A Measure of Phase Randomness for the Harmonic Model in Speech Synthesis

Gilles Degottex[1] *and Daniel Erro*[2]

[1]University of Crete and Foundation for Research and Technology-Hellas, Heraklion, Greece
[2]Aholab, University of Basque Country, IKERBASQUE, Basque Science Foundation, Bilbao, Spain
degottex@csd.uoc.gr, derro@aholab.ehu.es

Abstract

Modern statistical speech processing frameworks require the speech signals to be translated into feature vectors by means of vocoders. While features representing the amplitude envelope already exist (e.g. MFCC, LSF), parametrizing the phase information is far from straightforward, not only because it is a circular data, but also because it shows an irregular behaviour in noisy time-frequency regions. Thus, many vocoders reconstruct speech by using minimum phases and random phases, relying on a previous voicing decision. In this paper, a phase feature is suggested to represent the randomness of the phase across the full time-frequency plan, in both voiced and unvoiced segments, without voicing decision. Resynthesis experiments show that, when integrated into a full-band harmonic vocoder, the suggested randomization feature is slightly better, on average, to STRAIGHT's aperiodicity. In HMM-based synthesis, the results show that the suggested vocoder reduces the complexity of the analysis and statistical modelling by removing the voicing decision, while keeping the perceived quality.

Index Terms: Speech synthesis, harmonic model, phase modelling, parametric speech synthesis.

1. Introduction

Applications like speech coding [1] require representations using a set of features yielding all the perceived characteristics in synthesis. In speech transformation [2, 3] and synthesis [4], speech models must also allow high-quality manipulation. Accordingly, features must be linked to the way speech is produced and perceived. Moreover, recent statistical trends are also encouraging speech representations with a constant number of parameters and with good mathematical properties [4].

Sinusoidal models [2] are one of the most popular speech representations. They decompose the signal into a number of sinusoids given by their frequency, amplitude and instantaneous phase. Among them, harmonic models [5] assume the frequencies to be integer multiples of a fundamental frequency $f_0(t)$, thus solving the frequency matching between successive frames. Harmonic models have been widely used in different applications [6, 7]. Recently, continuous full-band harmonic models have been shown to achieve excellent quality in resynthesis [8, 9], while representing both voiced and unvoiced segments in the same way, thus, simplifying processing techniques [9]. Unfortunately, sinusoidal parameters are not suitable for modern statistical frameworks because amplitude and phase parameters lie on the harmonic grid which depends on $f_0(t)$ [10]. This problem is easily overcome for amplitudes, which can be interpolated and parametrized as spectral envelopes [11]. However, the instantaneous phase constantly wraps across time because of the frequency integral; in addition, it exhibits an irregular behaviour in noisy time-frequency regions. This makes phase envelopes tricky to estimate [12]. Consequently, standard parametrizations used in statistical frameworks tend to oversimplify the phase information, using a minimum-phase component derived from the amplitude envelope and complementary parameters related to the degree of harmonicity in different time-frequency regions [13, 14, 11], the voiced/unvoiced decision being of crucial importance [15, 14, 12, 11]. To exploit the potential of the continuous full-band harmonic model in statistical frameworks, phase information needs to be coded in a uniform way regardless of the local degree of harmonicity. Ultimately, ideal phase features should convey both waveform-related information and noisiness information. However, even though the perceptual importance of phase has been demonstrated in the general case [16, 17, 18, 19], that of waveform-related phase information is still source of controversy in speech processing [20, 21]. Therefore, in the first stage of this work we decided to focus only on the representation of noisy time-frequency regions by means of phase.

In this paper, we suggest to model the noisiness of the speech signal in a continuous manner across time and frequency by controlling the randomness of the instantaneous phase. The method is inspired by phase randomization techniques that already existed for coding [3, 22]. To estimate the instantaneous phase from the waveform, the adaptive Harmonic Model (aHM) is employed in this work [8]. We use the concept of Phase Distortion (PD) [23, 24], which is related to the Relative Phase Shift (RPS) [25, 20], to extract meaningful characteristics from the instantaneous phase. We use PD because it has been shown to be directly linked to the shape of the glottal pulse through the maximum-phase component of the signal [23]. We then estimate the phase randomness feature across the full time-frequency space by computing the PDs standard-Deviation (PDD) on a short-term sliding window. This novel approach has the following advantages: (i) During analysis, it allows to represent harmonic and noise components that overlap in both time and frequency. This avoids binary voiced/unvoiced decisions, which are error-prone and result in misclassification and then audible artifacts [26]. (ii) During synthesis, since we generate noise through phase randomization, it avoids an independent synthesis of harmonics and noisy components [13, 27]. This provides a solid and uniform framework, thus avoiding artifacts near the voicing boundaries and risks of synthesizing perceptually independent sounds[10, 28]. (iii) It can be easily made compatible with statistical frameworks. For statistical parametric synthesis, given the continuous nature of the suggested feature, the use of multi-space distributions (MSD) [29] can be avoided, as suggested in [30]. In that sense, the training and generation procedures are simplified.

