

RESEARCH

A uniform phase representation for the harmonic model in speech synthesis applications

Gilles Degottex^{1,2*†} and Daniel Erro^{3,4}

Abstract

Feature-based vocoders, e.g. STRAIGHT, offer a way to manipulate the perceived characteristics of the speech signal in speech transformation and synthesis. For the harmonic model, which provide excellent perceived quality, features for the amplitude parameters already exist (e.g. LSF, MFCC). However, because of the wrapping of the phase parameters, phase features are more difficult to design. To randomize the phase of the harmonic model during synthesis, a voicing feature is commonly used, which distinguishes voiced and unvoiced segments. However, voice production allows smooth transitions between voiced/unvoiced states which makes voicing segmentation sometimes tricky to estimate. In this article, two phase features are suggested to represent the phase of the harmonic model in a uniform way, without voicing decision. The synthesis quality of the resulting vocoder has been evaluated, using subjective listening tests, in the context of resynthesis, pitch scaling and HMM-based synthesis. The experiments show that the suggested signal model is comparable to STRAIGHT or even better in some scenarios. They also reveal some limitations of the harmonic framework itself in case of high fundamental frequencies.

Keywords: Speech synthesis; Harmonic model; Phase modeling; Voice transformation; Parametric speech synthesis

1 Introduction

Parametric speech signal representations are necessary in almost every field of speech technologies: speech and speaker recognition [1, 2], speech and speaker transformation [3], synthesis [4], diarization [5], etc. Each of these fields, however, requires a particular type of parametrization scheme. Thus, while low-dimensional filter bank based Mel-frequency cepstral coefficients (MFCC) [6] are sufficiently accurate for recognition purposes, they are not suitable for speech reconstruction by themselves. Indeed, applications involving spoken outputs, such as speech coding [7], require the speech signals to be represented by a set of features yielding almost transparent analysis/resynthesis. Voice transformation and speech synthesis impose even stricter requirements, since the parametric speech representations they deal with must provide a solid and flexible framework to sculpt all the characteristics of the speech sounds through direct manipulation of the features (see, for instance, [8, 9, 10]). Interestingly, recent statistical trends are also encouraging research

on parametric speech representations with a constant number of parameters and with good mathematical properties [4].

Sinusoidal models represent the speech signal by means of a sum of sinusoids given by their instantaneous frequency, amplitude and phase [11]. These models have been widely used for speech analysis, resynthesis and modification [12, 13, 14]. Sinusoidal models have evolved over the years [3, 15] and recently the so called adaptive Harmonic Model (aHM) [16] has also been shown to yield practically transparent analysis/resynthesis and excellent modification performance [17, 18]. Despite the inherent assumption that speech can be represented only by harmonic sinusoidal components, even in unvoiced segments, aHM succeeds at capturing the relevant spectral information and noisy nature of a speech signal, thus, representing the speech signal in a uniform way, without using any voicing decision. As long as the perceptual information carried in the phase is preserved, the uniform way of describing and manipulating signals is a remarkable practical advantage of aHM with respect to alternative models involving an explicit separation between harmonics and noise [19] for two main reasons: (i) locating the voicing boundaries is an error-prone process; (ii) in voice

*Correspondence: degottex@csd.uoc.gr

¹Computer Science Department, University of Crete (UOC-CSD), Heraklion, Greece

Full list of author information is available at the end of the article

[†]Equal contributor

transformation, such a separation implies the need for two independent modification procedures, one for each component [3, 20], which increases the risk that listeners perceive them as two independent output signals.

The features handled by aHM are not directly compatible with methods involving statistical modeling because the amplitude and phase parameters lie on the harmonic grid which is dependent on the fundamental frequency f_0 [21]. To avoid this issue, amplitude and phase information have to be isolated from f_0 and translated into independent parameters. However, while amplitude envelopes are relatively easy to obtain through interpolation between sinusoidal amplitudes [22, 23], the representation of phase remains an open problem. Recent attempts of obtaining a consistent phase envelope [24, 25, 26, 27] provide features which are theoretically valid in voiced time-frequency regions but are not informative in unvoiced ones. Thus, standard speech parametrization systems used in statistical frameworks tend to discard the phase information. Instead, they rely on a minimum-phase component derived from the amplitude envelope, along with complementary parameters related to the degree of harmonicity in different time-frequency regions, such as band aperiodicities [28] or maximum voiced frequency [29].

This article presents a novel phase representation that has been designed to handle, in a uniform manner across time, all the relevant information conveyed by the phase parameters of a full-band aHM model, namely the maximum-phase component and the noisiness. This is done through the following steps: first, aHM analysis [16] is performed to obtain the instantaneous phase from the waveform; then, the minimum-phase term is subtracted from the measured phases and the local Phase Distortion (PD) [25] is calculated; finally, the short-time mean and standard deviation of the PD are computed in the neighborhood of each frame, the former being highly correlated to the maximum-phase component and the latter to the degree of noisiness. Among the advantages of this novel approach, we can mention the following: (i) It is valid to analyze signals exhibiting harmonic and noise components that overlap both in time and in frequency, thus avoiding binary voiced/unvoiced decisions which are error-prone and result in annoying artifacts, especially in synthesis [30, 31]. (ii) Since it helps avoiding an explicit separation between harmonics and noise, it provides a solid and uniform framework for speech manipulation, thus avoiding artifacts near the voicing boundaries [21]. (iii) It can be easily made compatible with statistical frameworks. Moreover, given the continuous nature of the feature streams, the use of multi-space distributions (MSD) [32] can be avoided. In that sense, the involved training and generation

procedures can be simplified. In addition to these advantages, the suggested phase representation facilitates the study of the perceptual importance of the maximum-phase component and the degree of noisiness, which are linked to separate features. Indeed, phase perception is still source of controversy in speech processing [33, 34, 35].

The next section first summarizes the low-level analysis of the speech signal using the aHM model. Then, the novel phase features based on the mean and standard-deviation of PD are described in details, which is followed by the description of the synthesis step. Finally, the evaluation section will show the importance of the features and demonstrate the feasibility of the suggested representation in the context of voice transformation and speech synthesis.

2 The Adaptive Harmonic Model (aHM)

Given a speech waveform $s(t)$, we assume that its continuous fundamental frequency curve $f_0(t)$ can be known a priori, thanks to numerous existing methods. For the experiments described in this article, the STRAIGHT method [10] has been used, which allows fair comparisons during evaluation. The speech waveform is first segmented into analysis frames centered around time instants t_i . For reason of clarity, a constant step size will be first assumed (e.g. 5ms). Pitch synchronous analysis will be used later on, for statistical characterization in Sec. 3.4. At each time instant t_i , the main goal is to represent the frequency content of each frame using features capturing independent characteristics of the speech signal. For this purpose, in a Blackman window of 3 pitch periods around each t_i , the aHM model [16] is first used to decompose the analytic signal of $s(t)$ into harmonic frequency components:

$$s_i(t) = \sum_{h=1}^{H_i} a_{i,h} \cdot e^{j(h\phi_0(t) + \phi_{i,h})} \quad (1)$$

where i is the frame index, $H_i = \lfloor 0.5f_s/f_0(t_i) \rfloor$ is the number of harmonic up to Nyquist frequency in the frame i , f_s denotes the sampling frequency, $a_{i,h}$ is the real-valued amplitude of the h th harmonic at frame i , $\phi_{i,h}$ is the instantaneous phase parameter and $\phi_0(t)$ is a real function which adapts the frequency basis of the harmonic model to the waveform frequency modulations [15]:

$$\phi_0(t) = \frac{2\pi}{f_s} \int_{t_i}^t f_0(\tau) d\tau \quad (2)$$

For this work, the Adaptive Iterative Refinement (AIR) algorithm presented in [16] is used to refine

the $f_0(t)$ values and the sinusoidal parameters ($a_{i,h}$ and $\phi_{i,h}$) are estimated using the Least Square (LS) solution.

Conversely to the conventional harmonic model [3], the aHM model uses a full-band non-stationary frequency basis. This mainly allows to represent a whole speech recording using a single and continuous harmonic structure during both analysis and synthesis steps [15]. This structural property is very convenient for the phase models and processing used in this work. Also, in unvoiced segments, assuming that an $f_0(t)$ curve can be obtained without substantial erratic jumps, it has been shown that aHM can represent both voiced and unvoiced segments uniformly, without voicing decision [16]. Given the goal of this work, this property is obviously a necessary prerequisite. Additionally, together with its harmonic tracking algorithm (i.e. AIR), this model provides almost always the most accurate and precise sinusoidal parameters compared to state-of-the-art methods [16]. Eventually, this good accuracy and precision might not be critical for obtaining the results of this article. However, this allows to minimize the influence of the sinusoidal parameter estimation on the results and, thus, strengthening the link between the suggested phase processing techniques and the results obtained. Finally, like the conventional harmonic model [3], the resynthesis obtained by aHM is almost indistinguishable from the original recording [16]. This ensures that the aforementioned properties come with no perceptual degradation.

Despite the advantages of aHM, the sinusoidal parameters $a_{i,h}$, $\phi_{i,h}$ and $f_{i,h} = hf_0(t_i)$ lie on the harmonic grid, which is not convenient for manipulation of the perceived characteristics or for statistical modeling. Moreover, the instantaneous phase parameter $\phi_{i,h}$ constantly wraps from one instant to the next, which makes its modeling far from straightforward. In the following steps, we aim at building amplitude and phase features, which are independent of the harmonic structure and we focus on the modeling of the instantaneous phase.

