



WNSP05

Heraklion
Crete, Greece
20-23 September
2005

Tuesday, September 20

- 9:00-10:15 Registration
- 10:15-10:30 Welcome
- 10:30-11:15 **Invited Talk: *Alpha-Stable Distributions: Theory & Applications***
Panos Tsakalidis
- 11:15-12:00 **Invited Talk: *Multiple Description Coding for Speech Signals***
Yannis Agiomyrgiannakis
- 12:00-12:15 *Coffee Break*
- 12:15-13:00 ***Method for exact performance of forward and Inverse Wavelet Transform in Real Time***
Pavel Rajmic
- 13:00-15:00 *Lunch*
- 15:00-15:45 ***Enhancement of speech from noisy background, using single-channel method of spectrogram mapping***
Zdenek Smekal
- 15:45-16:30 ***The Information Transmitted by the Verbal and Non Verbal Communication Modes on the Emotional States: Some perceptual Data***
Anna Esposito
- 16:30-17:00 *Coffee Break*
- 17:00-17:45 ***Video Conference: The set of logic context-based features for the classification of segmentation instants of speech signal***
Arimantas Raskinis



WNSP05

Heraklion
Crete, Greece
20-23 September
2005

Wednesday, September 21

- 9:30-10:15 **Invited Talk: Voice Transformation.**
Thanassis Mouchtaris
- 10:15-11:00 **Making Speech Synthesis Sound More Human: Introducing Variation.**
Eric Keller
- 11:00-11:30 *Coffee Break*
- 11:30-12:15 **Cepstral Liftering Techniques for Voice Quality Assessment**
Peter Murphy
- 12:15-13:00 **Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals.**
Thierry Dutoit
- 13:00-15:00 *Lunch*
- 15:00-15:45 **A Review of Advanced Methods for Glottal Inverse Filtering.**
Peter Murphy
- 15:45-16:30 **A Review of Methods for Detecting Audible Discontinuities in Speech Synthesis.**
Yannis Pantazis
- 16:30-16:45 *Coffee Break*
- 16:45-17:30 **Support Vector Machines in Speech Recognition.**
Dario Martin Iglesias
- 17:30-18:15 **Towards Automatic Speech Recognition in Adverse Environments.**
Nassos Katsamanis



WNSP05

Thursday, September 22

- 9:30-10:15 ***Wave-shaping models of cyclic signals.***
Jean Schoentgen
- 10:15-11:00 ***From black boxes to gray boxes: A look behind the curtain of nonlinear speech models***
Gernot Kubin
- 11:00-11:30 *Coffee Break*
- 11:30-11:45 *Transfer to Knossos Palace*
- 11:45-13:15 *Knossos Palace*
- 13:15-13:30 *Transfer to Archanes Village*
- 13:30-16:30 *Lunch – Relax – Visit Archanes*
- 16:30-16:45 *Transfer to Boutari Winery*
- 16:45-18:00 *Boutari Winery*
- 18:00-18:30 *Back to Heraklion*
- 21:00-23:30 *Concert*



WNSP05

Heraklion
Crete, Greece
20-23 September
2005

Friday, September 23

- 9:30-10:15 ***Invited Talk: Analysis of clusters of correlated activity in fMRI data by Hopfield Neural Networks.***
Marotessa Voultzidou
- 10:15-11:00 ***Non Linear Predictive Models: Overview and possibilities in speaker recognition.***
Marcos Faundez Zanuy
- 11:00-11:30 *Coffee Break*
- 11:30-12:15 ***Voice disguise and automatic detection.***
Guido Aversano
- 12:15-13:00 ***Audio-visual forgery in identity verification.***
Gerard Chollet
- 13:00-15:00 *Lunch*
- 15:00-15:45 ***Automatic Speaker Verification: state of the art and current issues.***
Dijana Petrovska
- 15:45-16:15 ***A New Approach for Speech Feature Extraction Based on Genetic Algorithms.***
Christophe Charbuillet
- 16:15-16:30 *Coffee Break*
- 16:30-17:30 *Meeting COST277*
- 17:30-18:00 *Updates regarding the proposals for New COST actions*

Method for Exact Performance of Forward and Inverse Wavelet Transform in Real Time

Pavel Rajmic

Department of Telecommunications,
Faculty of Electrical Engineering and Communication Technologies,
Brno University of Technology,
Purkyňova 118, 61200 Brno, Czech Republic
`rajmic@feec.vutbr.cz`

Abstract. The new method of segmented wavelet transform (SegWT) makes it possible to compute the discrete-time wavelet transform of a signal segment-by-segment. This in fact means that the method could be utilized for wavelet-type processing of a signal in “real time”, or in case we need to process a long signal (not necessarily in real time), but there is insufficient computational memory capacity for it (for example in the signal processors). Then it is possible to process the signal part-by-part with low memory costs by the new method.

The method is suitable also for the speech processing, e.g. denoising the speech signal via thresholding the wavelet coefficients or for speech coding.

In the paper, the principle of the forward segmented forward wavelet transform is explained and the algorithm is described in detail, with objective illustrations. The latest findings about the algorithm of the inverse segmented wavelet transform are also presented – for example we have found that this stage of the algorithm inevitably introduces a time-lag of one segment’s length, which was not the problem of the forward stage.