The next sections describe the feature based on PD's standard-Deviation (PDD). Then, the evaluation section shows the importance of the feature and demonstrates the feasibility of the suggested representation for speech synthesis.

2. Instantaneous phase representations

Given the speech waveform $s(t)$, we assume that its $f_0(t)$ curve is known a priori. In this work, the STRAIGHT method is used, which allowed fair comparisons during evaluations. Sinusoidal parameters are first estimated, N times per period, at analysis instants:

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

In this work, we chose $N = 4$, which ensures a minimal number of analysis instant for short-term statistical characterization. In a Blackman window of 3 pitch periods around each t_i, the aHM model represents the full-band of the analytic signal $s_i(t)$ by [8]:

$$s_i(t) = \sum_{h=1}^{H_i} a_{i,h} \cdot e^{j(h\phi_0(t)+\phi_{i,h})} \tag{2}$$

where i is the frame index, $H_i = \lfloor 0.5 f_s / f_0(t_i) \rfloor$, $a_{i,h}$ is the real-valued amplitude of the h^{th}-harmonic, $\phi_{i,h}$ is the instantaneous phase and $\phi_0(t)$ is a real function:

$$\phi_0(t) = \frac{2\pi}{f_s}\int_{t_i}^{t} f_0(\tau)d\tau \tag{3}$$

where f_s denotes the sampling frequency. In [8], it has been shown that aHM can represent both voiced and unvoiced segments uniformly, without voicing decision, assuming that an $f_0(t)$ curves can be obtained in unvoiced segments in a controlled range ([60Hz, 500Hz] in our study). Additionally, together with its harmonic tracking algorithm[8], this model provides almost always the most accurate and precise sinusoidal parameters compared to state-of-the-art methods. Eventually, this accuracy might not be critical for obtaining the results presented in this paper. However, this allows to minimize the influence of the sinusoidal parameter estimation on the results and, thus, to strengthen the link between the suggested phase feature and the results obtained. Finally, the resynthesis obtained by aHM is almost indistinguishable from the original recording [8], which ensures that the aforementioned properties come with no perceptual degradation. However, aHM cannot be used, as it is, in statistical approaches because $\phi_{i,h}$ wraps constantly across time. In this paper, using the parameters $f_0(t_i)$, $a_{i,h}$ and $\phi_{i,h}$, the goal is to represent the extent of phase randomness into a feature, that we assume to be the most perceived characteristic of $\phi_{i,h}$. Modelling the amplitude is not the subject of this work. Thus, we used a simple linear interpolation of $a_{i,h}$ across frequency in order to build an amplitude spectral envelope $A_i(f)$. We also assume that $A_i(f)$ approximates the Vocal Tract Filter (VTF) response, which is assumed to be minimum-phase. Thus, $\angle A_i(f)$ can be retrieved through the real cepstrum [31]. The following sections address the construction of the phase feature, which has to carry the property of randomness of $\phi_{i,h}$.

2.1. Model of the instantaneous phase

Models of $\phi_{i,h}$ have been already suggested, for phase synchronization between frames [32] and speech coding [3, 22]. In this paper, we suggest to use a model similar to that in [22]:

$$\phi_{i,h} = \overbrace{\underbrace{\theta_{i,h}}_{\text{source shape}} + \underbrace{h\frac{2\pi}{f_s}\int_{c_i}^{t_i} f_0(\tau)d\tau}_{\text{linear phase}}}^{\text{voice source}} + \overbrace{\angle A_i(h f_0(t_i))}^{\text{filter}} \tag{4}$$

whose terms are described here below. In voiced segments, the glottal pulse has a shape which is mainly a maximum-phase component [31, 23]. In unvoiced segments, one can assume that this shape is basically random. Moreover, from one frame to the next, this random shape also changes randomly. The *source shape* term $\theta_{i,h}$ represents this shape for both voiced and unvoiced cases. In speech processing techniques, a reference time instant c_i is often used for each glottal pulse (e.g. Glottal Closure Instant (GCI), energy local maximum of a residual signal, pitch pulse onset [2, 3]). Even though such a definition is necessary for many approaches, we will show that this value is discarded when using the RPS [25, 20], as employed in this work. Nevertheless, since the analysis is not pitch synchronous ($c_i \neq t_i$), the delay between t_i (the window's center) and the position of the source shape has to be represented by a *linear phase*. Here, the integral form is used since $f_0(t)$ is not constant in the analysis window when using aHM. Finally, according to the voice production, the *voice source* is convolved by the vocal tract impulse response. Thus, the minimum-phase $\angle A_i(\omega)$ adds to the model.