2.1 A simple representation of the amplitude

For this study, we assumed that the voice production consists of a spectrally flat source and a filter [36]. In frequency domain:

$$S_i(f) = G_i(f) \cdot V_i(f) \quad (3)$$

where $S_i(f)$ is the Fourier transform of $s_i(t)$, described in hertz for reason of clarity, $G_i(f)$ is the spectrum of the voice source and $V_i(f)$ is the vocal tract filter response. Therefore, the harmonic amplitudes $a_{i,h}$ can

be considered as discrete samples of the filter amplitude response $|V_i(f)|$:

$$\text{if } |G_i(f)| = 1 \quad \forall f \quad \Rightarrow \quad a_{i,h} = |V_i(hf_0(t_i))| \quad (4)$$

We also assumed that $V_i(f)$ is minimum-phase so that $\angle V_i(f)$ is linked to $|V_i(f)|$ through the Hilbert transform [37]. The phase response $\angle V_i(f)$ can also be retrieved through the imaginary part of the Fourier transform of the minimum-phase cepstrum $\hat{v}_-(t)$ of $V_i(f)$:

$$\hat{v}_-(t) = \begin{cases} 0 & t < 0 \\ 2\hat{v}(t) & t > 0 \\ \hat{v}(t) & t = 0 \end{cases} \quad (5)$$

where $\hat{v}(t)$ is the real cepstrum of $|V_i(f)|$, i.e. $\hat{v}(t) = \mathcal{F}^{-1}(\log |V_i(f)|)$, as described in [37]. Modeling the amplitude envelope is a well investigated subject and it is out of the scope of this article. In order to estimate $|V_i(f)|$ in a robust and simple way, we used a linear interpolation of $a_{i,h}$ across frequency, as used in [38], on a discrete scale of 512 frequency bins up to the Nyquist frequency. However, for reason of clarity, the continuous notation in hertz will be used in the following.

Even though the assumption of spectrally flat source is widely used, it is also known that this hypothesis is basically wrong since the glottal pulses have a low-pass characteristic [39]. Therefore, in this work, $|V_i(f)|$ encompasses the amplitude spectra of both the glottal source and the vocal tract. Nevertheless, using PD, it has been shown that this assumption allows to extract glottal source information which is almost independent of the vocal tract filter [40]. Indeed, this property was critical for estimation of glottal model parameters using phase minimization [40, 25]. For the work presented in this article, this same property ensures that the impact of the vocal tract filter on the phase features representing the source is minimal. On the contrary, the impact of the glottal source on the vocal tract feature is far from negligible, which is not convenient. However, a robust separation of the vocal tract filter and the glottal source is far from straightforward [41, 42, 43, 44, 31]. Thus, in this work, we chose to favor again robustness and simplicity, in order to focus, beforehand, on the phase features. Interposing a separation process within the presented phase feature extraction can be part of future works.

3 Representations of the phase

In this section, we first describe the analytical model of the instantaneous phase used in this work. State-of-the-art phase processing are then described and discussed analytically using this model. Finally, the novel

characterization of the short-term statistics of PD is described.

3.1 Theoretical model of the instantaneous phase

In order to represent the instantaneous phase parameter $\phi_{i,h}$, models have been already suggested, for phase synchronisation between frames [45, 46] and speech coding [14, 47]. In this work, we suggest to represent the measured $\phi_{i,h}$ using a model similar to that in [47]:

$$\phi_{i,h} = \underbrace{\theta_{i,h}}_{\text{source shape}} + \underbrace{h \frac{2\pi}{f_s} \int_{c_i}^{t_i} f_0(\tau) d\tau}_{\text{linear phase}} + \underbrace{\angle V_i(h f_0(t_i))}_{\text{filter}} \quad (6)$$

whose terms are described here below. In voiced segments, each glottal pulse of the glottal source has a shape which has mainly maximum-phase characteristics [39, 27]. This glottal pulse shape has also a position in time c_i . Speech processing techniques often define c_i as glottal closure instants [48, 49], or as energy local maxima of a residual signal [50, 51], or as pitch pulse onsets [12, 14, 27] for centering windows and to synchronize instantaneous phase parameters. Even though such a definition is necessary for many approaches, we will show below that it is not necessary when using the Relative Phase Shift (RPS)[24, 33] or PD, which avoids an extra estimation procedure and its potential misestimation errors. In unvoiced segments, one can assume that this shape is basically random for each frame. In (6), the *source shape* term $\theta_{i,h}$ represents this pulse shape in both voiced and unvoiced segments. Since the analysis windows are not centered on each c_i (i.e. $c_i \neq t_i$), a *linear phase* term is also necessary in order to represent the time delay between t_i (the window's center) and the position of the source shape c_i . In the literature, assuming the frequency structure is stationarity within a frame, i.e. $f_0(t) = f_0(t_i)$, this term is often simplified to a term which is linear in both frequency and time, i.e. $h(2\pi f_0(t_i)/f_s)(t_i - c_i)$. Conversely, in (6), we use the integral form since the harmonic structure is not stationary in the aHM model. Finally, according to the voice production (3), the *voice source* is convolved by the vocal tract filter impulse response. Thus, we add the minimum-phase $\angle V_i(\omega)$ to the model.

The following sections describe the suggested way to characterize $\phi_{i,h}$ for speech processing using statistics of PD and using the RPS as an intermediate step.

3.2 From phases to Relative Phase Shift (RPS)

The linear phase component in (6) constantly wraps the instantaneous phase $\phi_{i,h}$ from one time instant to

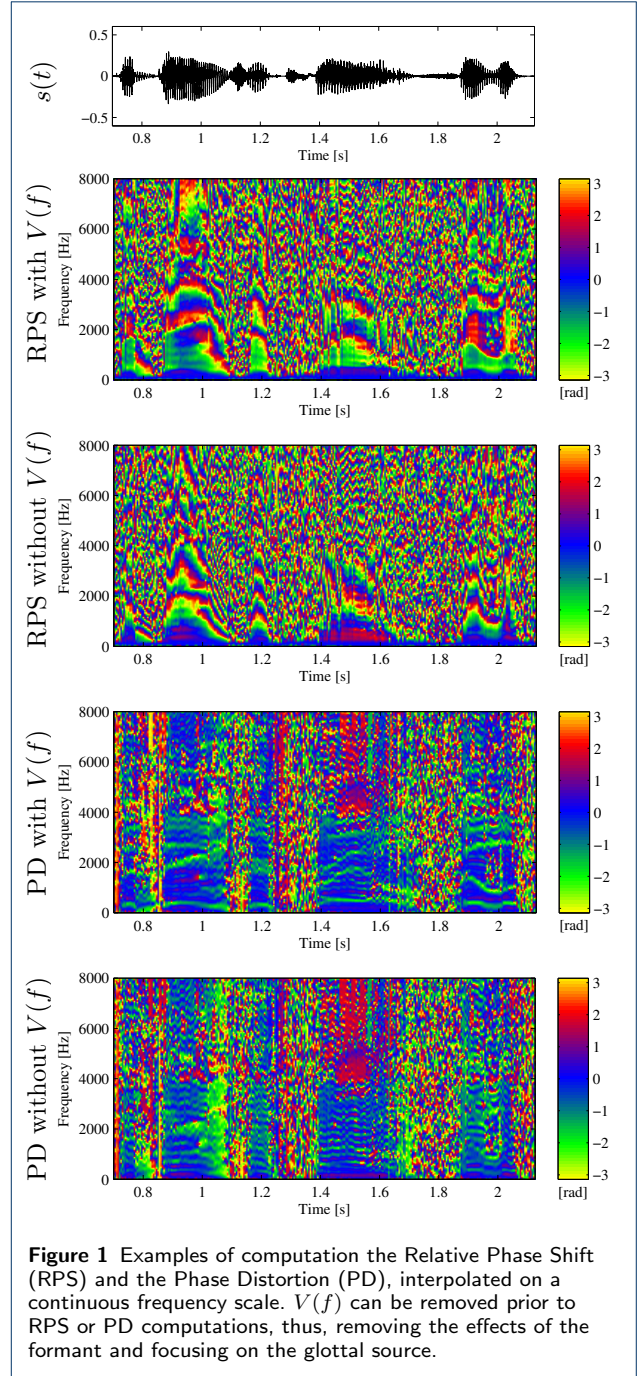


Figure 1 Examples of computation of the Relative Phase Shift (RPS) and the Phase Distortion (PD), interpolated on a continuous frequency scale. $V(f)$ can be removed prior to RPS or PD computations, thus, removing the effects of the formant and focusing on the glottal source.

the next. This constitutes a major issue in phase modeling [27]. To alleviate this issue, the Relative Phase Shift (RPS) has been suggested [33], which is expressed as:

$$\text{RPS}_{i,h} = \phi_{i,h} - h\phi_{i,1} \quad (7)$$

The second row of Fig. 1 shows an example of RPS computation (the harmonic values have been interpo-

lated on a continuous frequency axis for reason of presentation). To further analyze the results of the RPS computation, one can replace the estimated instantaneous phase parameter $\phi_{i,h}$ by its models (6):

$$\begin{aligned} \text{RPS}_{i,h} &= \theta_{i,h} + h \frac{2\pi}{f_s} \int_{c_i}^{t_i} f_0(\tau) d\tau + \angle V_i(hf_0(t_i)) \\ &\quad - h \cdot \left(\theta_{i,1} + \frac{2\pi}{f_s} \int_{c_i}^{t_i} f_0(\tau) d\tau + \angle V_i(f_0(t_i)) \right) \end{aligned} \quad (8)$$

which reduces to:

$$\text{RPS}_{i,h} = \theta_{i,h} - h\theta_{i,1} + (\angle V_i(hf_0(t_i)) - h\angle V_i(f_0(t_i))) \quad (9)$$

Eq. (9) shows that the RPS computation discards the linear phase terms. It remains only the source shape at each harmonic relative to that of the 1st harmonic and the contribution of the minimum-phase envelope $\angle V_i(f)$ relative to that at the 1st harmonic. In voiced segments, these two remaining terms can be easily assumed to evolve smoothly across time because the shape of the glottal pulse and the vocal tract do so. Therefore, this property of RPS basically solves the issue of phase wrapping. Additionally, c_i is also discarded in (9) so that there is no need to estimate any GCI or pitch pulse onset. This avoids misestimation of such time instants and the consequences on speech processing techniques.

