Wrapped Gaussian Mixture Models for Modeling and High-Rate Quantization of Phase Data of Speech

Yannis Agiomyrgiannakis and Yannis Stylianou, *Member, IEEE*

***Abstract*—The harmonic representation of speech signals has found many applications in speech processing. This paper presents a novel statistical approach to model the behavior of harmonic phases. Phase information is decomposed into three parts: a minimum phase part, a translation term, and a residual term referred to as *dispersion phase*. Dispersion phases are modeled by wrapped Gaussian mixture models (WGMMs) using an expectation-maximization algorithm suitable for circular vector data. A multivariate WGMM-based phase quantizer is then proposed and constructed using novel scalar quantizers for circular random variables. The proposed phase modeling and quantization scheme is evaluated in the context of a narrowband harmonic representation of speech. Results indicate that it is possible to construct a variable-rate harmonic codec that is equivalent to iLBC at approximately 13 kbps.**

***Index Terms*—Circular statistics, , phase quantization, sinusoidal models, speech analysis, speech coding, voice-over-IP, wrapped Gaussian mixture models (WGMMs).**

I. Introduction

TYPICALLY, the spectrum of a speech sound is typically split into two parts: an *amplitude spectrum* and a *phase spectrum*. The statistical behavior and properties of amplitude spectra and related parameterizations are well known and widely used in many speech processing applications. Amplitude spectra are well modeled by *spectral envelopes* which are continuous parametric curves describing the coarse structure of the spectrum. Unfortunately, the concept of the "spectral envelope" has not found its counterpart in phase spectra due to the intrinsic difficulties associated with the accurate and robust modeling of phases in voiced speech. Currently, there is no widely accepted parameterization of phase spectra. However, there are several studies that indicate the importance of phase in speech perception [1]–[3]. In this paper, we address the problem of modeling and quantization of phases for speech coding in the context of harmonic representation of speech.

In CELP speech codecs, the excitation of the AR filter is typically encoded with a closed-loop codebook search. Therefore, phases and fine-spectral details which are not captured by the AR spectral envelope are blindly encoded together [4]. On the contrary, sinusoidal coders represent speech using a series of harmonically related sinusoids and rely on a phase model to facilitate efficient coding. For example, in codecs based on sinusoidal transform coding (STC) [5] and multiband excitation (MBE) [6], the harmonics are classified as voiced or unvoiced; the voiced harmonics are constructed using the assumption that the excitation is a sequence of zero phase pulses while the unvoiced harmonics are constructed using random phases [4], [7], [8]. Zero phase assumption is, however, not accurate because according to the source-filter model of speech production, the excitation corresponds to the derivate of a maximum phase signal, the glottal air flow [8, p. 151]. This poses an upper bound to the quality of encoded speech at higher bit-rates but, in practice, it works well at low bit-rate coders (below 4 kbps). As a consequence, many researchers argue that high-quality sinusoidal speech coding requires the encoding of phases [9]–[14].

A *model-based* approach is to fit a deterministic model to the excitation signal or directly to the harmonic phases of the speech signal. In [15], the excitation is constructed using the Rosenberg glottal pulse model [16]. Another idea is to use all-pass filters to correct the phase response of the minimum phase AR spectral envelope [9], [10]. A drawback of the latter methods is that the resulting all-pass filters may be unstable. The parameters of the all-pass filter can also be computed on the frequency domain [17] by minimizing a squared-error criterion that is used directly on the phase values but this distortion measure is prone to errors due to the modulo-2π behavior of the phases.

Harmonic phases may directly be quantized without the requirement of a deterministic model. In [11] and [12], the *phase residual*, the difference between the phase of the current frame, and its prediction from the previous frame is quantized. Vector quantization of phases was proposed in [13] for the quantization of the harmonic phases of the slowly evolving waveform (SEW) in the context of waveform interpolation (WI) coders. An important contribution of the latter work is the introduction of a distortion measure that takes into account the modulo-2π behavior of phases and the derivation of the corresponding k-means algorithm for codebook training. However, codebook-based phase quantizers cannot operate at increased bit-rates. A GMM-based phase quantization algorithm capable of operating at high rates was provided in [14], where a K-dimensional GMM was used to model the source

Manuscript received June 26, 2008; revised September 12, 2008. Current version published April 01, 2009. This work was supported by the General Secretary of Research Technology Hellas and ICS FORTH under an ARISTEIA grant. A short version of the suggested model and quantization of phase data has been presented at ICASSP'07, Honolulu, HI. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hiroshi Sawada.

The authors are with the Institute of Computer Science, Foundation of Research and Technology Hellas (FORTH), GR-700 13 Heraklion, Crete, Greece, and Multimedia Informatics Lab, Department of Computer Science, University of Crete, 71409 Heraklion, Crete, Greece (e-mail: jagiom@csd.uoc.gr; yannis@csd.uoc.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2008229

pdf while scalar quantization of each dimension was made with a bounded-support compander/quantizer that was restricted to operate in $(-\pi, \pi]$. Although the design of the latter quantizer is hindered by ad-hoc choices like the application of a conventional GMM to bounded-support data, the suboptimal bit-allocation procedure and the necessity of having companding/quantization operating at a low rate/dimension where it is suboptimal, the main objection against this approach arises from the fact that it does not take into account the modulo-2π behavior of phase and may require as many as 2^K Gaussians to capture the statistics of a single class centered near the border of the $(0, 2\pi]^K$ hypercube. In other words, it is too dependent on the structure of the source probability density function (pdf); an undesirable property for a quantizer.

The quantization of phases has also been used in concantenative sinusoidal-based text-to-speech (TTS) systems. These systems use a database that contains segments of speech (usually referred to as "units"). The size of the database can be rather large and efficient storage may require the compression of the phase data. In [18], phase is encoded with a very high bit rate of 7 bits/harmonic. Furthermore, small footprint implementations suitable for low-end terminals like handheld devices may require further compression. A typical approach is to quantize only the phases of the first harmonics [19], [20] which define the coarse shape of the waveform. Another approach is to fit the unwrapped phases to a sinusoidal basis [21].

A direct comparison of phase quantization algorithms is not always possible because of the lack of a widely accepted phase distortion measure and to the strong coupling between the phase quantizer and the analysis/synthesis procedure of the sinusoidal coder. An important limitation is that the typically used squared-error distortion measure between the original and the reconstructed waveform does not correlate well with the perceived distortion at low/medium rates [22]. Another option is to compute the distortion directly on the phase data. A psychoacoustic study of a simple difference phase distortion measure is made in [23] to facilitate perceptual weighting of the harmonic phases.

Summarizing, up to the knowledge of the authors, the problem of phase quantization for speech coding has not yet been addressed in a way that is suitable for high-rates and takes into account the peculiarities associated with the circular nature of phase parameters. This paper proposes a way to reveal that speech phases actually have nonuniform distribution—thus a justification for phase vector quantization—a statistical approach to model these phases and a suitable high-rate quantization method for circular spaces. From the source coding aspect, this work recasts state-of-the-art GMM-based quantization [24] to circular spaces, while retaining the advantages of rate scalability and computational efficiency.

Initially, the harmonic phases are decomposed into three parts: a *minimum phase* part that accounts for the phase response of a minimum phase system, a linear phase part using a *translation term* that aligns the harmonic signal according to a reference point within the glottal cycle, and a residual part referred to as *dispersion phase*. The dispersion phases exhibit a covariation that is clearly observed on bivariate scatter plots. These intraphase dependencies of the dispersion phases can be used for their efficient quantization. Assuming that the minimum phase component and the linear phase term can be easily modeled, our focus is on the modeling and the quantization of the dispersion phases. The construction of a quantizer that is suitable for phases is developed in two steps: first, the statistics are captured via a suitable model, and then, a quantizer is constructed according to these statistics.

Phase data exhibit a modulo-2π behavior that bounds them on the surface of the n-Torus manifold $\mathbb{T}^K = \mathbb{R}^K/2\pi\mathbb{Z}^K$, which is the extension of the unit circle to K dimensions. The corresponding statistics are called *circular* or *directional statistics* [25]. We suggest to model the dispersion phases using the so-called *wrapped Gaussian mixture model* (WGMM) which is able to model variables that exhibit a modulo-2π behavior. Only a few recent publications utilize wrapped mixture models to model circular data: In [26], *wrapped Hidden Markov Models* (HMM) are used to track the trajectories of sound sources inside a room. In [27], wrapped (*Normal*, *Cauchy*) mixture models are used to study time series with linear and circular variables. An expectation-maximization (EM) algorithm for wrapped multivariate Gaussians and an extension to HMM is presented in [26] for the case of Gaussian components with diagonal covariance matrices. The EM algorithm provided in [26] estimates the parameters by performing the EM steps one dimension at a time, a restriction that is not necessary as it will be shown. This paper presents a vector-based EM algorithm for a WGMM with full covariance matrices and then, focuses on the more tractable case where the Gaussian components have diagonal covariance matrices.

The WGMM is then used to construct a quantizer for the *dispersion* phase data. Initially, we define a distortion measure that is suitable for circular spaces. Then, we extend the idea of GMM-based quantization [24] to WGMM. The extension requires the construction of a scalar quantizer for wrapped Gaussian random variables. Such a task is not trivial because the shape of the wrapped Gaussian pdf is changing with the variance. Two solutions are suggested: the first, simply *wraps* the codepoints of a linear Gaussian to the circumference of the unit circle, while the second introduces the concept of "*Polynomial Code-Functions*" (PCF) to facilitate efficient quantization. In PCF, each codepoint is replaced by a polynomial function of variance. The construction of a wrapped Gaussian codebook is made by sampling a set of PCF at the corresponding variance. Finally, the WGMM-based quantizer is used in a phase quantization scheme for a prototype narrowband sinusoidal codec and evaluated in terms of PESQ-MOS [28]. It is shown that an MOS score of 3.88 can be achieved with an average rate of $\sim$13 kbps for all speech parameters, a result that is similar to the performance of the iLBC[1] codec [29] (at 15.2 kbps).