2.2. Motivations for using the Phase Distortion (PD)

In order to address the wrapping of the harmonic phase values $\phi_{i,h}$ from one t_i to the next (due to the linear phase term), the RPS has been suggested[20], which is expressed as:

$$\text{RPS}_{i,h} = \phi_{i,h} - h\phi_{i,1} \tag{5}$$

To further analyze the results of this computation, the estimated $\phi_{i,h}$ in (2) can be replaced by its model from (4):

$$\begin{aligned} \text{RPS}_{i,h} &= \theta_{i,h} + h\tfrac{2\pi}{f_s}\textstyle\int_{c_i}^{t_i} f_0(\tau)d\tau + \angle A_i(h f_0(t_i)) \\ &\quad -h\cdot\left(\theta_{i,1} + \tfrac{2\pi}{f_s}\textstyle\int_{c_i}^{t_i} f_0(\tau)d\tau + \angle A_i(f_0(t_i))\right) \\ &= \theta_{i,h} - h\theta_{i,1} + \angle A_i(h f_0(t_i)) - h\angle A_i(f_0(t_i)) \end{aligned} \tag{6}$$

Eq. (6) shows that the linear phase is discarded. This is convenient since this term has no significant perceptual content besides its derivative, the fundamental frequency $f_0(t)$, which is already known. Moreover, c_i is also discarded, so that there is no need to estimate any GCI or pitch pulse onset, which avoids drawbacks of misestimation of such time instants. The RPS can also be computed on the Linear Prediction (LP) residual, thus, removing most of the contribution of the VTF. Similarly, let's define: $\widetilde{\phi}_{i,h} = \phi_{i,h} - \angle A_i(h f_0(t_i))$. Thus, (6) reduces to:

$$\widetilde{\text{RPS}}_{i,h} = \widetilde{\phi}_{i,h} - h\widetilde{\phi}_{i,h} = \theta_{i,h} - h\theta_{i,1} \tag{7}$$

In (7), only the source shape and the harmonic number h remain. This harmonic number is still inconvenient because it belongs to the harmonic structure which should be removed. Additionally, the harmonic number increases the RPS variance towards high frequencies, thus, drowning the variance of $\theta_{i,h}$ into that of $\theta_{i,1}$. Computing the phase difference between harmonics alleviates this issue. The phase difference between two components is known as Phase Distortion (PD) [18], whose general perceived characteristics are already known [16, 17, 18]. Using a harmonic model, the PD between consecutive harmonics is equal to a finite difference, which is also similar to the group-delay. Additionally, we recently suggested to use the PD on (7) for glottal parameter estimation [23, 24]. The rather complicate equation in [23] is actually equal to:

$$\text{PD}_{i,h} = \underset{h}{\Delta}\widetilde{\text{RPS}}_{i,h} = \left(\widetilde{\phi}_{i,h+1} - (h+1)\widetilde{\phi}_{i,1}\right) - \left(\widetilde{\phi}_{i,h} - h\widetilde{\phi}_{i,1}\right) \tag{8}$$

where $\underset{h}{\Delta}$ is the finite difference operator. Replacing $\phi_{i,h}$ (2) by its model (4) leads to:

$$\text{PD}_{i,h} = \theta_{i,h+1} - \theta_{i,h} - \theta_{i,1} \tag{9}$$

Compared to (7), the harmonic number h is removed. Consequently, PD is only related to the source shape.

Even though the influence of h is drastically reduced in $\text{PD}_{i,h}$, the values are still defined on harmonic indices. In order to obtain a continuous frequency scale, $\text{PD}_{i,h}$ is unwrapped and linearly interpolated across frequency, i.e. $\text{PD}_{i,h} \Rightarrow \text{PD}_i(f)$, assuming that 512 frequency bins are sufficient to represent a continuous scale for $f_s = 44$kHz. For reason of simplicity, the continuous notation will be used in the following.