Enhancement of speech from noisy background, using single-channel method of spectrogram mapping

Zdenek Smékal, Petr Sysel, Ivan Koula

Department of Telecommunications, Brno University of Technology,
Purkynova 118, 612 00 Brno, Czech Republic,
E-mail: smekal@feec.vutbr.cz

Methods for the enhancement of useful audio signal (speech or musical signal) from interfering background such as low-frequency noise, high-frequency noise, periodic or impulse interference, are of much importance in particular in mobile communications, in voice over the Internet transmission, in various means of transport, etc.

Methods for suppressing undesirable interference in speech signals can be divided into two groups. They are either the single-channel methods (the signal is detected by one microphone) or the multi-channel methods (several microphones or an array of sensors). At present the single-channel methods are given preference in practice since they are considerably simpler as regards the realization of detection elements or the implementation of algorithms. The existing single-channel methods make use of methods for spectral subtraction, adaptive filtering with different modifications of type LMS algorithm, the wavelet transform or digital filter bank, etc. In most cases these methods assume the presence of additive noise, whose properties are close to those of additive noise. A number of firms manufacturing mobile phones and hands-free sets apply the RASTA (RelAtive SpecTrAl) method [1]. This method uses the filtering of variable temporal trajectories of the harmonic components of the Fourier spectrum of speech signal by a low-pass or band-pass digital filter. The type of filter is chosen on the basis of the useful signal-to-noise ratio. The filters suppress the modulation components below 1 Hz and above 16 Hz. Problems arise when noise components are in a frequency range that corresponds to the speech modulation spectrum, and by their magnitude are comparable with the level of speech signal.

In the Department of Telecommunications of Brno University of Technology a novel single-channel method has been developed that makes use also of the short-time spectra of noisy speech signal, similar to the RASTA method. But there is a difference in the way that noise and interference are suppressed. In the time-frequency Fourier representation the regions of speech activity are found out and the interference threshold is selected adaptively. A binary mask is then made, which is used to suppress noise and interference [2, 3]. The method can suppress stationary and non-stationary noise and interference. It is known that speech contains, in unvoiced segments in particular, also noise that must be preserved in order to maintain good intelligibility and satisfactory quality of the reconstructed speech signal. The basic problem of the proposed method consists in determining a threshold that will decide which noise and interference components are undesirable and, vice versa, which have to be preserved since they are part of the speech signal. In the paper, the fundamental principle of the proposed method and ways of determining the adaptive threshold for different types of noise and interference will be described. In conclusion, the ways and problems of implementing the method in type VLIW digital signal processor (TMS 320C6711) will also be discussed.

References:

- [1] Hermansky, H., Wan, E.A., Avenando, C.: Speech Enhancement Based On Temporal Processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1995, pp.405-408.
- [2] Smékal, Z et al. "Non-Linear Methods of Speech Enhancement". Partial research report on the solution of international project No OC277.002 for the year 2003, Dep. of Telecommunications, Brno University of Technology, 2003, 92 pages.
- [3] Smékal, Z et al. "Non-Linear Methods of Speech Enhancement". Partial research report on the solution of international project No OC277.002 for the year 2004, Dep. of Telecommunications, Brno University of Technology, 2004, 107 pages.

The Information Transmitted by the Verbal and Non Verbal Communication Modes on the Emotional States: Some Perceptual Data

Anna Esposito^(a,b),

^(a) Dipartimento di Psicologia, Seconda Università di Napoli, Via Vivaldi 43, Caserta, Italy

^(b) Istituto Internazionale per gli Alti Studi Scientifici (IIASS), Via Pellegrino 19, Vietri, Salerno, Italy
e-mail: iiass.annaesp@tin.it; anna.esposito@tin.it

Abstract

Synaesthesia is a singular sensorial phenomenon (from Greek, *syn* = Together + *aisthesis* = Perception) where a sensation is produced in one sensorial modality when a stimulus is applied to another sensorial modality, as when the hearing of a certain sound induces the visualization of a certain colour. Such a phenomenon is one of the most intriguing examples of the crossing of sensory systems. We will love to say more about synaesthesia but this talk will not discuss on it. The word “synaesthesia” has been used here to introduce the idea of the cross-modal interaction of the sensory systems that mainly happen when we feel a given emotional state. In fact, in expressing our emotional states, it seems that we are trying to generate in our interlocutor a synaesthetic experience since he is receiving the information of our feeling through different sensorial channels, and this bring to the focus of the talk that will discuss on the amount of emotional information transmitted by the several communication modes. These modes are referred to as the verbal (the semantic content of our message) and non verbal (the gesture, the gaze, the tonal expression) modalities. From an engineering point of view a such transmission of the information content is redundant, since the same information is transferred through several channels. How much information about the speaker emotional state is transmitted by each channel and which channel play the major role in transferring such information? This work try to answer the above questions through a perceptual experiment that evaluates the subjective perception of different emotional states in the single (either visual or auditory channel) and the combined channels (visual and auditory). Results seem to show that, taken separately, the semantic content of the message and the visual content of the message bring the same information amount of the combined channels, suggesting that each channel performs a robust encoding of the emotional features that results very helpful in recovering the perception of the emotional state when one of the channel is degraded by the noise.