The outline of the paper is as follows. Section II presents the harmonic sinusoidal model and the phase decomposition procedure. Section III provides a brief introduction to circular statistics and presents WGMM. An EM algorithm for WGMM is then derived and discussed in Section IV. Section V focuses on the construction of a scalar quantizer for wrapped Gaussian random variables. The scalar quantizer is then used to construct

[1] Internet Low Bit-Rate Codec: an open source standard codec created by GlobalIPSolutions for voice-over-IP.

a WGMM-based quantizer that is described in Section VI. Section VII presents a WGMM-based phase quantization scheme suitable for narrowband speech. The evaluation of the proposed scheme is provided in Section VIII. Finally, Section IX concludes the paper.

II. Harmonic Phase Decomposition

The harmonic representation of the speech signal is a high quality parametric model used for analysis/synthesis and modification of speech [8]. The speech signal is typically analyzed in short intervals referred to as *frames*, where it is assumed to be stationary. Within each frame, the signal is represented as a weighted sum of harmonically related sinusoids

$$x(n) = \sum_{k=1}^{K} A_k \cos(k\omega_0 n + \phi_k) \tag{1}$$

where ω_0 is the fundamental frequency (in radians), K is the number of the harmonics, A_k and ϕ_k are the amplitude and the phase of the kth harmonic, respectively, $x(n)$ the analyzed speech signal, and n is the time index. The amplitudes and the phases can be obtained using least squares methods [30].

Harmonic amplitudes A_k maybe well represented using a minimum phase real cepstrum coefficients (RCC) spectral envelope with 20 dimensions. The cepstral envelope fits the log-spectra, $\log(A_k)$, at the Mel-scale by solving a regularized least squares problem [30].

Let $H_s(\omega)$ be the frequency response of the RCC envelope. Phases ϕ_k may be decomposed into a minimum phase term $\angle H_s(k\omega_0)$, a *linear phase term* $k\omega_0\tau$ and a *dispersion term* ψ_k

$$\phi_k = k\omega_0\tau + \angle H_s(k\omega_0) + \psi_k. \tag{2}$$

Therefore, phase term $\psi_k + k\omega_0\tau$ corresponds to the phase of an excitation-like signal at the corresponding frequency $k\omega_0$ since the subtraction of the minimum phase term corresponds to inverse filtering with the minimum phase system $H_s(\omega)$. Assuming that all amplitude spectrum information is captured by the magnitude response of $H_s(\omega)$, the excitation signal $e(n)$ can be reconstructed according to the following formula:

$$e(n) = \sum_{k=1}^{K} \cos(k\omega_o n + k\omega_0\tau + \psi_k). \tag{3}$$

The linear phase term $k\omega_0\tau$ corresponds to the translation of the center of the analysis window by τ samples with respect to a reference point inside the pitch period. As a reference point, the peak of the excitation $e(n)$ within a single pitch period was used. The peak-picking is made to a uniformly sampled version of the excitation $e(n)$, using 128 samples (7 bits). We found that this procedure provides robust reference points within the glottal cycle and observed that the dispersion phases $\psi_k, k = 1, \ldots, K$ have a distribution that exhibits structure (covariation). In this paper, we use a WGMM to model the statistics of the underlying pdf of dispersion phases. A two-dimensional example of phase intra-correlation is shown in Fig. 1. Note that this structure is evident only in voiced frames while in unvoiced frames the phases are noisy having a uniform distribution. Moreover, the phases were extracted from the harmonic analysis of 20-ms voiced frames with a signal-to-noise-ratio (SNR) above 8 dB. The analysis was made in the training set of TIMIT speech database [31], which consists of a large number of speakers and a phonetically balanced corpus under clean recording conditions. It is worth to mention that phase structure has also been reported in [14], where it was observed that excitation phases, however, extracted with a different process tend to concentrate around a point of reference (i.e., zero).

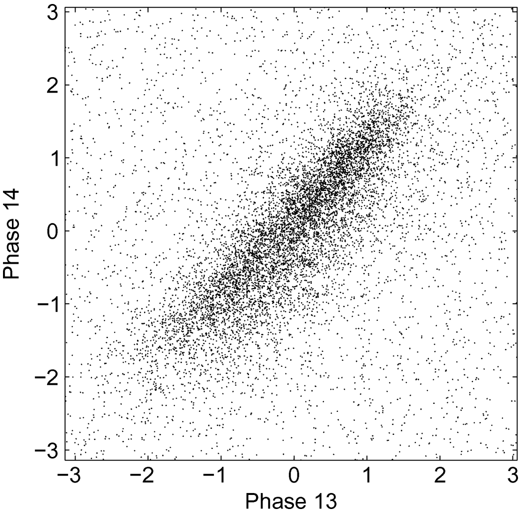

Fig. 1. Scatter plot of the dispersion phases of the 13th and 14th harmonic derived from voiced speech frames with pitch between 95 and 115 Hz. The mean of the dataset is removed.

III. Circular Statistics

Let θ be a circular random variable defined on the circumference of the unit circle $\mathbb{T}^1$. The corresponding circular pdf $f(\cdot)$ is periodic with period $2\pi : f(\theta) = f(\theta + w2\pi), \forall w \in \mathbb{Z}$. The function $f(\cdot)$ integrates to unit in $(0, 2\pi]$. For notational simplicity, we will henceforth assume that all circular variables are constrained to their *principal values* in $(0, 2\pi]$ which is obtained by taking the following modulo operation:

$$\theta \leftarrow \theta \; mod \, 2\pi. \tag{4}$$

Let $\theta_n, n = 1, \ldots, N$ be N samples drawn from $f(.)$. The *circular mean* μ_c and the *circular variance* σ_c^2 of θ are defined as [25] (pg. 20):

$$\begin{aligned} \text{Circular Mean} : \mu_c &= \arg\left(\mathrm{E}\{\mathrm{e}^{\mathrm{j}\theta_\mathrm{n}}\}\right) \\ \text{Circular Variance} : \sigma_c^2 &= 1 - \left\|\mathrm{E}\{\mathrm{e}^{\mathrm{j}\theta_\mathrm{n}}\}\right\| \end{aligned} \tag{6}$$

where $\mathrm{E}\{.\}$ denotes the expectation operator and j is the imaginary unit ($j^2 = -1$). The circular mean $\mu_c \in (0, 2\pi]$ measures the mean direction of the data and $\sigma_c^2 \in [0, 1]$.

The statistics of θ can be captured either by distributions which are directly defined on the unit circle, like the von Mises distribution, or by wrapping a pdf of a linear random variable to the circumference of the unit circle [25]. The corresponding distributions are called *wrapped*. In practice, the von Mises and the wrapped Gaussian distribution are very similar, therefore, we chose to work with the wrapped Gaussian distribution because it is more convenient for the purpose of quantization since it is

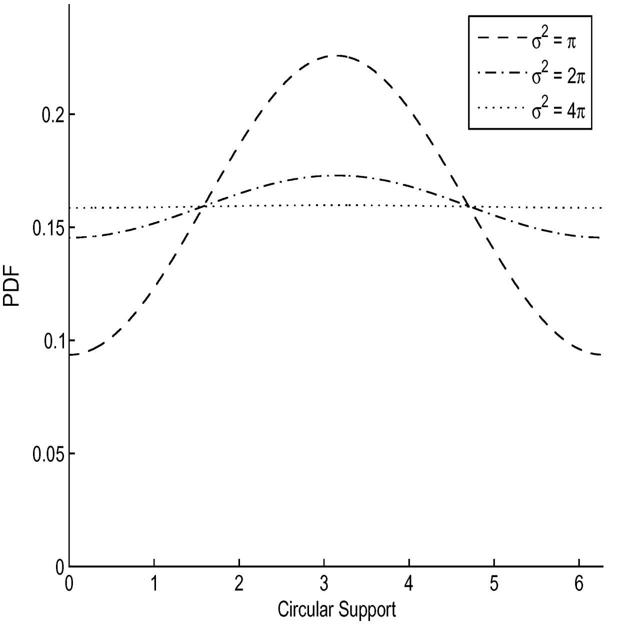

Fig. 2. Examples of scalar wrapped Gaussian pdf with variance $\sigma^2 = \{\pi, 2\pi, 4\pi\}$.

directly linked to the linear Gaussian which is well studied. For example, if $N(\theta;\mu,\sigma^2)$ is the pdf of the linear Gaussian distribution, the *wrapped Gaussian* is given by equation

$$N_w(\theta;\mu,\sigma^2) = \sum_{w=-\infty}^{\infty} N(\theta - w2\pi;\mu,\sigma^2) \tag{7}$$

where μ and σ^2 is the mean and the variance of the wrapped Gaussian. Note that the mean μ and the variance σ^2 of the wrapped Gaussian is related to the circular mean μ_c and the circular variance σ_c^2 by

$$\mu = \mu_c \; mod \; 2\pi \tag{8}$$
$$\sigma^2 = -2\log\left(1-\sigma_c^2\right). \tag{9}$$

The wrapped-Gaussian pdf can be approximated by the linear Gaussian pdf at small variances $\sigma^2 \leq 1$ and by the uniform distribution at large variances $\sigma^2 \geq 2\pi$. This is illustrated in Fig. 2, where three examples of a wrapped Gaussian pdf are shown. The wrapped Gaussian pdf is constructed by *infinite wrappings* of the linear Gaussian pdf inside the interval $(0, 2\pi]$. In practice, though, a summation over ±2 tilings provides a sufficient approximation even for large variances $\sigma^2 \leq 2\pi$, because in that case, the pdf of the truncated Gaussian closely approximates the pdf of the unconstrained Gaussian.