2.3. Short-term PD standard-Deviation (PDD) feature

On a frame-by-frame basis, it has been shown that the sole information carried by $\text{PD}_i(f)$ is sufficient to reconstruct an instantaneous phase which has all the perceived characteristics of $\phi_{i,h}$ [20]. However, through manipulation of $\text{PD}_i(f)$ (time or frequency scaling, statistical models for speech synthesis, voice conversion), the original statistical characteristics of $\text{PD}_i(f)$ might not be preserved. Therefore, in this paper, in order to preserve the noisiness in time-frequency regions, we suggest to preserve the short-term standard-deviation of $\text{PD}_i(f)$ in a feature used for synthesis. The short-term Phase Distortion's standard-Deviation (PDD) is computed over a sliding time window of $\text{PD}_i(f)$ values.

In voiced segments, the shape of the glottal pulse smoothly changes. Thus, $\text{PD}_i(f)$ has a trend at each t_i. In order to properly estimate PDD, which should represent only the noisiness and not the trend, this trend has first to be removed from $\text{PD}_i(f)$. Otherwise, PDD would be systematically overestimated. The trend $\widehat{\text{PD}}_i(f)$ is first obtained by computing an average PD value over 2 periods, using the formula for circular data [33]:

$$\widehat{\text{PD}}_i(f) = \angle\left(\frac{1}{M}\sum_{m\in C} e^{j\text{PD}_m(f)}\right) \tag{10}$$

where $C = i - \frac{M-1}{2} \cdots i + \frac{M-1}{2}$, and $M = 9$ (i.e. 2 periods). According to informal listening of resynthesis (presented in the next section), using 2 periods is necessary to quickly adapt the PDD estimate to the variations of the speech signal, especially in transients. Then, the PDD is computed by removing the trend and using the formula for circular data [33]:

$$\begin{aligned}\sigma_i(f) &= \operatorname*{std}_{i\in C}\left(\text{PD}_i(f) - \widehat{\text{PD}}_i(f)\right) \\ &= \sqrt{-2\log\left|\frac{1}{M}\sum_{m\in C} e^{j(\text{PD}_m(f) - \widehat{\text{PD}}_m(f))}\right|}\end{aligned} \tag{11}$$

with C and M as in (10). Finally, $\sigma_i(f)$ is still dependent on the time sampling imposed by $f_0(t)$ in (1). In order to remove this dependence $\sigma_i(f)$ is resampled each 5ms. Fig. 1 shows an example of PDD computation.

2.4. Synthesis

The analysis described above provides, each 5ms, features $f_0(t_i)$, $A_i(f)$ and $\sigma_i(f)$, which allow to synthesize a speech signal by the following way. Basically, the synthesis method is similar to that used originally for aHM [8]. Each amplitude and phase tracks, $\hat{a}_h(t)$ and $\hat{\phi}_h(t)$ respectively, are first synthesized independently of the others on a continuous time axis. Then, they are added up all together, without using any windowing scheme:

$$\hat{s}(t) = \sum_{h=1}^{H} \hat{a}_h(t)\cdot\cos(\hat{\phi}_h(t))\cdot\chi_{[hf_0(t)<0.5f_s]}(t) \tag{12}$$

where $H = \text{argmax}_i\lfloor 0.5f_s/f_0(t_i)\rfloor$ and the indicator function $\chi_{[hf_0(t)<0.5f_s]}(t)$ discards any harmonic segment whose frequency is higher than Nyquist. In order to synthesize each amplitude harmonic track $\hat{a}_h(t)$, the amplitude envelope is first sampled at each harmonic frequency $|A_i(hf_0(t_i))|$ and then interpolated across time on a logarithmic scale. The synthetic continuous phase track $\hat{\phi}_h(t)$ is reconstructed using $f_0(t_i)$ and $\sigma_i(f)$. A synthetic PD is first built:

$$\hat{\text{PD}}_{i,h} = \mathcal{WN}(0, \sigma_i(hf_0(t_i))) \tag{13}$$

where $\mathcal{WN}(0,\sigma)$ generates random values which obey a zero-mean wrapped normal distribution of standard-deviation σ [22]. The synthetic RPS is then obtained:

$$\hat{\text{RPS}}_{i,h} = \underset{h}{\Delta}^{-1}\hat{\text{PD}}_{i,h} + \angle A_i(hf_0(t_i)) \tag{14}$$