The RPS can also be computed on the Linear Prediction (LP) residual [33], which removes the minimum-phase contribution of $V_i(f)$. Similarly, let's define:

$$\tilde{\phi}_{i,h} = \phi_{i,h} - \angle V_i(hf_0(t_i)) \quad (10)$$

where $\tilde{\phi}_{i,h}$ is the instantaneous phase where the minimum-phase frequency response corresponding to the amplitude envelope has been removed. Consequently, (9) becomes:

$$\widetilde{\text{RPS}}_{i,h} = \tilde{\phi}_{i,h} - h\tilde{\phi}_{i,1} = \theta_{i,h} - h\theta_{i,1} \quad (11)$$

The third row of Fig. 1 shows an example of $\widetilde{\text{RPS}}$ (with, again, the interpolation on a continuous frequency axis). In (11), only the source shape and the harmonic number h remains. Ideally, we want to extract features from the speech waveform which are independent from each other as much as possible. However, the harmonic number h belongs to the harmonic structure which is already handled by $f_0(t)$ and the property of harmonicity of the harmonic model. Therefore, this harmonic number is still inconvenient for characterizing the phase properties independently from $f_0(t)$. Interpolating the $\widetilde{\text{RPS}}_{i,h}$ values on a continuous frequency

axis (as depicted in Fig. 1) removes the harmonic sampling. However, the harmonic number is still present in the interpolated values. Note also that h increases the RPS variance towards high frequencies and drowns the variance of $\theta_{i,h}$ into that of $h\theta_{i,1}$, which is not convenient for characterizing the source shape in mid and high frequencies.

3.3 From Relative Phase Shift to Phase Distortion (PD)

The problem cited above can be solved by simply computing the finite difference of $\widetilde{\text{RPS}}_{i,h}$ with respect to h . In the general context of signal processing, the relative phase difference between two frequency components is known as Phase Distortion (PD), whose perceived characteristics are already studied and known [52, 53, 54]. In the particular context of speech analysis, we already used the PD for the estimation of glottal model parameters [25, 40] and for emotion valence detection [55]. Moreover, in [40], we have shown that the PD is directly linked to the maximum-phase component of the glottal source. This sole property allows to estimate parameters of glottal models as presented in [40, 25]. Therefore, PD is also a strong correlate of the maximum-phase component of the voice source. The rather complicate definition of PD in [25, 40] is actually equal to:

$$\begin{aligned} \text{PD}_{i,h} &= \Delta_h \widetilde{\text{RPS}}_{i,h} = \widetilde{\text{RPS}}_{i,h+1} - \widetilde{\text{RPS}}_{i,h} \\ &= \left(\tilde{\phi}_{i,h+1} - (h+1)\tilde{\phi}_{i,1} \right) - \left(\tilde{\phi}_{i,h} - h\tilde{\phi}_{i,1} \right) \\ &= \tilde{\phi}_{i,h+1} - \tilde{\phi}_{i,h} - \tilde{\phi}_{i,1} \end{aligned} \quad (12)$$

where Δ_h denotes the finite difference operator. The fifth row of Fig. 1 shows an example of PD (with the interpolation on a continuous frequency axis). Basically, the PD measures the phase desynchronization which exists between each sinusoidal components of the voice source. Additionally, this desynchronization is centered on the 1st harmonic phase, like for the RPS. The finite difference makes also the PD similar to the group-delay, whose perceived characteristics have been already studied and demonstrated [56] and whose applications are numerous [57, 58, 48, 59, 60, 61].

By replacing $\phi_{i,h}$ by its model (6), the PD computation leads to:

$$\text{PD}_{i,h} = \theta_{i,h+1} - \theta_{i,h} - \theta_{i,1} \quad (13)$$

Since the linear and filter terms cancel, only the source shape terms remain in (13). Eq. (13) also shows that the computation of the PD represents the phase desynchronization of the source shape between each harmonic, centered on that of the 1st harmonic. Compared to (11), the harmonic number h is also removed,

as expected, by using the finite frequency difference. Consequently, when h increases, it adds to the RPS measurement, but does not influence the PD measurement. For example, when using PD in fourth and fifth rows of Fig. 1, one can see red patterns appearing around 1.5s between 4kHz and 8kHz. On the contrary, no clear pattern appears in the same time frequency region using the RPS (second and third rows). Using RPS, the region concerned actually seems as blurred as in noisy time frequency regions (e.g. around 1.8s). This is explained by the presence of the harmonic number h in RPS which increases the wrapping of the phase values.

3.4 Statistical features of the Phase Distortion

As shown in (13), the phase distortion represents basically the source shape. In voiced segments, the source shape accounts mainly for the shape of the glottal pulse. In unvoiced segments, the time evolution of this shape throughout adjacent frames reproduces the noisiness of the voice source. Therefore, in this section, we suggest to statistically characterize the phase distortion in a short-term window in order to extract a feature related to the shape at a given time and another feature representing the local variation of this shape around that same time. This characterization will allow to manipulate the components of the speech in voice transformation and HMM-based synthesis.

We first assume that the information carried in PD is independent of the fundamental frequency. As a consequence, we interpolate $PD_{i,h}$ on a linear frequency scale (as done for the previous Figures), like a phase spectral envelope [62, 63], thus, removing the harmonic number from the representation. To achieve this phase envelope, we first unwrap $PD_{i,h}$ and then interpolate it linearly on a discrete scale of 512 frequency bins up to the Nyquist frequency, thus, $PD_{i,h}$ becomes $PD_i[k]$. Here, the unwrapping function is necessary to avoid meaningless values during the interpolation process. Nevertheless, the resulting $PD_i[k]$ is still a circular data. Instead of the discrete notation in bins, the continuous notation in hertz will be used in the following descriptions and sections for reason of clarity, i.e. $PD_i[k] \Leftrightarrow PD_i(f)$, like for the amplitude spectral envelope $V_i(f)$.

On a frame-by-frame basis in a analysis/synthesis procedure, the sole information carried by $PD_i(f)$ might be sufficient to reconstruct an instantaneous phase which has the same perceived characteristics than those of the instantaneous phase $\phi_{i,h}$. This property has actually been shown through listening tests in [33]. However, through manipulation of $PD_i(f)$, by time scaling, pitch scaling or statistical modeling, the short-term statistical characteristics of the analyzed

voice might not be preserved. For example, stretching $PD_i(f)$ over time would automatically reduce its temporal variance, thus, changing the extent of randomness in the voice, which is not the purpose of a time stretching transformation. In this article, we suggest to preserve the short-term mean and short-term standard-deviation of $PD_i(f)$ in speech processing applications using features that represent these two moments. In order to estimate the mean and standard-deviation, we assume that the distribution of $PD_i(f)$ obeys a normal distribution. Moreover, since $PD_i(f)$ is a circular data defined in $(-\pi, \pi]$, we make use the wrapped normal distribution [64, 47].

In order to ensure that the short-term estimate of PD's variance is not influenced by the number of periods, it is first necessary to use the same number of time instants t_i in each glottal cycle and not a constant step size as previously assumed up to here. In order to have enough values for computing a reliable short-term variance in the following, we used 4 analysis instants per period:

$$t_i = t_{i-1} + \frac{1}{4 f_0(t_{i-1})} \quad \text{with } t_0 = 0 \quad (14)$$

3.4.1 The short-term mean of PD

Since the values $PD_i(f)$ are circular data, the wrapped normal distribution [64, 47] has been used in this work to model $PD_i(f)$ over a few periods. The mean is estimated with [64]:

$$\mu_i(f) = \text{mean}_i(PD_i(f)) = \angle \left(\frac{1}{N} \sum_{n \in B} e^{jPD_n(f)} \right) \quad (15)$$

where $B = \{i - \frac{N-1}{2}, \dots, i + \frac{N-1}{2}\}$ and we used $N = 25$ frames in this work, which corresponds to 6 periods. This averaging of $PD_i(f)$ is necessary for separating the randomness characteristics of the phase from its smoothly varying behaviors. Even though 6 periods might appear to be substantial, it ensures that the mean does not model the randomness of the phase, which has to be modeled by the feature described below.

3.4.2 The short-term standard-deviation of PD

According to [64], the standard-deviation should be estimated by :

$$\begin{aligned} \sigma_i(f) &= \text{std}_i(PD_i(f)) \\ &= \sqrt{-2 \log \left| \frac{1}{N} \sum_{n \in B} e^{j(PD_n(f))} \right|} \end{aligned} \quad (16)$$

where $B = \{i - \frac{N-1}{2}, \dots, i + \frac{N-1}{2}\}$. However, in our context of application, two issues arise with (16).