The extension to multivariate data is straightforward. Let $\vec{\psi} \in \mathbb{T}^K$ be a circular random vector. The *linear multivariate Gaussian PDF* is given by

$$N(\vec{\psi};\vec{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^K|\Sigma|}} e^{-\frac{1}{2}(\vec{\psi}-\vec{\mu})^T\Sigma^{-1}(\vec{\psi}-\vec{\mu})} \tag{10}$$

where $\vec{\mu}$ is the mean and Σ the covariance matrix. The K-dimensional *multivariate wrapped Gaussian PDF* is then obtained by *tiling* the linear Gaussian kernel over all possible $(0,2\pi]^K$ hypercubes, as shown as

$$N_w(\vec{\psi};\vec{\mu},\Sigma) = \sum_{\vec{w}\in\mathbb{Z}^K} N(\vec{\psi} - \vec{w}2\pi;\vec{\mu},\Sigma) \tag{11}$$

where $\vec{w}$ is a vector that holds integer displacements. The corresponding pdf is a periodic function that is defined (and integrates to unit) in the $(0,2\pi]^K$ hypercube. Multivariate wrapped Gaussian pdfs have also been proposed in [32] for the purpose of handwriting recognition.

The Wrapped Gaussian Mixture Model can now be defined as

$$p(\vec{\psi};\Omega) = \sum_{m=1}^{M} \alpha_m N_w(\vec{\psi};\vec{\mu}_m,\Sigma_m) \tag{12}$$

where α_m, $\vec{\mu}_m$, and Σ_m are the posterior probability, the mean, and the covariance matrix of the mth wrapped Gaussian component, respectively, and Ω is the set that holds all these parameters. Note that (11) requires a summation over an infinite number of tilings. In the scalar case, it suffices to perform the summation over ±2 tilings. The number of hypercubes which are two hops away from the principal hypercube (at $\vec{w} = \vec{0}$) is increasing exponentially with dimensionality. In practice, it is not possible to consider more than two dimensions. This problem can be resolved if the covariance matrices are constrained to be diagonal: $\Sigma_m = \mathrm{diag}(\sigma_m^2(1), \sigma_m^2(2), \ldots, \sigma_m^2(K))$. In this case, the wrapped pdf can be computed as

$$p(\vec{\psi};\Omega) = \sum_{m=1}^{M} \alpha_m \sum_{\vec{w}\in\mathbb{Z}^K} \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\sigma_m^2(k)}} \times e^{-\frac{(\vec{\psi}(k)-\vec{\mu}_m(k)-\vec{w}(k)2\pi)^2}{2\sigma_m^2(k)}}$$
$$\Rightarrow p(\vec{\psi};\Omega) = \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K} \sum_{w\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma_m^2(k)}} \times e^{\frac{-(\vec{\psi}(k)-\vec{\mu}_m(k)-w2\pi)^2}{2\sigma_m^2(k)}}$$

where $\vec{\psi}(k)$, $\vec{\mu}_m(k)$ are the kth element of $\vec{\psi}$ and $\vec{\mu}_m$, respectively. An example that proves the interchange for the three-dimensional case is provided in the Appendix. The interchange between the product over the dimensions and the summation over the wrappings allows a significant complexity reduction.

IV. An Expectation-Maximization Algorithm for Wrapped Gaussian Mixture Models

The estimation of WGMM parameters from N phase vectors $\vec{\psi}_n$ can be made using the *maximum-likelihood* criterion. The maximization of the log-likelihood function

$$L(\Omega) = \sum_{\vec{\psi}} \log p(\vec{\psi}|\Omega)$$

over all Ω is a difficult optimization task. However, it can be easily addressed with a two-step procedure that belongs to the class of *expectation-maximization* algorithms [33]. According to the WGMM, a sample $\vec{\psi}$ is generated by a single tiling of a single wrapped Gaussian component. Using a slight abuse of notation, let m be the random variable of the wrapped Gaussian component and $\vec{w}$ the random variable of the tiling. Both variables are *latent variables* (hidden from us). Under these considerations, we can derive an EM algorithm for WGMM:

Initialization: Set Ω_0.

Step 1: Expectation

$$p(m,\vec{w}|\vec{\psi}_n) \leftarrow \frac{\alpha_m N(\vec{\psi}_n - \vec{w}2\pi; \vec{\mu}_m, \Sigma_m)}{p(\vec{\psi}_n; \Omega_0)}.$$

Step 2: Maximization

$$\alpha_m \leftarrow \frac{1}{N}\sum_{n=1}^{N}\sum_{\vec{w}\in\mathbb{Z}^K} p(m,\vec{w}|\vec{\psi}_n)$$

$$\vec{\mu}_m \leftarrow \frac{\sum_{n=1}^{N}\sum_{\vec{w}\in\mathbb{Z}^K} p(m,\vec{w}|\vec{\psi}_n)(\vec{\psi}_n - \vec{w}2\pi)}{\sum_{n=1}^{N}\sum_{\vec{w}\in\mathbb{Z}^K} p(m,\vec{w}|\vec{\psi}_n)}$$

$$\Sigma_m \leftarrow \frac{\sum_{n=1}^{N}\sum_{\vec{w}\in\mathbb{Z}^K} p(m,\vec{w}|\vec{\psi}_n)(\vec{\psi}_n - \vec{\mu}_m - \vec{w}2\pi)(\vec{\psi}_n - \vec{\mu}_m - \vec{w}2\pi)^T}{\sum_{n=1}^{N}\sum_{\vec{w}\in\mathbb{Z}^K} p(m,\vec{w}|\vec{\psi}_n)}.$$

Check Convergence: Repeat steps 1 and 2 until the convergence of $L(\Omega)$.

where $p(w,\vec{\psi}_n|m) = N(\vec{\psi}_n - \vec{w}2\pi; \vec{\mu}_m, \Sigma_m)$ according to (11).

The initialization step is computed using several trials with random WGMM. A couple of EM iterations were used to improve the initial guess of each random WGMM and the model parameters that provided the best log-likelihood were chosen. Each random WGMM is computed using random Gaussian means, variances that correspond to a fraction of the variance of the dataset and equal posterior probabilities $\alpha_m = 1/M$.

The complexity of the algorithm increases exponentially with dimensionality and becomes intractable for more than two dimensions. However, as it is indicated in the previous section, if we constrain the covariance matrices Σ_m to be diagonal, we can interchange the summation over the tilings with the product over the dimensions and obtain an algorithm with tractable complexity. For computational efficiency, the expectation and the maximization steps are combined to a full EM iteration. Initially, we will define some accessory variables that hold the information related to the expectation step

$$\delta_{k,m,n,w} = N\left(\vec{\psi}_n(k) - w2\pi; \vec{\mu}_m(k), \vec{\sigma}^2_m(k)\right) \quad (13)$$

$$\beta_{k,m,n} = \sum_{w\in\mathbb{Z}} \delta_{k,m,n,w} \quad (14)$$

$$\beta^{(\mu)}_{k,m,n} = \sum_{w\in\mathbb{Z}} \delta_{k,m,n,w}\left(\vec{\theta}_n(k) - w2\pi\right) \quad (15)$$

$$\beta^{(\sigma^2)}_{k,m,n} = \sum_{w\in\mathbb{Z}} \delta_{k,m,n,w}\left(\vec{\theta}_n(k) - \vec{\mu}_m(k) - w2\pi\right)^2 \quad (16)$$

$$\beta_{m,n} = \alpha_m \prod_{k=1}^{K} \beta_{k,m,n} \quad (17)$$

$$\omega_m = \sum_{n=1}^{N} \frac{\beta_{m,n}}{\sum_{m'=1}^{M} \beta_{m',n}}. \quad (18)$$

The update equations can then be written as

$$\alpha_m \leftarrow \frac{1}{N}\omega_m \quad (19)$$

$$\vec{\mu}_m(k) \leftarrow \frac{1}{\omega_m}\sum_{n=1}^{N} \frac{\beta_{m,n}}{\beta_{k,m,n}}\beta^{(\mu)}_{k,m,n} \quad (20)$$

$$\sigma^2_m(k) \leftarrow \frac{1}{\omega_m}\sum_{n=1}^{N} \frac{\beta_{m,n}}{\beta_{k,m,n}}\beta^{(\sigma^2)}_{k,m,n}. \quad (21)$$

In practice, the summations in (14)–(16) are not taken over the whole $\mathbb{Z}$; only ±2 tilings are sufficient. At large variances, the univariate wrapped Gaussian approximates the uniform distribution as illustrated in Fig. 2. Therefore, it is reasonable to constrain the variances in $(0, 2\pi]$ during the training stage in order to ensure that the approximation that is made using only ±2 tilings is valid.

An EM-algorithm for WGMM with diagonal covariance matrices has also been proposed in [26]. The latter algorithm performs the EM process independently for each dimension, an optimization strategy that is suboptimal in general. On the contrary, the proposed algorithm handles all dimensions simultaneously.