The set of logic context-based features for the classification of segmentation instants of speech signal

Dr. Arimantas Raškinis, Dainora Kuliešienė
Vytautas Magnus University, Kaunas, Lithuania

Abstract

In this work, the problem of data driven speech signal segmentation is addressed. Our approach is based on the symbolic machine learning (rule induction) techniques and on the idea of two-level feature set. The idea of the two level feature set states that given some particular ML system any good feature set capable of adaptation to specificity of the domain should be constructed in two steps. The so called first-order features represent transformations of learning material resulting in characteristics relevant to all learning samples regardless of their class assignment. The second-order features are derivative characteristics of the first order features. They are built as to achieve optimum class discrimination.

In this paper, we present an overview of recognition systems based on the two level feature set approach that solve a variety of classification tasks including discrimination among geometrical figures, arithmetic objects, and segments of a vocal signal.

In this paper, the problem of speech signal segmentation is stated as a machine learning problem based on the two-level feature set. The set of the second order features describing candidate segment boundaries are *context-based*. It describes not only instantaneous (inside a frame) signal characteristics at a candidate time but include characteristics of some other time instants as well.

The system was tested on real acoustic data. The rule learning algorithm RIPPER k was used. Rule sets for discriminating voiced/unvoiced speech segments and for spotting silence present in speech recordings were built. Error rate varied between 9-14% for various parameter settings. The relevance of automatically induced rules is investigated. Future research directions and perspectives are discussed.

Contact person: dr. Arimantas Raškinis
arimantas_raskinis@fc.vdu.lt

Making Speech Synthesis Sound More Human: Introducing Variation

Eric Keller, LAIP, IMM Lettres, University of Lausanne, Switzerland
eric.keller@unil.ch

Abstract

Current speech synthesis systems show increasing technical perfection, but are still recognized as non-human when producing extended passages. This perception is a major hindrance in the general acceptance of speech synthesis technology in areas such as film and television. We propose a review of issues that contribute to the "perceived humanness" of speech. Central to this review will be the issue of regularity and variation, particularly with respect to an utterance's temporal structure. We have new evidence to show that human speakers show considerable interspeaker agreement with respect to the placement of strong vowel onsets, but that interspeaker agreement is much less for weak vowel onsets. Strength or weakness is determined automatically from the acoustic signal, and is thus likely to correlate with both motoric and perceptual saliency within the utterance. This suggests that current statistical or neural network models for temporal structuring of speech may well be fundamentally flawed. Instead of a rigid, totally predictable structure, temporal prediction systems should probably provide a set of main "temporal anchor points" within the utterance, and introduce "motivated variation" for the remaining aspects of temporal structure. We will pass in review the different types of psycholinguistic and pragmatic events that can motivate such variation, and we will present a modified linear prediction system that can handle these new requirements.

Cepstral Liftering Techniques for Voice Quality Assessment

Specific regions of the cepstrum of voiced speech are selected for analysis with a view to determining a harmonics-to-noise ratio (HNR) estimate. Firstly, the low quefrequency portion of the cepstrum is analysed. The Fourier transformed liftered cepstrum approximates a noise baseline from which the harmonics-to-noise ratio is estimated. The present study highlights the manner in which the cepstrum-based noise baseline estimate is obtained, essentially behaving like a moving average filter applied to the power spectrum for voiced speech. As such, the noise baseline, which is taken to approximate the noise excited vocal tract, is shown to be influenced by the window length and the shape of the glottal source spectrum. Two approaches (a harmonic pre-emphasis technique and a symmetric baseline technique) are implemented to overcome the glottal source and window length dependences. The results indicate accurate HNR estimation using the new methods.

The high quefrequency region of the cepstrum has been investigated with a view to providing a correlate of voice quality. Specifically, the high quefrequency region is characterised by rahmonics peaks spaced at the pitch period and its sub-multiples. It is known that the amplitude of the first rahmonic, R1 has a correspondence with the richness of the harmonic spectrum for voiced speech, however a formal description has, to date, remained absent. The present study provides a theoretical description of rahmonic analysis of voiced speech containing aspiration noise and hence derives a definition for R1. It is shown that R1 is proportional to a geometric mean harmonics-to-noise ratio (gmHNR), where the gmHNR is defined as the mean of the individual spectral (i.e. at specific frequency locations) harmonics-to-noise ratios in dB. The technique is tested using synthesized voice signals. R1 is found to be sensitive to all forms of waveform aperiodicities and is shown to depend on analysis window length and fundamental frequency, f_0 . A pitch-synchronous, harmonic-limited spectral analysis is implemented to alleviate the window length/ f_0 dependence of the measure. A discussion of the results of previous studies employing R1 in the analysis of human voice signals is given in light of the definition and present results on synthesis.

Title: Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals

Authors: Baris Bozkurt and Prof. Thierry Dutoit,

Address: Circuit Theory and Signal Processing Laboratory (TCTS Lab), Faculte Polytechnique de Mons, Parc Initialis, 1, Copernic Avenue, B7000 Mons, Belgium, <http://tcts.fpms.ac.be>, {bozkurt,dutoit}

@tcts.fpms.ac.be

This study presents new chirp group delay processing techniques for analysis of resonances of a signal. The theoretical part of the study is based on analysis of the Zeros of Z-Transform (ZZT), which is an all-zero representation of the z-transform of the signal. Given the two spectral representations, effective algorithms are developed for: source-tract decomposition of speech, glottal flow parameter estimation, formant tracking and feature extraction for speech recognition.