Firstly, as shown by (13), $PD_i(f)$ is related to the source shape in voiced segments. One can easily assume that this shape is non-constant and evolve smoothly across time. Thus, over a few periods, $PD_i(f)$ has also a non-constant trend. Fig. 2 shows an example of $PD_i(2 \cdot f_0(t))$ and the corresponding estimation of $\sigma_i(2f_0(t))$. One can see that the waveform is drastically changing from 1s to ~ 1.07 s. This phenomenon is revealed by $PD_i(2f_0(t))$ which tends towards low phase values. The $\sigma_i(f)$ estimate is overestimated by

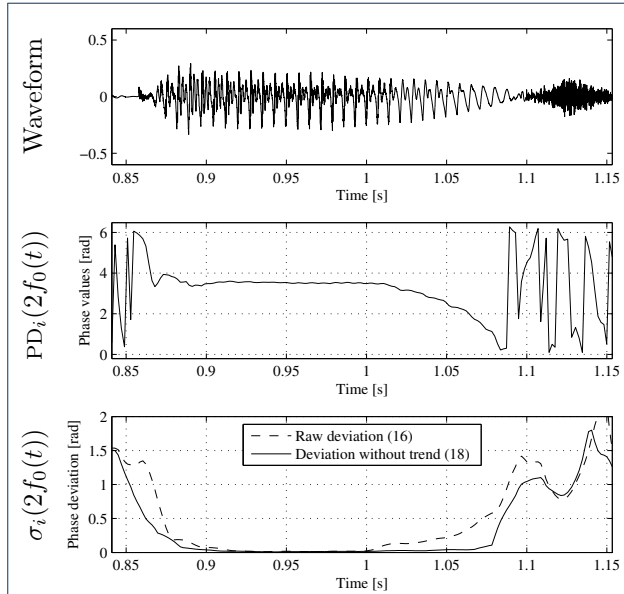


Figure 2 Example of estimation of the standard-deviation of PD, using (16) and (18).

the presence of this trend whereas $\sigma_i(f)$ is supposed to represent only the noisiness of the voice source. Moreover, the time evolution of $\mu_i(f)$ already model this trend of the voice source. Consequently, there is no reason to keep this trend in the standard-deviation estimate. In Fig. 2, one can also see that the estimation of the variance using (16) reaches 0.5rad in the bottom plot, even though the waveform does not show any noise around 1.07s. Therefore, to alleviate this problem, we suggest to remove an estimate of the trend prior to the computation of the standard-deviation. The trend is first estimated by averaging $PD_i(f)$ over 2 periods:

$$\widehat{PD}_i(f) = \mathcal{L} \left(\frac{1}{M} \sum_{m \in C} e^{jPD_m(f)} \right) \quad (17)$$

where $C = \{i - \frac{M-1}{2}, \dots, i + \frac{M-1}{2}\}$, with $M = 9$. This trend is then removed before computing the standard-

deviation:

$$\begin{aligned} \sigma_i(f) &= \text{std}_i(PD_i(f) - \widehat{PD}_i(f)) \\ &= \sqrt{-2 \log \left| \frac{1}{M} \sum_{m \in C} e^{j(PD_m(f) - \widehat{PD}_m(f))} \right|} \end{aligned} \quad (18)$$

where $C = \{i - \frac{M-1}{2}, \dots, i + \frac{M-1}{2}\}$ and $M = 9$ frames. Using 2 periods for the standard-deviation and 6 for the mean is motivated by the following reason. A wider window for the standard-deviation could cover the end of a noisy segment and the beginning of a voiced segment, thus, overestimating the presence of noise at the beginning of the voiced segment. Therefore, a short window seems necessary to quickly adapt the standard-deviation estimate in transients. On the other hand, using a wider window for the mean allows to obtain a more robust estimate of the source shape in transients where harmonic sinusoidal parameters are less reliable than in voiced segments.

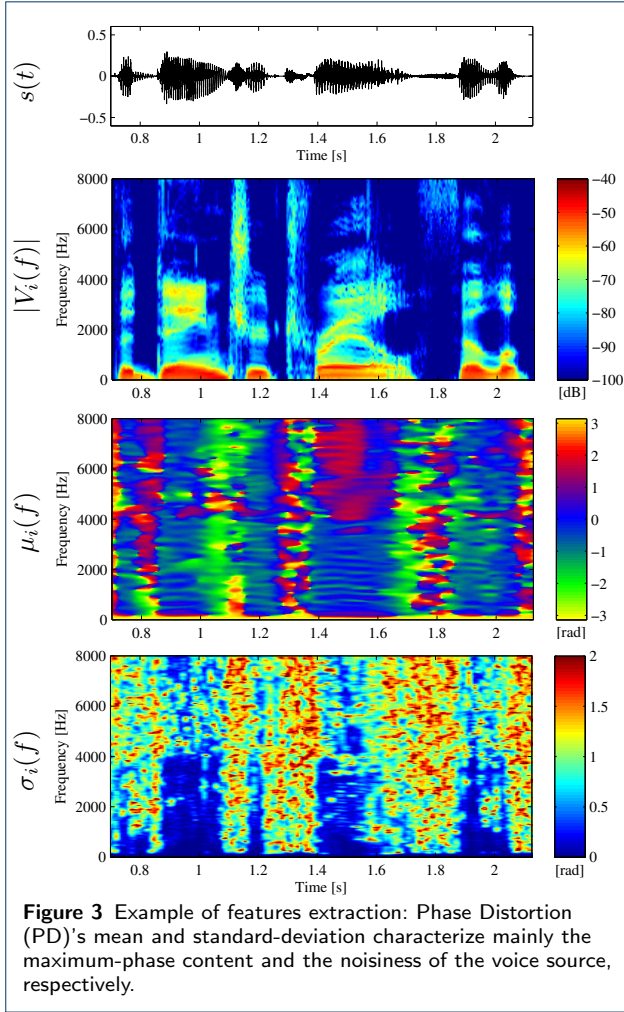
In order to have the same number of analysis instants in each period, the step size of the analysis instants was first adapted to $f_0(t)$ (see eq. 14). However, both mean and standard-deviation have to be independent from $f_0(t)$ so that each feature represents independent characteristics of the speech signal. Additionally, a variable step size is not desirable for many applications, like in statistical modeling, where a constant step size is necessary. Consequently, prior to any application, $\mu_i(f)$ and $\sigma_i(f)$ features are resampled at new time instants \hat{t}_i , with a constant step size, each 5ms. Fig. 3 shows an example of features extraction.

4 Synthesis

The analysis steps described above provide, each 5ms, the features $f_0(t_i)$, $V_i(f)$, $\mu_i(f)$ and $\sigma_i(f)$. This section describes the method used to resynthesize a full speech signal using these features. This synthesis method is similar to that used originally for the aHM model [16]. Basically, each harmonic track is first synthesized across time, independently of each other, using a sampling rate f_s . Since the synthetic signal is band-limited to the Nyquist frequency, the continuous notation for the time axis will be used in the following, for reason of simplicity (i.e. $x[n] \Leftrightarrow x(t)$). The synthetic harmonics are then added up all together, without using any windowing scheme.

In order to synthesize the continuous amplitude $\hat{a}_h(t)$ of each harmonic track, the amplitude envelope is first sampled at harmonic frequencies for each time instants \hat{t}_i :

$$\hat{a}_{i,h} = |V_i(hf_0(\hat{t}_i))| \quad (19)$$



Then, these anchor values $\hat{a}_{i,h}$ are interpolated across time on a logarithmic scale in order to obtain $\hat{a}_h(t)$.

In the aHM model, the continuous instantaneous phase $\hat{\phi}_h(t)$ was not directly interpolated from the measured instantaneous phase parameters $\phi_{i,h}$ because of the linear phase term which is present in $\phi_{i,h}$ and does not allow to compute a reliable interpolation. Instead, a *relative phase* $\check{\phi}_{i,h}$ was first computed in order to remove the influence of the linear phase [16]:

$$\check{\phi}_{i,h} = \phi_{i,h} - h \cdot \frac{2\pi}{f_s} \int_0^{t_i} f_0(\tau) d\tau \quad (20)$$

where the zero for the beginning of the integral means the beginning of the speech signal. Since the linear phase is removed in (20), the relative phase changes smoothly from one time instant to the next if the shape of the signal is also changing smoothly. Thus, any processing of $\check{\phi}_{i,h}$ is better conditioned than a processing of the raw instantaneous phase value $\phi_{i,h}$. This property explains the success of the simple processing tech-

niques presented in [17, 18]. In [16], the relative phases were then interpolated on a continuous time axis using splines, i.e. $\check{\phi}_{i,h} \Rightarrow \check{\phi}_h(t)$. Finally, the continuous instantaneous phase $\hat{\phi}_h(t)$ was recovered by adding back the linear phase previously removed:

$$\hat{\phi}_h(t) = \check{\phi}_h(t) + h \cdot \frac{2\pi}{f_s} \int_0^t f_0(\tau) d\tau \quad (21)$$

The final synthetic signal $\hat{s}(t)$ was generated by summing all the harmonic tracks together:

$$\hat{s}(t) = \sum_{h=1}^H \hat{a}_h(t) e^{j\hat{\phi}_h(t)} \cdot \chi_{[hf_0(t) < f_s/2]}(t) \quad (22)$$

where H is the maximum value of all H_i and the indicator function $\chi_{[hf_0(t) < f_s/2]}(t)$ discards any harmonic segment whose frequency is higher than Nyquist.