V. Scalar Quantization for Wrapped Gaussian Variables

The construction of a quantizer for circular data requires an appropriate distortion measure. In [22] and [34], the weighted squared-error between synthesized waveforms is formulated as a distance measure that involves trigonometric functions. Using trigonometric functions is a plausible way to avoid the pitfalls of phase unwrapping, but it increases the complexity of the quantizer. This section proposes a squared-error-like distortion measure that is suitable for circular spaces. We define the *Wrapped-Squared-Error* (WSE) as

$$d(\psi, \hat{\psi}) = \min_{w\in\mathbb{Z}}\left\{(\psi - \hat{\psi} - w2\pi)^2\right\}. \quad (22)$$

If both ψ and $\hat{\psi}$ are constrained to their principal values in $(0, 2\pi]$, then only ±1 wrappings are enough in (22). The extension of WSE to vectors is straightforward, as follows:

$$d(\vec{\psi}, \hat{\vec{\psi}}) = \sum_{k=1}^{K} d\left(\vec{\psi}(k), \hat{\vec{\psi}}(k)\right). \quad (23)$$

The quantization of data distributed according to the univariate wrapped Gaussian $N_w(\vec{\mu}, \sigma^2)$ is not a trivial task because the shape of the pdf changes with the variance, as shown in Fig. 2. Precomputing and storing a different codebook for each possible rate and variance is not a practical choice due to increased storage requirements. For example, for a 20-dimensional WGMM with 32 wrapped Gaussians, we need $20*32=640$ codebooks. This section proposes two methods to *construct* codebooks for univariate wrapped Gaussian random variables; the first one is fast but suboptimal while the second performs better at the expense of higher complexity.

A. Wrapped Linear Gaussian Codebooks

A simple and fast approach is to construct the wrapped Gaussian codebooks by wrapping the codepoints of a linear Gaussian codebook around the circumference of the unit circle. Let c_{linear} be a codepoint of a linear Gaussian $N(0,\sigma^2)$. The codepoint c_{linear} corresponds to a quantization level obtained by any quantization algorithm (for example k-means) of linear random variables. The corresponding wrapped codepoint then can be obtained using the modulo operation

$$c_{\text{wrapped}} \leftarrow c_{\text{linear}}\ mod\ 2\pi. \tag{24}$$

This solution works quite well for low variances $\sigma^2 \le 1$ because the interval $(0,2\pi]$ contains most of the pdf mass and the overlapping of the tilled Gaussian components is low. However, it becomes less accurate in higher variances. The corresponding quantizers will be referred to as WLGC quantizers.

B. Polynomial Code-Functions

The optimal codebook for a wrapped scalar Gaussian $N_w(\mu,\sigma^2)$ is a function of the variance σ^2 and the number of codepoints Q, since the translation term μ does not effect the shape of the pdf. Assuming that the codepoints evolve smoothly over σ^2, we can construct a codebook for a specific variance σ^2 by sampling a set of precomputed polynomial functions at that point. Let

$$c_q(\sigma^2)=\sum_{p=0}^{P} c_{q,p}\sigma^{2p}, \quad q=1,\dots,Q \tag{25}$$

be the set of polynomial functions that generate a codebook with Q entries for $N_w(0,\sigma^2)$, where each σ^2 belongs to a continuous interval S. These functions could be referred to as "*polynomial codepoint generator functions*," or—in short—as *polynomial code-functions* (PCFs). As it was mentioned previously, the PCF quantizer is based on the assumption that the optimal codepoints evolve smoothly over σ^2 in S. Let P be the order of the polynomial, $c_{q,p}$ its coefficients, and Q be the size of the codebook. Assume that $c_q(.)$ and all circular variables take values in $(-\pi,\pi]$, and that $c_q(.)$ are sorted so that $c_1(\sigma^2) > c_2(\sigma^2) > \dots > c_Q(\sigma^2)$, for all $\sigma^2 \in S$. Since $N_w(0,\sigma^2)$ is symmetric around zero, only $\lfloor Q/2 \rfloor$ PCF are needed, and the following holds:

$$c_q(\sigma^2)=-c_{Q-q+1}(\sigma^2), \quad q=1,\dots,Q. \tag{26}$$

If Q is odd, then the intermediate PCF is zero

$$c_q(\sigma^2)=0, \quad q=\lfloor Q/2 \rfloor+1. \tag{27}$$

Enforcing symmetry using (26) ensures that the partitioning of the unit circle made by the PCF has the angle of π (or equivalently of $-\pi$) at the boundary between the quantization cells of $c_1(\sigma^2)$ and $c_Q(\sigma^2)=-c_1(\sigma^2)$. In this case, no wrappings should be taken into account when the circular random variable $\theta \in (-\pi,\pi]$ is quantized, and then the linear squared error

$$d(\theta,\hat{\theta})=(\theta-\hat{\theta})^2$$

can be used instead of the wrapped-squared-error in (22).

1) Training PCF: An iterative k-means-like algorithm has been developed for the training of PCF for $\sigma^2 \in S$. Let $\sigma_l^2=\{\sigma_1^2,\sigma_2^2,\dots,\sigma_L^2\}$ be L samples of σ^2 in S. For example, $\sigma_l^2=\{0.5,0.52,0.54,\dots,0.68,0.7\}$. Let $\theta_{l,n}$, $l=1,\dots,L$, $n=1,\dots,N$ be N random samples from $N_w(0,\sigma_l^2)$, for each variance σ_l^2. Let $c_{q,p}^{(i)}$ be the PCF parameters resulting from the ith iteration of the algorithm. The algorithm is initialized with constant PCF, uniformly distributed over $(-\pi,\pi]$

$$c_{q,0}^{(0)}=2\pi\left(0.5+\frac{0.5-q}{Q}\right), \quad q=1,\dots,\left\lfloor\frac{Q}{2}\right\rfloor \tag{28}$$

$$c_{q,1}^{(0)}=c_{q,2}^{(0)}=\dots=c_{q,P}^{(0)}=0. \tag{29}$$

Each iteration consists of two steps: a *classification step* and an *optimization step*. The classification step labels each sample $\theta_{l,n}$ to a PCF function, and the optimization step uses these labels to estimate each PCF function. The PCF functions converge after 20 to 50 iterations.

Classification Step
At the ith iteration, the classification step finds the indices $I_{l,n}$, $l=\{1,\dots,L\}$, $n=\{1,\dots,N\}$ that minimize the squared error

$$I_{l,n}=\arg\min_q\left\{\left(\theta_{l,n}-c_q^{(i-1)}\left(\sigma_l^2\right)\right)^2\right\}, \quad q=1,\dots,Q. \tag{30}$$

Optimization Step
Let $\Theta_{l,q}=\{\theta_{l,n}:I_{l,n}=q\}$ be the set of samples that have been classified to the qth PCF for each variance σ_l^2. The optimized qth PCF is the polynomial that best fits the pairs of variables $\{(\sigma_l^2,\theta_l): l=1,\dots,L \;\&\; \theta_l \in \Theta_{l,q}\}$. In other words, we obtain the optimal qth PCF by minimizing the corresponding mean-squared error

$$c_{q,p}^{(k)}=\arg\min_{c_{q,p}}\left\{\sum_{l=1}^{L}\sum_{\theta\in\Theta_{l,q}}\left(\theta-\sum_{p=0}^{P}c_{q,p}\sigma_l^{2p}\right)^2\right\}. \tag{31}$$

The optimization can be made using typical polynomial least squares fitting methods [35].

2) Practical Considerations: The wrapped Gaussian is closely approximated by the linear Gaussian for small variances $\sigma^2 < 0.5$. Therefore, for $\sigma^2 < 0.5$ we can use the wrapped linear Gaussian quantizer presented in Section V-A, while for higher variances we use PCF quantizers. Furthermore, we limit the variance σ^2 to a maximum of 2π. The construction of PCF quantizers depends on two interrelated design parameters: the size of the variance interval S and the order of the PCF polynomial. We found that high-order polynomials cannot provide high-quality PCF for the whole range of interest $S=[0.5,2\pi]$. It is better to divide $[0.5,2\pi]$ into