The ZZT representation is mainly important for theoretical studies. Studying the ZZT of a signal is essential to be able to develop effective chirp group delay processing methods. Therefore, first the ZZT representation of the source-filter model of speech is studied for providing a theoretical background. The ZZT of windowed speech signals is also studied since windowing cannot be avoided in practical signal processing algorithms and the effect of windowing on ZZT representation is drastic. We show that separate patterns exist in ZZT representations of windowed speech signals for the glottal flow and the vocal tract contributions and review our source-tract decomposition method based on this property.

We define chirp group delay as group delay calculated on a circle other than the unit circle in z-plane. The need to compute group delay on a circle other than the unit circle comes from the fact that group delay spectra are often very noisy and cannot be easily processed for formant tracking purposes (the reasons are explained through ZZT representation). In this study, we propose methods to avoid such problems by modifying the ZZT of a signal and further computing the chirp group delay spectrum. New algorithms based on processing of the chirp group delay spectrum are developed for formant tracking and feature estimation for speech recognition. The proposed algorithms are compared to state-of-the-art techniques. Equivalent or higher efficiency is obtained for all proposed algorithms.

A Review of Advanced Methods for Glottal Inverse Filtering

Jacqueline Walker, Peter Murphy

Department of Electronic and Computer Engineering, University of Limerick, Limerick,
Ireland.

Glottal inverse filtering is a technique used to derive the glottal waveform during voiced speech. Closed phase inverse filtering (CPIF) is a common approach for achieving this goal. During the closed phase there is no input to the vocal tract and hence the impulse response of the vocal tract can be determined through linear prediction. However, a number of problems are known to exist with the CPIF approach. This review paper briefly details the CPIF technique and highlights certain associated theoretical and methodological problems. An overview is then given of advanced methods for IF; pole-zero modelling, model based, adaptive, higher order statistics and cepstral approaches are examined. The advantages and disadvantages of these methods are highlighted.

In addition to developing an optimum glottal identification algorithm some remaining issues concerning glottal inverse filtering include: evaluating what is considered to be a good result; what characteristics are perceptually relevant?; what characteristics are physically relevant? parametrization of the glottal volume velocity and the voice source (derivative glottal volume velocity); source-tract interaction; secondary excitation; recording conditions; time-varying glottal open phase transfer function. These issues are reviewed in the current study and a guide to applying a successful inverse filtering strategy is presented.

Support Vector Machines in Speech Recognition

José Miguel García Cabellos, Darío Martín Iglesias, Fernando Díaz de María, Jaume Padrell, Ascensión Gallardo Antolín and Carmen Peláez Moreno

Signal Theory and Communications Department
EPS-Universidad Carlos III de Madrid
Avda. de la Universidad, 30, 28911 – Leganés (Madrid), SPAIN

Hidden Markov Models (HMMs) are, undoubtedly, the most employed core technique for Automatic Speech Recognition (ASR). During the last decades, research in HMMs for ASR has brought about significant advances and, consequently, the HMMs are currently accurately tuned for this application. Nevertheless, we are still far from achieving high-performance ASR systems. Some alternative approaches, most of them based on Artificial Neural Networks (ANNs), were proposed during the last decade. Some of them tackled the ASR problem using predictive ANNs, while others proposed hybrid (HMM-ANN). However, despite some achievements, none of these approaches could outperform the results obtained with HMMs and, nowadays, the preponderance of Markov Models is a fact.

In the last decade, however, a new tool appeared in the field of machine learning that have proved its capability to overcome many of the problems of techniques as ANNs. The Support Vector Machines (SVMs) are effective discriminant classifiers capable of maximizing the error margin. As opposed to ANNs, they have the advantage of being capable to deal with samples of a very higher dimensionality. Also, their convergence to the minimum of the associated cost function is guaranteed as a simple problem of quadratic programming (QP). Besides, instead of only minimizing the empirical risk, they also try to minimize the “structural risk”, being the solution a compromise between empirical error and generalization capability.

These characteristics have made SVMs very popular and successful in many fields of application. Nevertheless, in order to use them in a problem of speech recognition, some limitations must be overcome. One of them is the number of training samples they can deal with that, in spite of the apparition of techniques as Sparse SVM, is still limited to a few thousands. Another problem of SVMs is that, in their original formulation, they are restricted to work with input vectors of fixed dimension (although nowadays there are some solutions to cope with this problem, as we will see). Finally, another limitation is that SVMs *only* classify, but they don't give us a reliable measure of the probability of the correctness of the classification. This, can cause problems in recognition, where without a concrete value of probability we can't carry out some algorithms as Viterbi, to look for the most probable sequence of recognition units.

In this chapter we will introduce SVMs in the speech recognition problem and will make a review of the current state-of-the-art techniques in this field. Specifically, we will study the use of HMM-guided segmentation to produce the fixed-size vectors needed for traditional SVMs. Afterwards we will explore more sophisticated techniques based on the use of different kernels capable to deal with sequences of different length. Among them is the DTAK kernel, simple and effective, which rescues an old technique of speech recognition: Dynamic Time Warping (DTW). After this, we will see some techniques used in text classification and biology, but useful in our problem, based on *score spaces*, such as the Fisher score.