For the analysis/synthesis method suggested in this article, the aHM synthesis summarized above has to be adapted in order to use $\mu_i(f)$ and $\sigma_i(f)$. Looking at equation (21), the linear phase term can be reconstructed because $f_0(t)$ is preserved through the analysis step. However, the relative phase $\check{\phi}_h(t)$ is lost. Therefore, the main goal is to reconstruct a synthetic relative phase which has the same perceived characteristics as the original one. For this purpose, we suggest to follow the phase model (6) while using the short-term mean and standard-deviation of the phase distortion in order to resynthesis a source shape. First, at each instant \hat{t}_i , we synthesize a phase distortion $\hat{\text{PD}}_{i,h}$ using the wrapped normal distribution [64, 47]:

$$\hat{\text{PD}}_{i,h} = \mathcal{WN}(\mu_i(hf_0(\hat{t}_i)), \sigma_i(hf_0(\hat{t}_i))) \quad (23)$$

where $\mathcal{WN}(\mu, \sigma)$ generates random values which obey a wrapped normal distribution of mean μ and standard-deviation σ . Note that this procedure is similar to known phase randomization techniques [65, 66, 67]. However, because of the finite difference used to compute PD, our approach is similar to randomizing the group-delay and not the instantaneous phase. Because of this difference, the randomization is always centered around the linear phase. Therefore, this approach ensures that the noise component is always merged with the deterministic component, which avoid these two components to be perceived separately. Based on equation (13), then, we suggest to approximate a source shape using $\hat{\text{PD}}_{i,h}$. However, because the PD values are centered around $\theta_{i,1}$ (see eq. (13)), this value is lost during the analysis step. Therefore, we assume $\theta_{i,1} = 0$ and thus:

$$\hat{\theta}_{i,h} = \Delta_h^{-1} \hat{\text{PD}}_{i,h} \quad (24)$$

where Δ_h^{-1} is the cumulative sum, which compensates for the finite difference in (13). Following the model of the instantaneous phase (6) we finally add the minimum-phase response of the envelope $V_i(f)$ in order to obtain the synthetic relative phase values $\phi_{i,h}$:

$$\check{\phi}_{i,h} = \hat{\theta}_{i,h} + \angle V_i(hf_0(t_i)) \quad (25)$$

The rest of the synthesis process is identical to that of aHM. The continuous relative phase values are interpolated across time using splines and the linear phase is added at the end in order to obtain the continuous instantaneous phase $\hat{\phi}_h(t)$ (21) which is finally used in equation (22).

The complete analysis/synthesis procedure is called Harmonic Model + Phase Distortion (HMPD).

4.1 Correction of $\sigma_i(f)$

When testing the resynthesis capabilities of the method through informal listening tests, we found that the perceived characteristics of the fricatives were not properly reproduced. Basically, no segments of the signal were fully randomized. After investigation, we found that $\sigma_i(f)$ was limited in its measure. To illustrate this phenomenon, Fig. 4 shows the average of $\sigma_i(f)$ measured from synthetic signals of known $\sigma_i(f)$. Synthetic signals of 4s are first generated using the HMPD synthesis method described above with $\mu_i(f) = 0 \forall i, f$. Each synthetic signal uses a different $\sigma_i(f)$ value between 0 and 2. The HMPD analysis method is then used to re-estimate the $\sigma_i(f)$ values from the synthetic signals. Fig. 4 shows the measured $\sigma_i(f)$ averaged along the 4s. It shows that the measured $\sigma_i(f)$ hardly reach 1.2, which is not sufficient for reconstructing a sound which is perceived as fully noisy. The following two reasons can explain this issue. Firstly, during analysis, a window of 3 periods is used to estimate the sinusoidal parameters (1). This window is necessary to obtain a frequency resolution which allows to distinguish harmonic peaks and estimate the sinusoidal parameters with a sufficient accuracy. However, this window also over-smooths the variance of the parameter estimates across time. Even though these parameters allow reconstruction of the signal on a frame-by-frame basis. This over-smoothing effect do not allow to estimate a short-term variance with sufficient accuracy. Secondly, the window size M in eq. (18), which has to be short enough to follow the time evolution of the speech signal, limits also the standard-deviation estimate. Note that this effect appears with or without removing the PD's trend $\widehat{\text{PD}}_i(f)$ during analysis, thus, using either (16) or (18).

To avoid overloading this presentation, we chose to alleviate this issue in a simple manner. We corrected

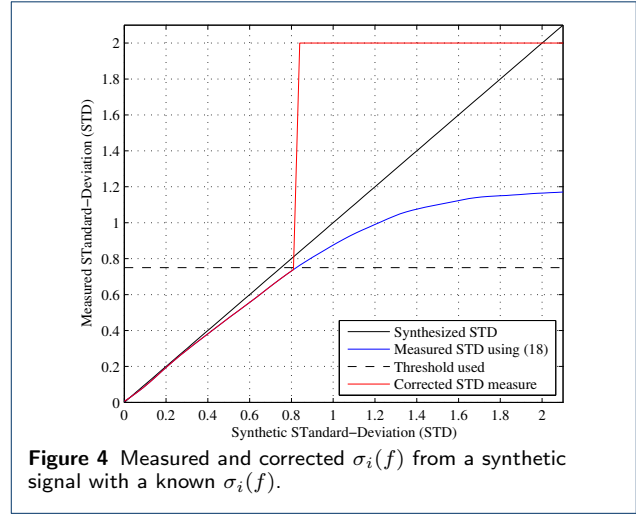


Figure 4 Measured and corrected $\sigma_i(f)$ from a synthetic signal with a known $\sigma_i(f)$.

the $\sigma_i(f)$ values prior to synthesis so that it was given a sufficient randomness when it was greater than an empirical threshold. Through informal listening tests, we found that a forced $\sigma_i(f)$ value of 2 used above a threshold of 0.75 properly reconstruct the noisiness of fricatives while preserving the voiced segments quality. Fig. 4 depicts this corrections. Fig. 5 shows an example of $\sigma_i(f)$ after correction, as it is used during the synthesis step. Future works are planned to study the influence of the window sizes and address this issue in a neater way.

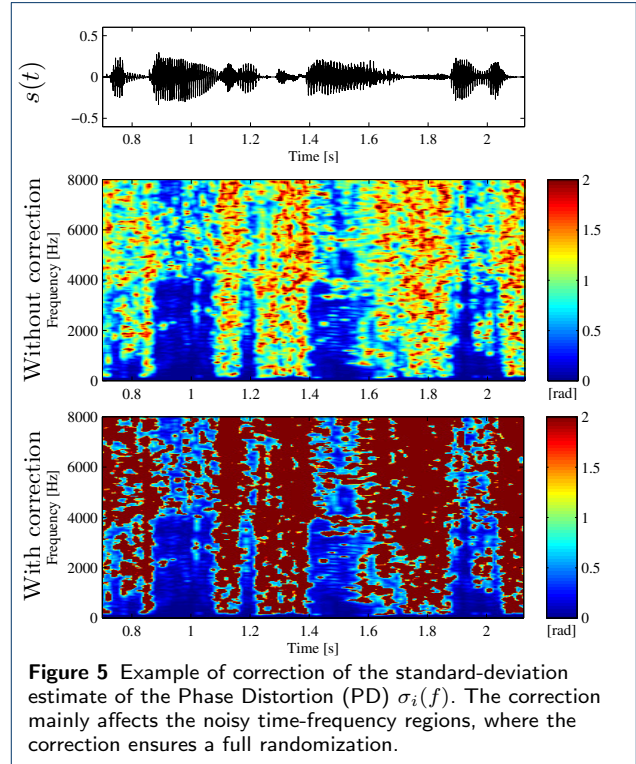


Figure 5 Example of correction of the standard-deviation estimate of the Phase Distortion (PD) $\sigma_i(f)$. The correction mainly affects the noisy time-frequency regions, where the correction ensures a full randomization.

5 Evaluation

This section aims at assessing the quality and versatility of the proposed phase representation. To this end, experiments have been conducted in three different scenarios: resynthesis with no modification (Sec. 5.1), pitch scale modification (Sec. 5.2), and HMM-based speech synthesis (Sec. 5.3).

Even in resynthesis, objective measures such as Signal to Reconstruction Error Ratio (SRER) or PESQ [68] are not suitable for evaluation as they are waveform-sensitive. While it is true that the suggested HMPD representation retains the waveform characteristics of the signal, it does not keep its linear phase term: the original linear phase removed between (8) and (9) and the synthetic one added in (21) are not necessarily the same but just have the same derivative, i.e. $f_0(t)$. Consequently, original and synthesized waveforms are not time-synchronous. Objective measures are also inconvenient for comparing different configurations of the HMPD vocoder, including those dropping the maximum-phase component given by $\mu_i(f)$, in which the shape of the glottal pulse is not preserved. Therefore, in this paper, all evaluations have been carried out by means of subjective listening tests, as suggested by [69]. All the sounds used in the following tests are available at [70].