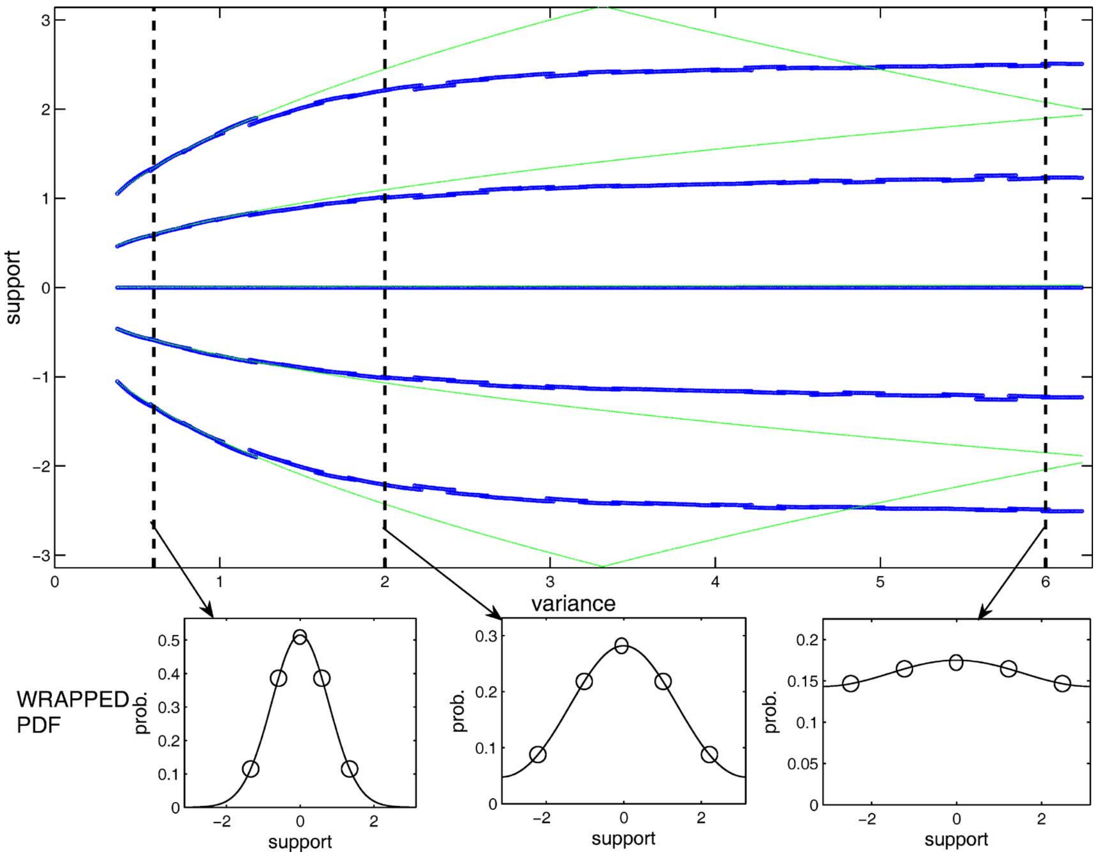

Fig. 3. Codepoint trajectories over σ^2. The thin lines correspond to wrapped linear Gaussian codepoints and the thick lines to PCF generated codepoints. Three wrapped Gaussian pdf's with variances $\sigma^2 = \{0.6, 2, 6\}$ are illustrated along with the corresponding PCF-generated codepoints.

smaller intervals, for example in segments of length 0.2, and to construct low-order (i.e., quadratic) polynomials for each of these intervals.

An example of the trajectories of wrapped linear Gaussian codepoints and PCF over σ^2 is shown in Fig. 3. The wrapped linear Gaussian codepoints are not well distributed for $\sigma > 1$ and they may occasionally coincide, like at $\sigma^2 = 2\pi$, resulting in a practical loss of some quantization points and a distortion penalty. On the other hand, the PCF codefunctions converge to the codepoint allocation of a uniform quantizer, as σ^2 increases.

VI. WGMM-Based Quantization

This section presents an extension of GMM-based quantization schemes for linear data [24] to WGMM-based schemes for circular data. Multivariate GMM-based quantizers provide state-of-the-art performance at low complexity. A GMM-based quantizer is actually a multicoder scheme where the source vector is quantized with a set of *transform coders* to yield a set of *candidate encodings*. Only the "best" candidate encoding is kept, together with the index of the corresponding transform coder. Each of these transform coders is constructed according to a multivariate Gaussian component of the GMM. The efficiency of GMM-based quantization arises from the fact that it separates the estimation of the statistics of the source from the allocation of codepoints. The same idea can be extended to circular data using WGMM instead of GMM. Therefore, in WGMM-based quantization, the source vector is encoded according to a set of transform coders and the encoding that provides the minimum vector-WSE [see (23)] is kept together with the index of the corresponding transform coder. The process is depicted in Fig. 4.

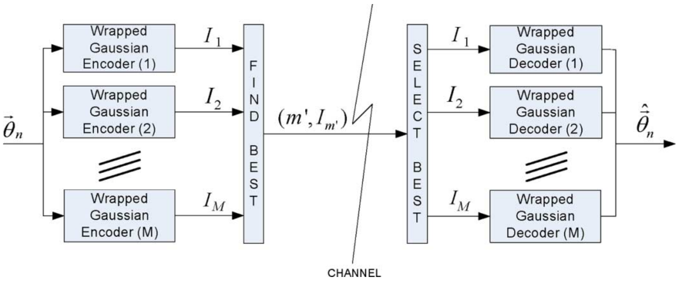
$\vec{\theta}_n$
Wrapped Gaussian Encoder (1)
Wrapped Gaussian Encoder (2)
Wrapped Gaussian Encoder (M)
I_1
I_2
I_M
FIND BEST
$(m', I_{m'})$
CHANNEL
SELECT BEST
Wrapped Gaussian Decoder (1)
Wrapped Gaussian Decoder (2)
Wrapped Gaussian Decoder (M)
$\hat{\vec{\theta}}_n$

Fig. 4. Basic scheme for WGMM-based vector quantization. I_m is the index of the encoding made by the mth encoder, while m' is the index of the "best" encoding.

The construction of a transform coder for the mth multivariate wrapped Gaussian is trivial if the covariance matrix is diagonal. In that case, each dimension k is quantized according to the mean $\vec{\mu}_m(k)$ and the variance $\sigma^2_m(k)$ of the corresponding univariate wrapped Gaussian $N_w(\vec{\mu}_m(k), \sigma^2_m(k))$. Scalar wrapped Gaussian quantization was discussed in Section V. In this section, the allocation of bits to each Gaussian component and each dimension will be discussed.

Let R be the rate of the WGMM-based quantizer and $N_m = \lfloor \alpha_m 2^R \rfloor$ be the number of quantization levels that is assigned to each of the M components of the WGMM. Within each Gaussian component, the N_m quantization levels are allocated with a greedy algorithm that minimizes the expected component distortion D_m

$$D_m = \sum_{k=1}^{K} D\left(N_{m,k}, \sigma_m^2(k)\right) \tag{32}$$

where $D(N_{m,k}, \sigma_m^2(k))$ is the expected WSE when the kth variable of the mth wrapped Gaussian component is encoded with $N_{m,k}$ quantization levels. The minimization is made subject to the rate constrain

$$\prod_{k=1}^{K} N_{m,k} \leq N_m.$$

When the variances $\sigma_m^2(k) \leq 0.5$, the wrapped univariate Gaussian is well approximated by a linear Gaussian and the well-known distortion-rate formula for linear Gaussians can be used [36, p. 228]

$$D(N, \sigma^2) = \frac{\sqrt{3}\pi}{2}\sigma^2 N^{-2}. \tag{33}$$

For higher variances, $\sigma_m^2(k) > 0.5$, we use linear interpolation of tabulated distortions, sampled for a wide range of quantization levels and variances. The same algorithm is applicable to both PCF-based and wrapped-linear quantizers that were presented in Section V. The distortions were computed using 100 000 samples of a wrapped $N_w(0, \sigma^2)$ and evaluated with the WSE, for quantization levels $l = 1, 2, \ldots, 2^6$ and for a dense sampling set of variances $\sigma^2 = \{0.5, 0.51, 0.52, \ldots, 2\pi\}$.

VII. Quantizing Harmonic Phases of Speech

The proposed WGMM-based scheme is used to quantize the phases of a prototype *narrowband* sinusoidal codec. Initially, phases are decomposed according to (2). The minimum phase part is obtained from the cepstral envelope that fits the harmonic amplitudes as discussed in Section II. The translation term τ is encoded in relation to the pitch using 7 bits/frame, while pitch is encoded with 8 bits/frame. Speech is analyzed and synthesized using 20-ms frames and Hanning windows with a 10-ms steps at a total rate of 100 frames/s. Each speech frame is classified as "*silent*," "*unvoiced*," "*transitional*," or "*voiced*" [37]. For transitional and voiced frames, the dispersion phase term $\vec{\psi}$ is quantized using the proposed WGMM-based scheme. For unvoiced and silent frames, the dispersion phase term is randomly set according to a uniform distribution. In all cases, the harmonics are synthesized up to the frequency of 3700 Hz. For notational simplicity, we will henceforth refer to the quantization of dispersion phases as "*phase quantization*."

An intrinsic difficulty in phase quantization is the variable dimensionality of the dispersion phase vectors $\vec{\psi}$. We address this problem by classifying pitch values in seven classes (continuous intervals), referred to as Q1 to Q7 in Table I, in order to reduce the variance of the dimensionality within each class. Note that this classification is just a plausible choice and that it is not critical for efficiency. The harmonics are separated into two bands; a low- and a high-frequency band, in order to provide more bits to the perceptually important low-frequency harmonics. A fixed number of low-frequency harmonics (depending on pitch class Q_i) are grouped together to form the lower-band dispersion phase vectors. For pitch classes Q1 and Q2, the lower-band consists of the first 24 harmonics. For pitch classes Q3 to Q6, the number of dimensions of the low-frequency harmonics is equal to the minimum size of the phase vectors of the corresponding class. For example, for class Q5, the number of harmonics for the low-frequency band is given by: $\lfloor 3700/217 \rfloor = 17$ harmonics, where 217 is the lower pitch in Q5 and 3700 is the bandwidth of the speech signal. The bandwidth of the lower frequency band varies with the number of harmonics and the pitch. For the first six classes, Q1 to Q6, 6 fixed-dimension low-frequency dispersion phase datasets are obtained from the TIMIT database [31]. The number of dimensions of each dataset is shown in Table I. Two more datasets are obtained for the high-frequency phases of pitch classes Q1 and Q2 with a size of 14 and 8 dimensions, respectively. These phases correspond to the first harmonics of the high-frequency band. An example is provided in Table II: assume that the pitch is 100 Hz, resulting to a total of $\lfloor 3700/100 \rfloor = 37$ phases. The first 24 phases are used to train the low-frequency WGMM, while phases ψ_{25} to ψ_{32} are used to train the high-frequency WGMM. Concluding, we derive six datasets from the low-frequency band and two datasets from the high-frequency band. These datasets are used to train the corresponding WGMM according to Section IV. The circular mean [(5)] of each dataset is removed prior to training. This procedure moves the wrapped multivariate Gaussians closer to the center of the principal hypercube $(0, 2\pi]^K$ and increases the accuracy of the approximation that is made using only ± 2 tilings of each scalar Gaussian dimension. The number of dimensions of each low-frequency and high-frequency WGMM is shown in the two right-most columns of Table I. For the training of WGMM, we used the training data set of the TIMIT database [31].

