locuteur de référence sont modifiées de telle façon qu’elles soient, en moyenne, semblables à celles du locuteur cible. Lorsque l’on dispose de la même phrase test prononcée par les deux locuteurs, il est possible de recopier intégralement les variations de la fréquence fondamentale et de la vitesse d’élocution.

La Fig.C.24 montre le diagramme du système proposé pour la modification du locuteur.

Figure C.24: Diagramme d’étape de conversion $t_a^i$: instants d’analyse, $t_s^i$: instants de synthèse.
Bibliography


BIBLIOGRAPHY


[HMC89] C. Hamon, E. Moulines, and F. Charpentier. A diphone Synthesis System based on
time domain prosodic modification of Speech. *Proc. IEEE Int. Conf. Acoust., Speech,


[IS88] K. Itoh and S. Saito. Effects of acoustical feature parameters on perceptual speaker


[IS95] N. Iwahashi and Y. Sagisaka. Speech spectrum conversion based on speaker interpolation
and multi-functional representation with weighting by radial basis function networks.
*Speech Communication*, 16(2):139–151, February 1995.

[JHH84] C. C. Johnson, H. Hollien, and J. W. Hicks. Speaker identification utilizing selected


[Kli87] F. Klingholz. The measurement of the Signal-to-Noise Ration (SNR) in Continuous

[KS95] H. Kuwabara and Y. Sagisaka. Acoustic characteristics of speaker individuality:

[LaR75] C. LaRiviere. Contributions of fundamental frequency and formant frequencies to


[LC91] M.Y. Liberman and K.W. Church. Text Analysis and Word Pronunciation in Text-