5.1 Quality of resynthesis

The first test was designed to evaluate mainly the importance of $\mu_i(f)$ and $\sigma_i(f)$ in terms of perceptual quality. The speech database used in this experiment contained a total of 32 utterances spoken in 16 different languages (2 utterances per language, one from a male speaker and one from a female speaker). Such a multilingual database had been thoroughly designed to exhibit a very wide phonetic variability and also an heterogeneous set of speakers. The sampling frequency of the signals, f_s , varied between 16 kHz and 44.1 kHz. All original recordings showed high signal-to-noise ratios as they had been collected from various synthesis databases. The test was conducted through a web-based interface. A total of 43 volunteer listeners were presented with the original recordings of randomly-selected signals along with their reconstructed versions using: aHM; the suggested HMPD using both $\mu_i(f)$ and $\sigma_i(f)$; HMPD using $\sigma_i(f)$ only; the well-known STRAIGHT vocoder, which was used as a hidden anchor. Then, they were asked to grade the quality of these sounds using a 5-points scale [69]. The order of the reconstruction methods was randomized too and the listenings were made through headphones or earphones. For consistency and to avoid the fatigue of the evaluators, each listener was asked to grade only the voices of two languages (both male and female voices) randomly selected among the 16.

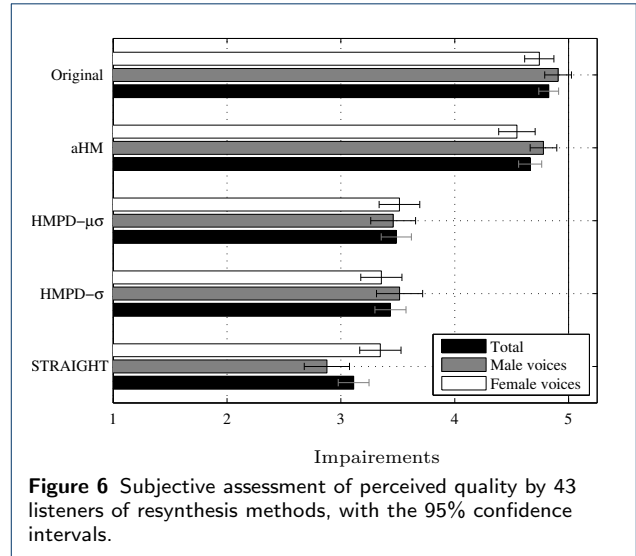


Fig. 6 contains the resulting mean opinion scores along with their 95% confidence intervals. The scores have been normalized according to the number of occurrence of each language and to the variance of each listener's answers, as suggested in [69]. The scores achieved by aHM are consistent with those reported in previous studies [16] and confirm the very good resynthesis potential of this signal model. However, one of the most remarkable observations is the significant performance gap between aHM and HMPD, which gives an idea of the relevance of phase in quality issues. This loss of quality could be partly explained by the following reason. First, the HMPD resynthesis is not time synchronous with that of aHM. Indeed, the synthetic linear phase term used in (21) is not that of the original speech signal. On the one hand, the voiced segments can be delayed or advanced of only a maximum of half of a period. This effect is surely impossible to perceive within a voiced segment, where the amplitude envelope and the phase characteristics are properly decorrelated using the suggested analysis/synthesis procedure. However, on the other hand, this time desynchronization might play a role in highly non-stationary speech segments like in transients from plosive to voiced segments or in creaky voice. Indeed, in such cases, the time amplitude envelope, which is driven by the mean of the spectral amplitude envelope, should be synchronous with the impulses triggered by the linear phase. However, the time desynchronization in HMPD may break this necessary correlation which can easily blur the perception of time events and degrade the overall quality. Note that this reasoning also holds for the STRAIGHT method where the linear phase is also fully artificial and neither synchronous with the original one. According to informal listen-

ing, the creaky voice segments seem, indeed, not properly reconstructed in both HMPD and STRAIGHT. Among other reasons, the measure of randomness using $\sigma_i(f)$ might not adapt quickly enough in transients, so that the beginning and the end of voiced segments can be sometimes over-randomized. Smoothing techniques and different separation procedures for estimation of $\mu_i(f)$ and $\sigma_i(f)$ should be investigated in the future.

Regarding the relative performance of HMPD- $\mu\sigma$ and HMPD- σ , the average scores indicate that, for the voices used in this experiment, the listeners were not able to perceive any difference between them. This suggests that the contribution of $\mu_i(f)$ is not perceptually significant in comparison with that of $\sigma_i(f)$. Even more, since the link between PDi, h and the maximum-phase of the glottal pulse has been shown and exploited [25, 40], this suggests that the maximum-phase information is hardly noticeable at this overall quality level. Admittedly, this could also be an indicator that $\mu_i(f)$ is not capturing the maximum-phase component properly. In any case, the average results also show that the quality provided by HMPD is at least as good as that of STRAIGHT and better for male voices. Note also that, compared to STRAIGHT, the difference of quality between genders is also clearly reduced using HMPD. In other words, the phase randomization technique suggested in this paper, which exploits $\sigma_i(f)$, might be a potential improvement and replacement for STRAIGHT's aperiodicity measures [28].

5.2 Quality of pitch shifting

A second experiment was conducted to check the consistency of HMPD in a more challenging scenario. In that sense, pitch scaling is preferable over time scaling because it can shed light on possible inaccuracies in isolating amplitude or phase information from periodicity information. Therefore, after the analysis step, $f_0(t_i)$ was multiplied by a factor of 2, or 0.5, in order to shift the pitch of the voice one octave upwards, or downwards, respectively. The signals in the database described in Sec. 5.1 were manipulated using three different methods: HMPD- $\mu\sigma$, HMPD- σ and STRAIGHT. In the case of HMPD, the pitch modification factor was applied to all $f_0(t_i)$ values, without any distinction between voiced and unvoiced segments, while in STRAIGHT unvoiced segments were obviously kept unvoiced.

Using a web-based interface, 30 listeners gave their pairwise preferences for the three possible combinations of methods using a 5-points scale [69]: strong preference for one method, preference for one method, preference for the other method, strong preference for the other method, or uncertainty. Again, each listener

assessed the quality of the upwards and downwards shifts of the recordings of 2 languages, for one female and one male speaker per language. The individual scores given by listeners were then aggregated into a single mean score for each method, which shows global preference of one method against all the others.

The results shown in Fig. 7 indicate that HMPD- $\mu\sigma$ is less preferred than HMPD- σ . Deeper investigation based on informal listening revealed that a low-frequency resonance could be perceived in some signals after HMPD- $\mu\sigma$ manipulation. This might corresponds to the glottal formant effect [71] which is not properly handled in our manipulation. Indeed, keeping the phase characteristics of the original glottal source, as $\mu_i(f)$ is supposed to do, does not make sense after pitch scaling by one octave. Finally, this can also be interpreted as a symptom that $\mu_i(f)$ is not reproducing the maximum-phase component properly. In any case, the resulting artifacts make signals more unnatural. By discarding the contribution of $\mu_i(f)$, HMPD- σ avoids this issue and achieves better quality. This is the reason why only HMPD- σ was considered for evaluation in the next Sec. 5.3.

Concerning the comparison between HMPD and STRAIGHT, for upwards pitch shifting STRAIGHT is clearly preferred over HMPD- σ . However, for downwards shifts, clear preferences are shown for HMPD- σ . Informal listening revealed that for upwards pitch shifting the speech signals modified by HMPD sound tenser and lack some noisiness. This is due to the inherent limitations of modeling speech exclusively through harmonics: even for an adequate phase variance across time, at high pitch values the frequency gap between every two consecutive harmonics does not allow a proper reconstruction of noise characteristics. STRAIGHT is not prone to this effect because it uses a wideband noise [28]. This is undoubtedly one issue in HMPD to be solved in future works.

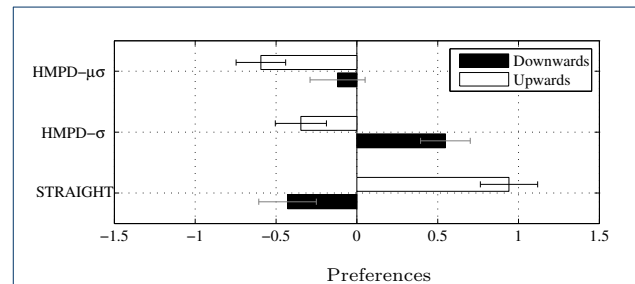


Figure 7 Preferences for 30 listeners of upwards and downwards pitch shifts of one octave, with the 95% confidence intervals.

5.3 Quality of statistical parametric speech synthesis

To assess the quality of HMPD- σ in statistical parametric speech synthesis, we built a system based on HTS, the open-source HMM-based speech synthesis system [72](v2.1.1). HTS learns a correspondence between labels containing phonetic, linguistic and prosodic information and one/many streams of vectors containing acoustic features. This correspondence is modeled at phone level through 5-state left-to-right context-dependent HMMs with explicit state duration distributions. The technology behind this well known system is explained in detail in [4].

Both HMPD and STRAIGHT were slightly modified to meet the requirements of HTS. In both of them, 39th-order Mel-cepstral (MCEP) coefficients were used to model the amplitude envelope $|V_i(f)|$ as suggested in [73], the only difference being that in HMPD these coefficients were obtained from discrete harmonic amplitudes as in [29]. To model the degree of noisiness, the aperiodicity measures provided by STRAIGHT were averaged within 5 meaningful bands, as detailed in [73], whereas HMPD's $\sigma_i(f)$, which takes values in the range $[0, \infty)$ like $|V_i(f)|$, was also translated into MCEP coefficients (order 12). For synthesis, the $\sigma_i(f)$ on linear scale was recovered from the corresponding MCEP coefficients, like the amplitude envelope. Given the importance of pitch artifacts in HMM-based speech synthesis, for a fair comparison, we used the same $f_0(t_i)$ values for both vocoders, namely those provided by STRAIGHT. In unvoiced segments, the continuous $f_0(t)$ curve required by HMPD- σ was simply obtained by linear interpolation of the non-zero $f_0(t_i)$ values. The resulting curve was then modeled using continuous HMMs with one Gaussian mixture per state instead of MSD-HMMs, as proposed by [74]. Before presenting the generated utterances to the evaluators, we manually checked that no significant prosodic differences were present between the two vocoders. During synthesis, all parameter streams were generated through the standard maximum likelihood parameter generation procedure with global variance enhancement. Table 1 summarizes the settings used for the streams.