TABLE I
Pitch Classes for WGMM-Based Vector Quantization of Phases. The Phases Are Classified According to Their Pitch Value (Second Column). Columns 3 and 4 Show the Number of Phases (Dimensions) Quantized in the Low-Frequency and the High-Frequency WGMM, Respectively. Whenever the Dimensionality Varies According to Pitch, the Symbols < and > Are Used to Indicate That the Following Number Is the Minimum and the Maximum Number of Phases, Respectively

Pitch Class	Pitch Range	Low-Freq. WGMM dims.	High-Freq. WGMM dims.
Q1	<95 Hz	24	>14
Q2	95-115 Hz	24	>8
Q3	115-142 Hz	24	>0
Q4	142-176 Hz	21	>0
Q5	176-217 Hz	17	>0
Q6	217-250 Hz	14	>0
Q7	>250 Hz	<14	0

In most frames, the trained wrapped Gaussian mixture models do not model all the harmonics and the statistics of a variable

TABLE II
THREE EXAMPLES OF DISPERSION PHASE VECTORS WITH PITCH VALUES $f_0 = 100, 150$, AND 300 Hz. PHASES IN BRACKETS ARE MODELED BY A WGMM TRAINED FROM DATA. PHASES OUTSIDE BRACKETS ARE MODELED BY THE "EXTENDED" WGMM

Pitch Class	f_0	Low Frequency Harmonics	High Frequency Harmonics
Q2	100 Hz	$\underbrace{[\psi_1, ..., \psi_{24}]}_{\text{Quantized}}$	$\underbrace{[\psi_{25}, ..., \psi_{32}], \psi_{33}, ..., \psi_{37}}_{\text{Quantized}}$
Q4	150 Hz	$\underbrace{[\psi_1, ..., \psi_{21}]}_{\text{Quantized}}$	$\underbrace{\psi_{22}, \psi_{23}, \psi_{24}}_{\text{Quantized}}$
Q7	300 Hz	$[\underbrace{\psi_1, ..., \psi_{12}}_{\text{Quantized}}, \underbrace{\psi_{13}, \psi_{14}}_{\text{Ignored}}]$	$\oslash$

number of high-frequency harmonics are not captured. In the examples provided in Table II these phases are shown to be outside the brackets. However, these harmonics are in high frequencies where the ear is less sensitive to individual phase distortions. Furthermore, we have observed that the bivariate marginal distributions of high-frequency harmonics have similar statistics. Therefore, for each frame we construct a high-band WGMM by replicating the means and the variances of the dispersion phase with the highest frequency that is modeled by a WGMM. Precisely, for pitch classes Q1 and Q2, where a high-frequency WGMM is already trained with 14 and 8 dimensions, respectively, the trained WGMM is *extended* to the total number of harmonics using the means and the variances of the last dimension of the latter WGMM. In the first example of Table II, this means that the statistics of phases ψ_{33} to ψ_{37} are obtained from the statistics of ψ_{32}. For pitch classes Q3 to Q6, the high-frequency WGMM is constructed using the statistics of the last dimension of the corresponding low-frequency WGMM. In the second example of Table II, the statistics of ψ_{22} to ψ_{24} are obtained from the statistics of ψ_{21}. A different procedure is used for the high-frequency class Q7 (above 250 Hz). Assuming that classes Q6 and Q7 have similar statistics, the phases of Q7 are modeled by *removing* the necessary number of higher frequency harmonics from the low-frequency WGMM of Q6. Therefore, we have *less than* 14 dimensions as it is stated in Table I and illustrated in Table II.

VIII. Evaluation

An objective evaluation of the performance of the quantizers is made using the SNR criterion between the original excitation signal [(3)] and the synthesized excitation signal with quantized dispersion phases, for a duration of two pitch periods. Fig. 5 depicts the measured SNR/rate relationship for the quantization of the low-band dispersion phases of pitch class Q2 using several types of quantizers: two WGMM-based that employ WLGC (WLGC) and PCF-based (PCF) scalar quantization, and four GMM-based with 16 or 32 classes and diagonal or full covariance matrices. Three of the latter are conventional GMM-based quantizers $(\mathrm{GMM16}_{\mathrm{full}}, \mathrm{GMM32}_{\mathrm{full}}, \mathrm{GMM32}_{\mathrm{diag}})$, similar to those proposed in [24], but with the modification that scalar quantization is made using precomputed codebooks instead of companding/quantization. This modification improves the performance in high-dimensional sources where companding/quantization is forced to operate at low rate/dimension where it is suboptimal. The fourth GMM-based quantizer $(\mathrm{GMM32}_{\mathrm{comp}})$ is the one presented in [14] and employs a bounded-support compander tuned to operate in $(-\pi, \pi]$ in order to avoid quantization outside the principal values of the phases. Following the strategy of [14], we employed a better bit-allocation algorithm that improves the performance of $\mathrm{GMM32}_{\mathrm{comp}}$. Unfortunately, due to the varying boundaries of the scalar quantizers in $\mathrm{GMM32}_{\mathrm{comp}}$, it is not possible to replace companding/quantization with codebook-based scalar quantization, something that is reflected in the performance of the latter method.

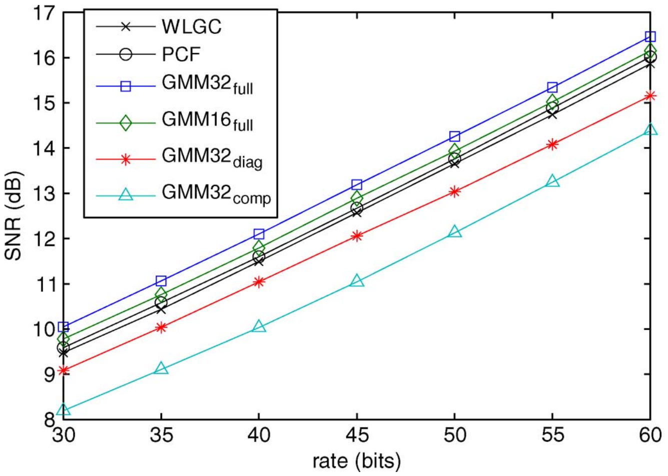

Fig. 5. Measured distortion-rate curves of the excitation signal of Q2 class using several quantizers; WGMM-based with WLGC or PCF scalar quantization, GMM-based with 16 or 32 classes and diagonal or full covariance matrices $(\mathrm{GMM16}_{\mathrm{full}}, \mathrm{GMM32}_{\mathrm{full}}, \mathrm{GMM32}_{\mathrm{diag}})$, and GMM-based with 32 classes, diagonal covariance matrices and bounded-support companding $(\mathrm{GMM32}_{\mathrm{comp}})$.

Fig. 5 shows that the two GMM-based quantizers with full covariance matrices (and 16, 32 classes) outperform WGMM-based quantization. This can be primarily attributed to fact that the shape of the underlying pdf has most of its mass concentrated around the center of the 24-dimensional hypercube and secondarily to the increased number of parameters that these quantizers use to model the source. As a limit case, we could argue that when the effective support of the underlying pdf is much smaller than the $(0, 2\pi]^K$ hypercube, the circular pdf can be well approximated by a linear pdf, allowing typical quantization methods to be employed. This is not the case for the particular experiment, however, since the GMM-based quantizer with 16 classes and full covariance matrices has only slightly better performance than the PCF-based circular quantizer, although it uses much more parameters.

A fair comparison can be made only between systems with the same degree of freedom; in our case, between both WGMM-based quantizers and the two GMM-based quantizers with 32 classes and diagonal covariance matrices. It can been seen that WGMM-based quantization offers a gain of about 0.5-dB SNR over conventional GMM-based quantization and a gain of about 1.2-dB SNR over the bounded-support method. The surprising (perhaps) observation that $\mathrm{GMM32}_{\mathrm{comp}}$ is outperformed by conventional GMM-based quantization can be partially attributed to that companding/quantization operates in low rates of about 1–2 bit/dimension where it is suboptimal (it

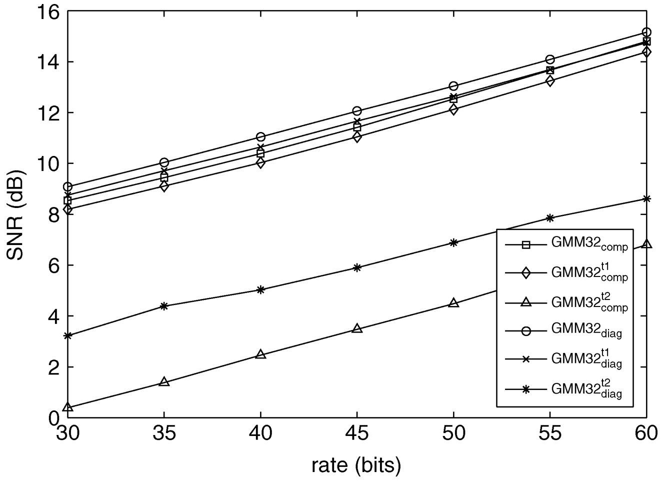

Fig. 6. Measured distortion-rate curves when the source is translated. The nomenclature follows the one in Fig. 6. The superscript $t1$ corresponds to a translation that restores the circular mean of the original dataset, while the superscript $t2$ corresponds to a translation by $\pi\vec{1}$.

requires rates above 4 bits/dimension to reach the performance of codebook-based scalar quantization). Finally, note that the PCF quantizer outperforms the WLGC quantizer in terms of SNR by less than 1 bit and that similar results are also obtained for the other pitch classes.