Towards Automatic Speech Recognition In Adverse Environments

D. Dimitriadis, N. Katsamanis, P. Maragos, G. Papandreou and V. Pitsikalis

Email: [ddim,nkatsam,maragos,gpapan,vpitsik]@cs.ntua.gr

1. Noise Robustness

1.1. Modulation Analysis

Motivated by strong evidence for the existence of amplitude and frequency (AM-FM) modulations in speech signals, a speech resonance can be modeled as

$$r_i(t) = A_i \exp\left(\int_0^t b_i(\tau)d\tau\right) \cos\left(2\pi \int_0^t f_i(\tau)d\tau\right) \quad (1)$$

where $b_i(t)$ and $f_i(t)$ are the instantaneous modulating signals. Correspondingly the total speech signal can be considered as a superposition of a small number of such AM-FM signals. The instantaneous Teager energy Ψ of such resonance signal is given by

$$\Psi[r_i(t)] \approx A_i^2 \exp\left(2 \int_0^t b_i(\tau)d\tau\right) 4\pi^2 f_i^2(t) \quad (2)$$

The modulating features based on such signals are proved to be both robust to noise and rich in acoustic information [2]. Herein, we are presenting improved recognition results on noisy tasks when compared to the typical linear MFCC features.

1.2. Dynamical Systems, Fractal Analysis

1.2.1. Phase Space Reconstruction

A speech signal segment $s(n)$ can be thought of as a 1D projection of the unknown *multidimensional* phase space of the speech production system. Possibly, this projection is responsible for a loss of information. By a reverse procedure a multidimensional phase space is reconstructed according to the *embedding* theorem [1]. The reconstructed space $Y(n)$ is formed by samples of the original signal delayed by multiples of a constant time delay T_D and constructs a trajectory in the multidimensional (D_E) space $Y(n) = [s(n), s(n + T_D), \dots, s(n + (D_E - 1)T_D)]$, that shares invariant characteristics with the original phase space. Thus, by studying the reconstructed space we can possibly uncover useful information about the original unknown dynamical system. In the unfolded phase space we measure invariant quantities of the set (attractor) that are conserved from the original phase space. Such measures are the fractal dimensions which correspond to the number of active degrees of freedom and the underlying system complexity.

1.2.2. Robust Processing

Robust processing of speech signals is indispensable for any application that is meant to work in real environments. Denoising in the reconstructed phase space is done in ‘agreement’ to the assumed system dynamics. Common methods for robust processing of speech signals especially in the framework of speech recognition at the feature level, are filterbank analysis, modulation based analysis [2], auditory inspired processing [5],

methods influenced by perceptual ideas [3]. In this work we present an alternative denoising scheme from the dynamical systems’ point of view for speech processing. Based on ideas mentioned above we employ methods to filter the signal on the reconstructed space assuming that the multidimensional signal is closer to the dynamics of the speech production system, compared to the scalar signal.

2. User Robustness

In an effort to achieve user robustness we applied and evaluated state of the art speaker adaptation/normalization methods, like Maximum Likelihood Linear Regression, Maximum A Posteriori Adaptation and Vocal Tract Length Normalization in various scenarios (online, offline, fast adaptation, non-native speech). The results presented exhibit important performance improvement. Our research mainly focused on speaker normalization techniques at the signal level and their proper combination with transformation-based adaptation (MLLR, MAP). Such techniques achieve the proper modification of the incoming utterance based on speaker-dependent properties so that the extracted features may be even more speaker-independent, better suiting to speech recognition models. We tried to apply concepts and algorithms from nonlinear and time-varying speech processing in this direction.

3. Audio-Visual Automatic Speech Recognition

The design of robust audio-visual ASR systems, which perform better than their audio-only analogues in all scenarios, poses new research challenges. Two new major issues arise in the design of audio-visual ASR systems [4], namely: 1) *Selection and robust extraction of visual speech features*. 2) *Optimal fusion of the audio and visual features*. In this paper, we will briefly describe our ongoing research in the area and how we have tried to address the aforementioned challenges in our audio-visual ASR system.

4. References

- [1] H. D.I. Abarbanel, *Analysis of Observed Chaotic Data*, Springer-Verlag, New York, 1996.
- [2] D. Dimitriadis, P. Maragos and A. Potamianos, “Robust AM-FM Features for Speech Recognition”, to appear, *IEEE SPL*, 2005.
- [3] H. Hermansky and N. Morgan and A. Bayya and P. Kohn “Compensation for the Effect of the Communication Channel in Auditory-like Analysis of Speech (RASTA-PLP)” *Eurospeech-91*, pp. 578–589, 1991.
- [4] G. Potamianos, C. Neti, J. Luetin, and I. Matthews. Audio-visual automatic speech recognition: An overview. In G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, editors, *Issues in Visual and Audio-Visual Speech Processing*, chapter 10. MIT Press, 2004.
- [5] S. Seneff, “Pitch and spectral estimation of speech based on an auditory synchrony model”, *ICASSP-84*, pp.3621-3624, 1984.
- [6] D.G. Stork and M.E. Hennecke, editors. *Speechreading by Humans and Machines*. Springer, Berlin, Germany, 1996.