We trained models for 4 different speech databases: one female and one male speaker in Spanish, containing 1.2k and 2k utterances respectively [75, 76], and one female and one male speaker in English, containing 1.1k and 2.8k utterances respectively [77, 78], all with $f_s = 16$ kHz). All the samples using STRAIGHT and HMPD are available at [70]. For the sack of completeness, samples using impulse-based glottal sources ($\mu_i(f) = 0$ and $\sigma_i(f) = 0 \forall i, f$ in the whole signal or only in the voiced segments, as often used in the literature as baseline systems [79, 80]) have also been

Harmonic Model + Phase Distortion (HMPD)			
	Stream 1	Stream 2	Stream 3
Content	f_0	Amp.Env.	$\sigma_i(f)$
Parameters	log	MCEP(39)	MCEP(12)
Model	Cont.HMM	Cont.HMM	Cont.HMM

STRAIGHT[10, 28]			
	Stream 1	Stream 2	Stream 3
Content	f_0	Amp.Env.	Aperiodicity
Parameters	log	MCEP(39)	Bandwise(5)
Model	MSD-HMM	Cont.HMM	Cont.HMM

Table 1 Summary of the streams used in the HMM-based synthesis system. Amp.Env. is the Amplitude Envelope, MCEP(39) stands for mel-cepstral coefficients of order 39, Cont.HMM stands for continuous HMM model, MSD-HMM stands for the Multi-Space Distribution (MSD)[32].

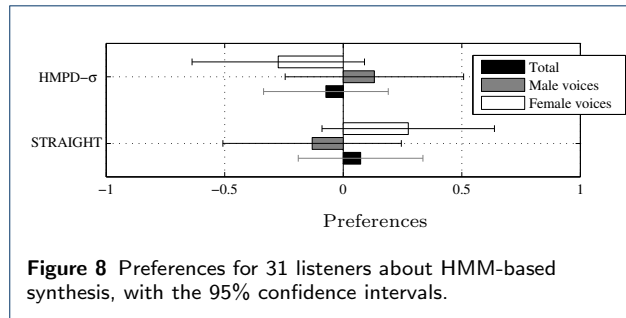
generated and are available on the demonstration page [70]. However, given their very poor quality, they have not been included in the following listening test in order to avoid their potential influence on the results of STRAIGHT and HMPD. Therefore, we conducted a pairwise preference test between STRAIGHT and HMPD only, similar to that of Sec. 5.2. For each voice, 31 listeners gave their preference for each method for one synthetic utterance randomly selected among 10.

Fig. 8 shows the global mean preferences. Since there are only two systems under comparison, the preferences are symmetrical. The results show no significant difference between the two systems despite the high number of evaluators. Therefore, the quality achieved by HMPD is comparable to that of the state of the art, while it uniformizes both the speech representation and the modeling process by discarding the voicing decision. More importantly, this preliminary experiment is a reliable confirmation that phase variance across time can inspire features that succeed at capturing the time- and frequency-varying degree of noisiness of speech in the aHM framework.

Interestingly, the gender dependencies observed in the previous experiments also arise in Fig. 8. Indeed, listeners seem to prefer the female voices of STRAIGHT and the male voices of HMPD- σ . As mentioned in Sec. 5.2, this phenomenon is due to the inherent limitations of harmonic modeling at high pitch values. Forthcoming works will address this issue.

6 Conclusions

In this paper, features based on mean and standard-deviation of the Phase Distortion (PD), have been suggested for analysis/synthesis of speech signals, leading to a new Harmonic Model + Phase Distortion (HMPD) vocoder. These features avoid voiced and unvoiced segmentation. Thus, the perceived quality of HMPD's synthesis is independent of the reliability of a voicing estimator. A first listening test has shown



that HMPD's resynthesis quality is as good as that of the STRAIGHT vocoder for female voices and better for male voices. A second preference test about pitch scaling has shown a limitation of HMPD when the harmonics are not dense enough to properly reproduce noise properties (e.g. with high f_0). Future works are planned to address this fundamental issue of the harmonic models. However, a clear preference has been shown for HMPD in downwards shifts, suggesting that additive wideband noise, often used in existing vocoders, is not necessary for low pitched voices. A last test has suggested that the quality of HMPD in HMM-based speech synthesis is similar to that of the state-of-the-art. Therefore, HMPD basically simplifies the signal representation, in terms of uniformity, by removing the voicing decision, without losing, on average, perceived quality.

Acknowledgements

G. Degottex has been funded by the Thanks to the Swiss National Science Foundation (SNSF) (grants PBSKP2_134325, PBSKP2_140021), Switzerland, and the Foundation for Research and Technology-Hellas (FORTH), Heraklion, Greece.

Author details

¹Computer Science Department, University of Crete (UOC-CSD), Heraklion, Greece. ²Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH-ICS), Heraklion, Greece.

³Basque Science Foundation (IKERBASQUE), Bilbao, Spain. ⁴Aholab, University of the Basque Country, Bilbao, Spain.

References

- Gales, M.J.F., Young, S.J.: The application of hidden markov models in speech recognition. *Foundations and Trends in Signal Processing* **1**(3), 195–304 (2007)
- Kinnunen, T., Li, H.: An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication* **52**(1), 12–40 (2010)
- Stylianou, Y.: Harmonic plus noise models for speech combined with statistical methods, for speech and speaker modification. PhD thesis, TelecomParis, France (1996)
- Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.: Speech synthesis based on hidden markov models. *Proceedings of the IEEE* **101**(5), 1234–1252 (2013)
- Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, O., Vinyals, O.: Speaker diarization: a review of recent research. *IEEE Trans. Audio, Speech, and Language Processing* **20**(2), 356–370 (2012)
- Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.

- IEEE Transactions on Acoustics, Speech, and Signal Processing* **28**, 357–366 (1980)
- Spanias, A.S.: Speech coding: a tutorial review. *Proceedings of the IEEE* **82**(10), 1541–1582 (1994)
- Scott, J.M., Assmann, P.F., Nearey, T.M.: Intelligibility of frequency shifted speech. *The Journal of the Acoustical Society of America* **109**(5), 2316–2316 (2001)
- Schweinberger, S.R., Casper, C., Hauthal, N., Kaufmann, J.M., Kawahara, H., Kloth, N., Robertson, D.M., Simpson, A.P., Zäske, R.: Auditory adaptation in voice perception. *Current Biology* **6**(9), 684–688 (2008)
- Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication* **27**(3–4), 187–207 (1999)
- McAulay, R., Quatieri, T.: Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing* **34**(4), 744–754 (1986)
- Quatieri, T., McAulay, R.J.: Speech transformations based on a sinusoidal representation. In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, vol. 10, pp. 489–492 (1985)
- Quatieri, T.F., McAulay, R.: Shape invariant time-scale and pitch modification of speech. *IEEE Trans. on Signal Processing* **40**(3), 497–510 (1992)
- Quatieri, T.F., McAulay, R.: Phase coherence in speech reconstruction for enhancement and coding applications. In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, pp. 207–210 (1989)
- Pantazis, Y., Rosec, O., Stylianou, Y.: Adaptive AM-FM signal decomposition with application to speech analysis. *IEEE Transactions on Audio, Speech, and Language Processing* **19**(2), 290–300 (2010)
- Degottex, G., Stylianou, Y.: Analysis and synthesis of speech using an adaptive full-band harmonic model. *IEEE Trans. on Audio, Speech and Lang. Proc.* **21**(10), 2085–2095 (2013)
- Kafentzis, G., Degottex, G., Rosec, O., Stylianou, Y.: Time-scale modifications based on a full-band adaptive harmonic model. In: *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8193–8197 (2013)
- Kafentzis, G., Degottex, G., Rosec, O., Stylianou, Y.: Pitch modifications of speech based on an adaptive harmonic model. In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)* (2014)
- Laroche, J., Stylianou, Y., Moulines, E.: HNS: Speech modification based on a harmonic+noise model. In: *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 550–553 (1993)
- Richard, G., d'Alessandro, C.: Analysis/synthesis and modification of the speech aperiodic component. *Speech Communication* **19**(3), 221–244 (1996)
- Banos, E., Erro, D., Bonafonte, A., Moreno, A.: Flexible harmonic/stochastic modelling for hmm-based speech synthesis. In: *Proc. V Jornadas en Tecnologías del Habla* (2008)
- El-Jaroudi, A., Makhoul, J.: Discrete all-pole modeling. *IEEE Transactions on Signal Processing* **39**(2), 411–423 (1991)
- Campedel-Oudot, M., Cappe, O., Moulines, E.: Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach. *IEEE Transactions on Speech and Audio Processing* **9**(5), 469–481 (2001)
- Saratxaga, I., Hernaez, I., Erro, D., Navas, E., Sanchez, J.: Simple representation of signal phase for harmonic speech models. *Electronics Letters* **45**(7), 381–383 (2009)
- Degottex, G., Roebel, A., Rodet, X.: Function of phase-distortion for glottal model estimation. In: *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4608–4611 (2011)
- Ohtani, Y., Tamura, M., Morita, M., Kagoshima, T., Akamine, M.: HMM-based speech synthesis using sub-band basis spectrum model. In: *Proc. Interspeech*, pp. 1440–1443 (2012)
- Maia, R., Akamine, M., Gales, M.J.F.: Complex cepstrum for statistical parametric speech synthesis. *Speech Communication* **55**(5), 606–618 (2013)
- Kawahara, H., Estill, J., Fujimura, O.: Aperiodicity extraction and