In the previous experiment, the linear-space methods benefited from a pdf that is concentrated around the center of the K-dimensional hypercube. It is interesting to investigate what happens when the mass of the pdf is not so well-behaved. This can be simulated, for example, by introducing a circular translation to the source vectors. Therefore, we constructed two other sources by translating the previous source by 1) the minus circular mean of the original dataset and 2) $\pi\vec{1}$. The phases are restricted to their principal value in the $(-\pi, \pi]^K$ hypercube by a modulo-2π operation. Recalling that the source we quantized was the original dispersion phases minus the corresponding circular mean, the first translation is actually restoring the source at its original position. The second translation is a worst case scenario where the probability mass is moved from the center of the hypercube to the edge where the modulo-2π operation scatters the data to all the edges of the hypercube. Fig. 6 shows the performance of $\mathrm{GMM32}_{\mathrm{diag}}$ and $\mathrm{GMM32}_{\mathrm{comp}}$ under these two translations. Note that the performance of WGMM-based methods is not effected by any translation of the source and therefore it is not depicted for the purpose of clarity. We can observe that the removal of the circular mean of the dataset prior to modeling/quantization gave to GMM-based methods a benefit of about 0.3-dB SNR. Furthermore, the degradation for a translation of $\pi\vec{1}$ is severe. This experiment reveals the pitfalls of conducting phase quantization using linear phases: the performance of the quantizer is *abnormally* dependent on the location and the structure of the mass of the pdf inside the hypercube. On the contrary, statistics and quantizers that are natively circular do not suffer from these deficiencies and can be used at any case.

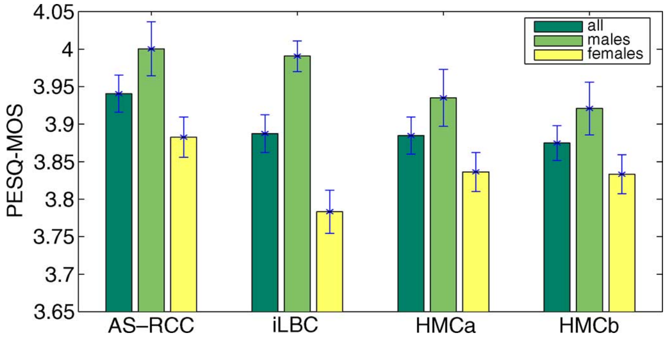

Fig. 7. PESQ-MOS evaluation (mean and 95% confidence interval) of the HMC codec, iLBC, and the analysis/synthesis system.

Finally, it should be emphasized that the SNR does not necessarily reflect the perceptual performance of the quantizer. Ideally, we would construct and evaluate the quantizer according to a perceptually motivated distortion measure. However, the determination of a perceptually motivated distortion metric for harmonic phases of speech is not trivial and, in the authors' perspective, is still an open issue. For a discussion regarding phase distortion and perception the interested reader is referred to [2], [22] and [23].

Another objective evaluation is made in terms of PESQ-MOS measured between the original speech signal and the speech signal that is synthesized by a prototype harmonic codec referred to as *Harmonic Model Codec* (HMC) [37], [38]. Pitch was quantized with 8 bits, frame energy with 8 bits, the linear phase term τ was quantized with 7 bits, and the voicing condition with 3 bits. The RCC parameters were encoded with 50 bits for transitional frames and 60 bits for unvoiced and voiced frames. Two different cases were examined: **HMCa** with 70 bits for the low-frequency WGMM and 30 bits for the high-frequency WGMM, and **HMCb** with 60 bits for the low-frequency WGMM and 20 bits for the high-frequency WGMM. Both cases used PCF-based quantization. Codec HMCa requires an average of 14.2 kbps (max. 18.6 kbps) and codec HMCb an average of 12.9 kbps (max. 16.6 kbps). The evaluation is made using PESQ-MOS [28] computed with a test-set of 64 male and 64 female utterances from the test-set of TIMIT database. For comparison purposes, a baseline system with unquantized harmonic amplitudes (although computed via the RCC spectral envelope) and phases has been developed (**AS-RCC**). Furthermore, we evaluated the PESQ-MOS provided by iLBC [29] codec (**iLBC**) at the 20 ms mode (15.2 kbps). The **AS-RCC** system was chosen in order to provide an upper-bound to the measured quality, since it uses the unquantized phases, while the iLBC codec was chosen as a reference codec that also encodes speech in a packet-independent manner (without using previous packets), as HMC does [37]. Results are shown in Fig. 7. We can observe that the 12.9 kbps HMCb codec is more-or-less equivalent to iLBC in terms of PESQ-MOS score. This observation is also supported by informal subjective listening tests.

A subjective evaluation was conducted using the Degradation Category Rating (DCR) test. The listeners were presented with two stimuli, the original and the encoded speech signal, and were asked to evaluate the degradation of the perceptual quality

TABLE III
DCR TEST SCALE

Description	Rating
Degradation is not perceived	5
Degradation is perceived but not annoying	4
Degradation is slightly annoying	3
Degradation is annoying	2
Degradation is very annoying	1

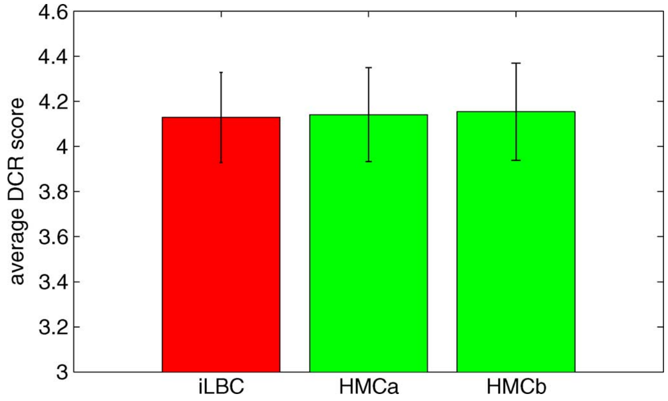

Fig. 8. Subjective evaluation (mean and 95% confidence interval) of the HMC codec and iLBC according to the DCR test.

that the quantization process introduced to the encoded signal. The degradation was graded according to the DCR scale, which is presented in Table III. A total of 16 listeners participated on this test. The samples were randomly drawn from a small database that was constructed from 15 male and 15 female utterances from the TIMIT test-set. All signals were low-pass-filtered and decimated to an 8-kHz sampling rate. The results are shown in Fig. 8. Both HMC codecs have similar performance to iLBC around 4.13 in terms of DCR score.

There are many ways to interpret these results; for example, note that the proposed phase quantization method is capable of providing high-quality phase quantization at a rate that is comparable to the rate needed for the quantization of spectral envelopes. This is a surprising result because phase information was always considered to require a much higher rate than spectral information that is readily structured. If we take into account that no perceptual weighting is made, we can speculate that there is a lot of space for improvement. From another point of view, and given the amount of knowledge that is now accumulated for CELP codecs, it is interesting to see that a sinusoidal codec based on the proposed phase quantization method competes a CELP codec that is equivalent in the sense that they both encode each packet independently of the previous packets. However, the intention of this paper is to demonstrate an application of the proposed WGMM-based quantization to speech coding and not to provide a thorough evaluation of the proposed method regarding speech coding. Details regarding the HMC framework of speech coding for Voice-over-IP are beyond the scope of this paper and can be found in [37].

Summarizing, the presented work provides a *robust* framework for addressing the pdf of phases natively in their circular space. Then, it incrementally recasts well-known source coding algorithms to the specifics of the circular space; circular distortion metrics, scalar quantization for wrapped Gaussians, transform coding, and bit-allocation for circular spaces, up to mixture-based quantization. Since the new algorithms operate natively in the circular space, they feature improved performance and they are significantly more robust than their linear-space counterparts, as it is demonstrated by the experiments conducted in this section. Finally, we can argue that phase information has a specific structure that is revealed after the determination of a linear phase term and allows us to benefit from vector quantization. A number of questions rises regarding the qualitative nature of phase information; for example, does it contain user specific information or phoneme specific information (i.e., for plosives and consonants) that is useful for speaker or speech recognition, respectively? The proposed work provides an analysis procedure and a statistical framework suitable for addressing these issues.

IX. CONCLUSION

This paper presents a methodology to model and quantize phases of speech. Phases are decomposed into three parts: a linear translation term, a minimum phase term, and a residual term referred to as *dispersion* phase. The statistics of dispersion phases are captured using WGMM. These models are then used to construct quantizers for phase data. Related issues like scalar quantization of circular data have also been addressed and solutions have been suggested. The effectiveness and the robustness of the proposed quantizer in comparison to alternative methods is demonstrated experimentally, while a practical evaluation is made in the context of sinusoidal coding of speech. The outcome of this work can find application in high-quality sinusoidal codecs suitable for VoIP and reduced footprint text-to-speech systems.