Wave-shaping models of cyclic signals

Jean Schoentgen

Department “Signals and Waves”, Faculty of Applied Sciences, Université Libre de Bruxelles, Brussels, Belgium, jschoent@ulb.ac.be

The presentation is devoted to the modelling of the glottal waveform, which is the acoustic signal that is generated by the vibrating vocal folds and pulsatile airflow. The formalism that is discussed is, however, generic and applicable to any band-limited cyclic signal the Fourier series of which exists.

A wave-shaping model is a non-linear memory-less representation of a cycle of a signal. Usually, the shaper is a pair of polynomials that transform a harmonic driving function into any desired cycle shape. The representation is exact when the desired waveform is band-limited and its Fourier coefficients are known.

Most often, the glottal airflow rate and its derivative are modelled by means of concatenated curves that mimic observed glottal cycle shapes. Although these models are very popular, their use may be problematic, especially when the model's parameters are manipulated so as to change cycle amplitudes, lengths or shapes rapidly in time. The reason is that the spectral bandwidths of arbitrary curve-based signal models are unknown a priori and the bandwidth may evolve in a non-obvious manner with the model parameters. This may be a problem in the framework of numerical syntheses of signals, when the bandwidth must be controlled to avoid spectral aliasing.

A possible solution consists in using curve-based models as templates of cycle shapes and approximating these by means of their Fourier series. The Fourier approximation enables controlling the bandwidth of the represented signals as long as these are periodic or pseudo-periodic. If, however, the fundamental frequency or amplitudes of the harmonics are strongly modulated to mimic signals the cycle lengths or shapes of which change rapidly, the modulation may broaden the bandwidth considerably in uncontrollable ways.

A formalism that enables an easier control of the bandwidth than the Fourier series is the wave-shaping representation. Its practical advantage stems from the predictable link between the bandwidth of the quasi-sinusoidal driving function and the total bandwidth. Checking the overall bandwidth of the output signal therefore amounts to restricting the bandwidth of a modulated sine, which is easier than controlling the full bandwidth of a time-evolving pulsatile signal.

In addition, wave shaping enables independently controlling the glottal cycle length, the spectral brilliance of the signal as well as the auditorily perceived speaker identity. This may be an advantage because letting evolve the parameters of curve-based models may re-touch all cycle properties at once.

The generic, mathematically based part of the presentation comprises the following items. First, a proof of the equivalence of the Fourier and wave shaper representations of harmonic signals; second, a formulation of the mathematical link between the bandwidth of the output signal and the bandwidth of the quasi-sinusoidal driving function; third, the generation of the derivative or integral of the represented signals by means of the shaper formalism of the original signal.

The relevance of the latter is that the glottal excitation signal is frequently modelled as the derivative with regard to time of the glottal airflow rate. The extension of the wave shaper formulation so as to automatically generate derivatives or integrals therefore enables simulating the glottal excitation and glottal airflow rate by means of the same core model.

The wave shaping-based model of the speech signal is illustrated in the framework of the simulation of disordered voices, which involves the rapid cycle-to-cycle evolution of properties such as the cycle length, amplitude and shape.

From black boxes to gray boxes: A look behind the curtain of nonlinear speech models.

Gernot Kubin,
Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria

Abstract - When we started our work on nonlinear dynamical modeling of speech signals some 15 years ago we did this based on a strong mathematical existence theorem due to F. Takens that would guarantee us to build signal models able to re-generate perfectly equivalent observables without even attempting to identify the structure of the underlying physical signal generation mechanism. Combining this existence theorem with automatic machine learning methodologies has naturally led to a framework where observed speech signals (from sustained, stationary sounds) can more or less automatically be converted to a nonlinear black-box oscillator with a parameterization adapted to the data, but with no insight into the relationship between these parameters and their physical meaning.

Many subsequent years have been devoted to incremental improvements of the structures used for representation of the black-box oscillators and to the tuning of the associated machine learning algorithms. However, it was only during the course of the past five years, where our work was embedded in the COST277 environment, that the constant challenge from our colleagues in speech synthesis made us relate our model back to the human speech signal generation apparatus. By now, we have converged to a gray-box model where some light is shed on the relationship of the model structure (in terms of a decomposition into a small number of interacting modules) to the underlying physics. The model is still no glass-box as we have maintained an emphasis on the ability to automatically identify all the parameters from data without other physical measurements than the speech waveform itself. This model has been referred to as the oscillator-plus-noise model and condenses what we have learnt in the COST action. At the eve of its completion, we are about to lift the curtain even further by working towards glass-box models that will use limited prior assumptions on model complexity to uncover more details of the internal system structure of the physical modules).

Analysis of clusters of correlated activity in fMRI data by Hopfield Neural Networks

Abstract

A variant of Hopfield's neural network model is presented as an approximative method for the identification of regions of interest from fMRI data. If a subset of pixels is mutually correlated it forms a cluster which can be interpreted as a functional unit. Correlation clusters are extracted from the data without any reference to the stimulus and thus various aspects of stimulus related activity can be distinguished and dependencies between stimulus related and background activity can be revealed. In this approach networks of pixels with correlated time courses represent functionally engaged regions of brain activity. The stationary configurations of the network dynamics represent such a kind of correlated clusters and can be assumed to indicate functional connectivity. The degree of connectivity within a retrieved cluster is allowed to range continuously between the extreme cases of cliques and connectivity components, allowing thus for clusters of intermediate connectivity. The graph connectivity can be fixed posteriorly by an intrinsic significancy measure or by correlation to the stimulus. These clusters are subsequently analyzed to identify spacial structures in the brain and more precisely task related regions. The computational complexity and disambiguation quality of the algorithm is compared to results from ICA and graph-theoretical algorithms.