- control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In: MAVEBA (2001)
29. Erro, D., Sainz, I., Navas, E., Hernaez, I.: Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE Journal of Selected Topics in Signal Processing* (2014)
30. Latorre, J., Gales, M.J.F., Buchholz, S., Knill, K., Tamurd, M., Ohtani, Y., Akamine, M.: Continuous f0 in the source-excitation generation for hmm-based tts: Do we need voiced/unvoiced classification? In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4724–4727 (2011)
31. Degottex, G., Lanchantin, P., Roebel, A., Rodet, X.: Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis. *Speech Comm.* **55**(2), 278–294 (2013)
32. Tokuda, K., Masuko, T., Myizaki, N., Kobayashi, T.: Multi-space probability distribution HMM. *IEICE Transactions on Information and Systems* **E85-D**, 455–464 (2002)
33. Saratxaga, I., Hernaez, I., Pucher, M., Sainz, I.: Perceptual importance of the phase related information in speech. In: *Proc. Interspeech* (2012). ISCA
34. Mowlae, P., Saiedi, R.: Iterative closed-loop phase-aware single-channel speech enhancement. *Signal Processing Letters, IEEE* **20**(12), 1235–1239 (2013)
35. Mowlae, P., Saiedi, R., Martin, R.: Phase estimation for signal reconstruction in single-channel speech separation. In: *Proceedings of the International Conference on Spoken Language Processing*, pp. 1–4 (2012)
36. Miller, R.L.: Nature of the vocal cord wave. *Journal of the Acoustical Society of America* **31**(6), 667–677 (1959)
37. Oppenheim, A.V., Schaffer, R.W.: *Digital Signal Processing*, 2nd edn. Prentice-Hall, New Jersey, U.S.A. (1978)
38. Paul, D.B.: The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech and Signal Processing* **29**(4), 786–794 (1981)
39. Doval, B., d'Alessandro, C., Henrich, N.: The voice source as a causal/anticausal linear filter. In: *Proc. ISCA Voice Quality: Functions, Analysis and Synthesis (VOQUAL)*, pp. 16–20 (2003)
40. Degottex, G., Roebel, A., Rodet, X.: Phase minimization for glottal model estimation. *IEEE Transactions on Audio, Speech, and Language Processing* **19**(5), 1080–1090 (2011)
41. Oppenheim, A., Schaffer, R., Stockham, T.: Nonlinear filtering of multiplied and convolved signals. *Proceedings of the IEEE* **56**(8), 1264–1291 (1968)
42. Bozkurt, B., Doval, B., d'Alessandro, C., Dutoit, T.: Zeros of Z-Transform (ZZT) decomposition of speech for source-tract separation. *ICSLP* (2004)
43. Drugman, T., Bozkurt, B., Dutoit, T.: Complex cepstrum-based decomposition of speech for glottal source estimation. In: *Proc. Interspeech*, pp. 116–119 (2009)
44. Drugman, T., Dubuisson, T., Moinet, A., d'Alessandro, C., Dutoit, T.: Glottal source estimation robustness. In: *Proc. International Conference on Signal Processing and Multimedia Applications (SIGMAP)* (2008)
45. Laroche, J., Dolson, M.: Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing* **7**(3), 323–332 (1999)
46. Stylianou, Y.: Removing linear phase mismatches in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing* **9**(3), 232–239 (2001)
47. Agiomyriannakis, Y., Stylianou, Y.: Wrapped gaussian mixture models for modeling and high-rate quantization of phase data of speech. *IEEE Trans. on Audio, Speech, and Lang. Proc.* **17**(4), 775–786 (2009)
48. Smits, R., Yegnanarayana, B.: Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Speech and Audio Processing* **3**(5), 325–333 (1995)
49. Ananthapadmanabha, T., Yegnanarayana, B.: Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech and Signal Processing* **27**(4), 309–319 (1979)
50. Moulines, E., Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* **9**(5–6), 453–467 (1990). *Neurospeech '89*
51. Hamon, C., Mouline, E., Charpentier, F.: A diphone synthesis system based on time-domain prosodic modifications of speech. In: *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 238–241 (1989)
52. Lipshitz, S.P., Pocock, M., Vanderkooy, J.: On the audibility of midrange phase distortion in audio systems. *J. Audio Eng. Soc.* **30**(9), 580–595 (1982)
53. Hansen, V., Madsen, E.R.: On aural phase detection: Part 1. *J. Audio Eng. Soc.* **22**(1), 10–14 (1974)
54. Hansen, V., Madsen, E.R.: On aural phase detection: Part 2. *J. Audio Eng. Soc.* **22**(10), 783–788 (1974)
55. Tahon, M., Degottex, G., Devillers, L.: Usual voice quality features and glottal features for emotional valence detection. In: *Proc. International Conference on Speech Prosody*, pp. 693–696 (2012)
56. Banno, H., Takeda, K., Itakura, F.: The effect of group delay spectrum on timbre. *Acoustical Science and Technology* **23**(1), 1–9 (2002)
57. Yegnanarayana, B., Saikia, D., Krishnan, T.: Significance of group delay functions in signal reconstruction from spectral magnitude or phase. *Acoustics, Speech and Signal Processing, IEEE Transactions on* **32**(3), 610–623 (1984)
58. Murthy, H.A., Yegnanarayana, B.: Speech processing using group delay functions. *Elsevier Signal Processing* **22**, 259–267 (1991)
59. Zhu, D., Paliwal, K.K.: Product of power spectrum and group delay function for speech recognition. In: *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference On*, vol. 1, pp. 125–81 (2004)
60. Naylor, P.A., Kounoudes, A., Gudnason, J., Brookes, M.: Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Transactions on Audio, Speech and Language Processing* **15**(1), 34–43 (2007)
61. Drugman, T., Dubuisson, T., Dutoit, T.: Phase-based information for voice pathology detection. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4612–4615 (2011)
62. Shiga, Y., King, S.: Estimation of voice source and vocal tract characteristics based on multi-frame analysis. In: *Proc. Eurospeech*, vol. 3, pp. 1749–1752 (2003)
63. Bonada, J.: High quality voice transformations based on modeling radiated voice pulses in frequency domain. In: *Proc. Digital Audio Effects (DAFx)* (2004)
64. Fisher, N.I.: *Statistical Analysis of Circular Data*. Cambridge University Press, U.K. (1995)
65. McAulay, R.J., Quatieri, T.F.: Sine-wave phase coding at low data rates. In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, vol. 1, pp. 577–580 (1991)
66. McAulay, R., Quatieri, T.F.: Multirate sinusoidal transform coding at rates from 2.4 kbps to 8 kbps. In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, vol. 12, pp. 1645–1648 (1987)
67. Sugiyama, A., Miyahara, R.: Phase randomization - A new paradigm for single-channel signal enhancement. In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, pp. 7487–7491 (2013)
68. Assembly, T.I.R.: ITU-T P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Technical report, ITU (2000)
69. Assembly, T.I.R.: ITU-R BS.1284-1: General methods for the subjective assessment of sound quality. Technical report, ITU (2003)
70. Degottex, G., Erro, D.: Demonstrations audio samples of HMPD-based synthesis (2014). <http://gillesdegottex.eu/Ex2013hmpd> Accessed 31 May 2014
71. Doval, B., d'Alessandro, C., Henrich, N.: The spectrum of glottal flow models. *Acta acustica united with acustica* **92**(6), 1026–1046 (2006)
72. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., Tokuda, K.: The HMM-based speech synthesis system (HTS) version 2.0. In: *Proc. ISCA Workshop on Speech Synthesis (SSW)* (2007)
73. Zen, H., Toda, T., Nakamura, M., Tokuda, K.: Details of the nitech HMM-based speech synthesis system for the blizzard challenge 2005. *IEICE Trans. Inf. Syst.* **E90-D**(1), 325–333 (2007)
74. Yu, K., Young, S.: Continuous f0 modeling for HMM based statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing* **19**(5), 1071–1079 (2011)
75. Sainz, I., Erro, D., Navas, E., Hernaez, I., Sánchez, J., Saratxaga, I.,

- Odriozola, I.: Versatile speech databases for high quality synthesis for basque. In: Proc. of European Language Resources Association (ELRA) (2012)
76. Rodriguez-Banga, E., Garcia-Mateo, C.: Documentation of the uvigo.esda spanish database. Technical report, Universidade de Vigo (2010)
77. Kominek, J., Black, A.W.: The CMU ARCTIC speech databases. In: Proc. ISCA Speech Synthesis Workshop, pp. 223–224 (2003)
78. Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., Tang, Y.: Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication* **55**(4), 572–585 (2013)
79. Erro, D., Sainz, I., Navas, E., Hernaez, I.: Hnm-based mfcc+f0 extractor applied to statistical speech synthesis. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP), pp. 4728–4731 (2011)
80. Lanchantin, P., Degottex, G., Rodet, X.: A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, USA, pp. 4630–4633 (2010)