APPENDIX
SUMMATION-PRODUCT INTERCHANGE

This appendix provides a three-dimensional example of the interchange between summation and product that can be used to simplify the computation of (12). The extension to more than three dimensions is straightforward. Let $p(\theta, u; \sigma, \mu) = (1/\sqrt{2\pi\sigma^2})e^{-((\theta-\mu-u2\pi)^2/2\sigma^2)}$, $\theta \in (0.2\pi]$ be a single wrap of the Gaussian kernel. Then

$$\begin{aligned}\sum_{w_1}\sum_{w_2}\sum_{w_3} & p\left(\theta_1, w_1; \sigma_1^2, \mu_1\right) p\left(\theta_2, w_2; \sigma_2^2, \mu_2\right) \\ & \times p\left(\theta_3, w_3; \sigma_3^2, \mu_3\right) \\ & = \sum_{w_1} p\left(\theta_1, w_1; \sigma_1^2, \mu_1\right) \sum_{w_2} p\left(\theta_2, w_2; \sigma_2^2, \mu_2\right) \\ & \times \sum_{w_3} p\left(\theta_3, w_3; \sigma_3^2, \mu_3\right) \\ & = \left(\sum_{w_3} p\left(\theta_3, w_3; \sigma_3^2, \mu_3\right)\right)\left(\sum_{w_2} p\left(\theta_2, w_2; \sigma_2^2, \mu_2\right)\right)\cdots \\ & \left(\sum_{w_1} p\left(\theta_1, w_1; \sigma_1^2, \mu_1\right)\right).\end{aligned}$$

References

[1] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 2117–2120.

[2] D.-S. Kim, "On the perceptual irrelevant phase information in sinusoidal representation of speech," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 900–905, Nov. 2001.

[3] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Commun.*, vol. 22, no. 4, pp. 403–407, 1997.

[4] A. M. Kondoz*, Digital Speech: Coding for Low Bit Rate Communication Systems*. New York: Wiley, 2004.

[5] R. J. McAulay and T. F. Quatieri, "Sinusoidal transform coding," in *Proc. Mobile Satellite Conf., Jet Propulsion Lab*, May 1988, pp. 503–508, see N88-25680 19-32.

[6] D. Griffin and J. Lim, "Multi-band excitation vocoder," in *Proc. ICASSP*, New York, Apr. 1988, vol. 36, pp. 1223–1235.

[7] B. W. Kleijn and K. K. Paliwal*, Speech Coding and Synthesis*. New York: Elsevier, 1995.

[8] T. F. Quatieri*, Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice-Hall, 2001.

[9] S. Ahmadi and A. Spanias, "A new sinusoidal phase modelling algorithm," in *Proc. ICASSP*, Munich, Germany, 1997, vol. 3, pp. 1675–1678.

[10] P. Hedelin, "Phase compensation in all-pole speech analysis," in *Proc. ICASSP*, New York, Apr. 1988, pp. 339–342.

[11] D. L. Thomson, "Parametric models of the magnitude/phase spectrum for harmonic speech coding," in *Proc. ICASSP*, 1988, vol. 1, pp. 378–381.

[12] J. S. Marques, L. B. Almeida, and J. M. Tribolet, "Harmonic coding at 4.8 kb/s," in *Proc. ICASSP*, 1990, vol. 1, pp. 17–20.

[13] O. Gottesman, "Enhanced waveform interpolative coding at low bitrate," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 786–798, 2001.

[14] J. Lindblom, "A sinusoidal voice over packet coder tailored for the frame-erasure channel," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 787–798, Sep. 2005.

[15] X. Sun, B. Cheetham, and W. Wong, "Spectral envelope and phase optimization for sinusoidal speech coding," in *Proc. IEEE Workshop Speech Coding for Telecomm.*, Annapolis, MD, 1995, pp. 75–76.

[16] A. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, no. 2, pp. 583–590, 1971.

[17] X. Sun, F. Plante, B. M. Cheetham, and K. W. Wong, "Phase modelling of speech excitation for low bit-rate sinusoidal transform coding," in *Proc. ICASSP*, Munich, Germany, Apr. 1997, vol. 3, pp. 1691–1694.

[18] Y. Stylianou, "On the implementation of the harmonics-plus-noise model for concantenative speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, 2000, vol. 2, pp. 957–960.

[19] D. Chazan, R. Hoory, Z. Kons, D. Silberstein, and A. Sorin, "Reducing the footprint of the IBM trainable synthesis system," in *Proc. 7th Int. Conf. Spoken Lang. Process.*, Denver, CO, 2002, pp. 2381–2384.

[20] G. Aguilar, J.-H. Chen, R. B. Dunn, and R. J. McAulay, "An embedded sinusoidal transform codec with measured phases and sampling rate scalability," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 141–144.

[21] D. Chazan, R. Hoory, Z. Kons, A. Sagi, S. Shechtman, and A. Sorin, "Small footprint concatenative text-to-speech synthesis system using complex spectral envelope modeling," in *Proc. Interspeech*, 2005, pp. 2569–2572.

[22] H. Pobloth and W. B. Kleijn, "Squared error as a measure of perceived phase distortion," *J. Acoust. Soc. Amer.*, vol. 114, no. 2, pp. 1081–1094, 2003.

[23] D.-S. Kim and M. Y. Kim, "On the perceptual weighting function for phase quantization of speech," in *Proc. IEEE Workshop Speech Coding*, Delavan, WI, 2000, pp. 62–64.

[24] A. D. Subramaniam, "Gaussian mixture models in compression and communication," Ph.D. dissertation, Univ. of California, San Diego, 2003.

[25] K. Mardia*, Statistics of Directional Data*. New York: Academic, 1972.

[26] P. Smaragdis and P. Boufounos, "Learning source trajectories using wrapped-phase Hidden Markov models," in *Proc. IEEE Workshop Applications of Signal Process. to Audio Acoust.*, New Paltz, NY, 2005, pp. 114–117.

[27] H. Holzmann, A. Munk, M. Suster, and W. Zucchini, "Hidden Markov models for circular and linear-circular time series," *J. Environ. Ecol. Statist.*, vol. 13, no. 3, pp. 325–347, 2006.

[28] "Perceptual evaluation of speech quality assessment of narrowband telephone networks and speech codecs," ITU-T, Feb. 2001.

[29] S. V. Andersen, W. B. Kleijn, R. Hagen, J. Linden, M. N. Murthi, and J. Skoglund, "iLBC—A linear predictive coder with robustness to packet losses," in *Proc. IEEE Workshop Speech Coding*, Tsukuba, Ibaraki, Japan, 2002, pp. 23–25.

[30] Y. Stylianou, "Harmonic-plus-noise models for speech, combined with statistical methods for speech and speaker modification," Ph.D. dissertation, Ecole Nationale. Superieure des Telecomm., Paris, France, 1996.

[31] J. S. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," in *Proc. Linguist. Data Consortium*, 1993.

[32] C. Bahlmann, "Directional features in online handwriting recognition," *Pattern Recognition*, vol. 39, pp. 115–125, 2006.

[33] G. J. M. T. Krishnan*, The EM Algorithm and EXTENSIONS*. New York: Wiley, 1997.

[34] Y. Jiang and V. Cuperman, "Encoding prototype waveforms using a phase codebook," in *Proc. IEEE Workshop Speech Coding for Telecommunications*, 1995, pp. 21–22.

[35] A. Bjõrck*, Numerical Methods for Least Squares Problems*. Philadelphia, PA: SIAM, Dec. 1996.

[36] A. Gersho and R. M. Gray*, Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.

[37] Y. Agiomyrgiannakis, "Sinusoidal coding of speech for Voice-Over-IP" Ph.D. dissertation, Dept. of Comput. Sci., Univ. of Crete, Crete, Greece, Jan. 2007 [Online]. Available: http://www.csd.uoc.gr/~jagiom/voip.htm

[38] Y. Agiomyrgiannakis and Y. Stylianou, "The harmonic model codec framework for VoIP," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1681–1684.

Yannis Agiomyrgiannakis received the B.Sc. degree in computer science and the M.Sc. degree in networks and telecommunication, and the Ph.D. degree regarding sinusoidal speech coding for VoIP from the University of Crete, Crete, Greece, in 1999, 2002, and 2007, respectively.

He is currently holding a postdoctoral position regarding speech synthesis in Orange Labs, Rennes, France. He has worked on low footprint DSP implementations of speech coding and speech processing algorithms. His research interests include digital signal processing, analysis/synthesis and coding of speech signals, source/channel coding, Voice-over-IP, glottal inverse filtering, text-to-speech synthesis, and voice conversion/transformation.

Yannis Stylianou (M'95) received the Diploma in electrical engineering from NTUA, Athens, Greece, and M.Sc. and Ph.D. degrees in signal processing from ENST, Paris, France.

He is an Associate Professor in the Department of Computer Science, University of Crete. He was with AT&T Labs and Bell Labs from 1996 to 2002. Since 2002, he has been with the University of Crete. He is member of the Technical Committee of the IEEE for Speech and Language Processing, Associate Editor of the *EURASIP Journal on Speech, Audio and Music Processing* and *EURASIP Research Letters in Signal Processing* and Vice-Chairman of the Cost Action 2103 on Voice Function Assessment. He holds nine U.S. patents in speech signal processing and speech synthesis. His current research focuses on speech signal processing algorithms for speech analysis, speech representation and modification, statistical signal processing, and time-series analysis and modeling.

He is member of the Technical Chamber of Greece, TEE.