Nonlinear predictive models: Overview and possibilities in speaker recognition

Marcos Faundez-Zanuy, Mohamed Chetouani

Escola Universitària Politècnica de Mataró (BARCELONA), SPAIN
Laboratoire des Instruments et Systèmes d'Ile-De-France, Université Paris VI
faundez@eupmt.es, mohamed.chetouani@lis.jussieu.fr
<http://www.eupmt.es/veu>

Abstract. In this paper we give a brief overview of speaker recognition with special emphasis on nonlinear predictive models, based on neural nets.

Voice disguise and automatic detection

Patrick Perrot ^{(*)(**)}, G. Aversano ^(*), G. Chollet ^(*)
perrot,aversano,chollet@tsi.enst.fr

^(*) CNRS-LTCL, GET-ENST, TSI Department

^(**) Institut de recherche criminelle de la gendarmerie nationale

Abstract

This study focuses on the question of voice disguise and the problem of its detection. The voice disguise is considered as a deliberated action of the speaker who wants to falsify or to conceal his identity. Lots of possibilities are offered to a speaker to change his voice and to false a human ear or an automatic system. He could transform his voice by electronic scrambling or more simply by exploiting the intra-speaker variability: modification of his own pitch, modification of the position of the articulators like lips or tongue which affect the formant frequencies. The proposed work is divided in three parts: the first one is a classification of the different possibilities available to change his voice, the second one presents a review of the different techniques used in the literature and the third one described the main clues proposed in the literature to distinguish a disguised voice from an original voice, before to propose some directions of research based on disordered and emotional speech.

Different means exist to change his voice : a classification of those techniques is proposed: electronic and non-electronic. The aim of this classification is to study separately each kind of disguise class and to determine some specific characteristics. A distinction is realized between electronic and non electronic conversion. In the case of electronic changes, the most sophisticated method consists in mimicking a specific voice. That is what is qualified as voice conversion. Different techniques of voice conversion are explored in the literature with more or less success. The main works on this field and their results are presented. The principle is to elaborate a conversion function between a source and a target voice and to apply it to the source voice. The aim is to obtain a transformed source voice that sounds like a target voice. The second class of change in the electronic field is what is qualified as voice transformation. It consists in modifying his voice by some specific methods. These methods could be separated in two categories: parametric methods based on an accurate signal model and non-parametric method based on temporal or frequency field.

The second main part of the classification is the non-electronic changes. In the field of voice conversion the different studies on the work of a professional impersonator are presented to understand the main features used to imitate a voice. At last in the register of non-electronic voice transformation, two categories is described: alteration of the voice by using a mechanic mean like a pen in the mouse for instance, and prosody alteration. This last category is very large because the impostors can modify lots of parameters of his voice. The position of the different articulators affect vowel sounds for instance, the use of a foreign accent changes some voice features, but also the modification of the pitch or the formants position and so on. Before to propose some specific parameters to study, a presentation of different works on the detection of voice disguise is presented. Most of the studies available in the literature concerns the most common voice disguise that is to say the prosody alteration. In order to organize our work some particular disguises has been chosen. Our aim is to be able to recognize automatically a disguise, to identify it and if it is possible to link the disguised voice to the original speaker. A description of the method that we plan in order to satisfy those different objectives is presented. The method is based for one part on the characterization of some specific features, like the pitch, the position of the formants... and for a second part on a clustering technique dedicated to elaborate specific models for each kind of disguise.

Audio-visual forgery in identity verification

G. Chollet, B. Abboud, G. Aversano

Verification of identity is commonly achieved by looking at the face of a person and listening to his (her) speech.

Automatic means of achieving this verification has been studied for several decades. A speaking face offers many features to achieve a robust verification of identity.

The current deployment of videophony offers new opportunities for a secured access to remote servers (banking, certification, ...)

The synchrony of the speech signal and lip movements is a necessary condition to check that the observed speaking face has not been manipulated and/or synthesized.

This paper will review face, speaker and speaking face verification, face morphing and voice transformation techniques. It is demonstrated that a dedicated impostor needs limited information from a client to fool state of the art audio-visual identity verification systems.

A visual forgery system has been designed allowing to artificially animate a static image in such a way that it reproduces the facial movements (and in particular the lip movements) of a talking face shown in a video. This allows elaborating a forgery scenario where an impostor utters a number of words and uses his own lip movements to artificially animate the lips of a static image of the client in such a way that it reproduces the exact lip movements of the impostor.

Moreover, to modify the voice characteristics of the impostor in such a way that it would resemble to the voice of the client, two different voice transformation techniques have been developed and tested. Forgery attacks against a speaker verification system have been simulated on widely known databases and protocols such as BANCA and NIST.

The choice of the transformation technique is made according to the available quantity of client voice data.

If only a limited amount of client data is available for training (as in the case of BANCA protocol), a spectral conversion technique should be adopted. Considering a sequence of spectral vectors pronounced by the impostor, $X = [x_1, x_2, \dots, x_n]$, and a sequence composed by the same words, pronounced by the client, $Y = [y_1, y_2, \dots, y_n]$, a spectral transformation can be performed by finding the conversion function F that minimizes the mean square error:

$$\varepsilon_{\text{mse}} = E[\|y - F(x)\|^2],$$

where E is the expectation.

This conversion method requires a preliminary word-level segmentation of the training sentences.

If more client data are available (e.g. approximately 1 hour of client's speech), we could use a voice encoder based on the "Automatic Language Independent Speech Processing" (ALISP) approach. The principle of this system is to encode speech by recognition and synthesis in terms of basic acoustic units that can be derived by an automatic analysis of the signal. Such analysis is not based on a priori linguistic knowledge.

In the context of speaking face cloning a system has been implemented that allows automatic lip movement tracking for a speaking face. The same algorithm is also used to automatically place a certain number of MPEG-4 compatible fiducial points around the lip contour of the static image we wish to animate. An image warping technique is then used to move these fiducial points on the static image in such a way that their new position matches the position of the tracked feature points on each frame of the driving video sequence.

Current limitations of the described transformation techniques are the difficulties to achieve real-time performance.

Automatic Speaker Verification: state of the art and current issues

Dijana Petrovska-Delacrétaz, Asmaa El-Hannani and Gérard
Chollet

Speech is often the only available modality to recognize the identity of a person (over the telephone, the radio, in the dark, ...). Automatic speaker recognition has been studied for several decades. Few applications are deployed. This paper analyses the individual characteristics of human voice, its variability, possibilities of disguise and mimicry, and automatic techniques to capture robust characteristics and use them to identify speakers.

In this paper the state of the current speaker recognition research is reviewed. Basic principles of speaker recognition are first summarized. The choice of the speech features and speaker models are mostly related to the individual characteristics (variability) of the speakers' voices. Besides the speaker's variability, we are faced with other factors, such as channel variability, silence detection, or score normalization, that influence the performance of speaker verification algorithms. All these issues are illustrated on recent NIST speaker evaluation databases. The utility of open source reference (baseline) algorithms and controlled evaluation campaigns is also pinpointed.

The field of speaker recognition is also reviewed in relation to speech recognition, focussing on the usage of this new source of information for the speaker recognition task. This relationship has to be seen as an important issue in the development of new services based on speaker and speech recognition. Overview of recent developments in this field is given. With the combination of speech and speaker recognition, we introduce also the issue and challenges of multimodal biometrics (i.e. talking faces).

A New Approach for Speech Feature Extraction Based on Genetic Algorithms

C. Charbuillet, B. Gas, M. Chetouani, J. L. Zarader

**Laboratoire des Instruments et Systèmes d'Ile-de-France
Université Pierre et Marie Curie, Paris, FRANCE**

christophe.charbuillet@lis.jussieu.fr gas@ccr.jussieu.fr
mohamed.chetouani@lis.jussieu.fr zarader@ccr.jussieu.fr

Feature extraction plays a major role in a pattern recognition system. Its aim is to extract useful information in order to increase the classifiers' performances for a given task. Current speech feature extraction methods rely mainly on properties of speech production (LPC, LPCC) and perception (MFCC, PLP). State of the art methods in speech feature extraction don't take into consideration the specific information about the task to accomplish. However this field has recently been investigated using LVQ and MLP based nonlinear predictive methods [1] and giving interesting results.

We propose in this article to use genetic algorithms (GAs) to design a new feature extraction method adapted to a speaker recognition system.

Genetic algorithms were first proposed by Holland in 1975 [2] and have become widely used in various disciplines as a new mean of complex systems optimization. The basic idea of GAs is based on "natural selection", the principle of "survival of the fittest". A GA operates on a population of chromosomes, each generating a potential solution to the studied problem.

GAs most attractive quality is certainly their aptitude to avoid local minima. However, our study relies on another quality which is the capacity to optimize a complex system without needing any knowledge about it. This allows us to realize a feedback between the recognition system's output and the feature extractor to optimize.

In this article we present an application of our algorithm to the adaptation of the common MFCC extractor to the speaker clustering task. This adaptation consists in designing a filter bank, with optimized center frequencies and band-widths. In order to evaluate the performances of the developed algorithm, we test it on a sub-database from the ESTER [3] corpus of broadcasted radio emissions: two hours for the evolution stage and eight hours for the testing stage. Results showed that the obtained filter bank gives significant improvements compared to the Mel one, reducing the error rate from 7.68% to 5.52%.

More details about the experiments will be given in the final article and results will be discussed.

- [1] *"Non-linear Speech Feature Extraction for Phoneme Classification and Speaker Recognition"* M. Chetouani, M. Faundez, B. Gas and J.L. Zarader. in "Nonlinear speech processing : Algorithms and Analysis". Eds. G. Chollet, A. Esposito, M. Faundez, M. Marinaro. Springer Verlag (2005).
- [2] *"Adaptation in Natural and Artificial Systems"* J. H. Holland, University of Michigan Press, 1975, Ann Arbor.
- [3] *"The Ester Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News"* S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, G. Gravier , Proc.Eurospeech/Interspeech, 2005.