where $\underset{h}{\Delta}^{-1}$ is the cumulative sum which compensates for $\underset{h}{\Delta}$ in (8). The continuous counterpart $\hat{\text{RPS}}_h(t)$ is obtained by spline interpolation of $\hat{\text{RPS}}_{i,h}$ and $\hat{\phi}_h(t)$ finally results in adding the linear-phase:

$$\hat{\phi}_h(t) = \hat{\text{RPS}}_h(t) + h\frac{2\pi}{f_s}\int_0^t f_0(\tau)d\tau \tag{15}$$

where the low limit of the integral corresponds to the beginning of the signal. Finally, the synthetic signal is obtained by (12). In the following, the complete analysis/synthesis procedure will be called Harmonic Model + Phase Distortion (HMPD).

In (2), the window used to estimate the sinusoidal parameters over-smoothes their variance across time. Additionally, the value M in eq. (11), which has to be small enough to follow the time evolution of the PDD, limits the window of the PDD estimate. Therefore, in practice, $\sigma_i(f)$ is often underestimated. This issue becomes perceptually important in fricatives where the phase has to be fully randomized. To alleviate this problem, we simply suggest to amplify $\sigma_i(f)$ when it passes a given threshold. Through informal listening, we found that an amplification of 10 above a threshold of 0.75 properly reconstruct the noisiness of fricatives while preserving the voiced segments quality.

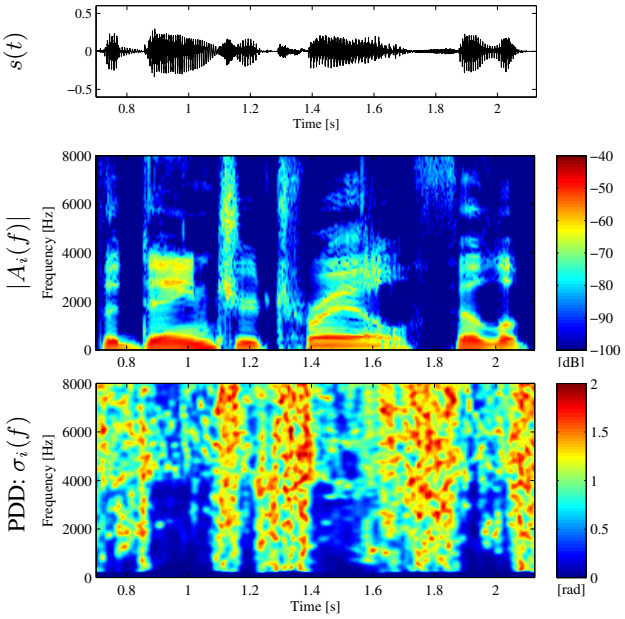

Figure 1: Example of feature extraction: The Phase Distortion standard-Deviation (PDD) characterizes the noisiness of the voice source.

3. Evaluation

An original signal and its resynthesis are not perfectly time synchronous, because the original linear phase is lost in (5) and is only approximated by the integral of $f_0(t)$ in (15). Indeed, the original and synthetic linear phase have the same derivative, i.e. $f_0(t)$. However, an unknown constant exists between the two, since the linear phase is built from the $f_0(t)$ integral. Consequently, signal to Reconstruction Error Ratio (SRER) or objective measurements (e.g. PESQ) can, thus, hardly assess the resynthesis quality as they are sensitive to the waveform. Therefore, in this paper, the evaluations are done using subjective listening tests [34]. The sounds used in the tests are available at: `http://gillesdegottex.eu/Ex2014hmpd`

3.1. Quality of resynthesis

This first test was designed to mainly evaluate the perceived quality in a simple analysis/resynthesis procedure, without any further modelling of the features. Through a web-based interface, listeners were first asked to listen to one original recording among 32 utterances picked up randomly from 16 different languages with both male and female voices. The sampling frequency f_s of the samples varies between 16kHz and 44.1kHz. Then, they were asked to grade, using a 5-points scale [34], the quality of 5 sounds randomly ordered which are: the resynthesis using aHM; that of HMPD; that of STRAIGHT; and the original recording for verification purpose. To obtain a moderate test duration, each listener was asked to grade only the voices of two languages (both male and female voices) randomly selected among the 16. 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 [34], which are averaged in Fig. 2. First, the results show that, in average, the HMPD's quality is slightly better than that of STRAIGHT. Thus, the suggested PDD feature might be a potential improvement and replacement for STRAIGHT's aperiodicity [14]. More precisely, compared to STRAIGHT, the quality of male voices are clearly improved, but not that of female voices. However, the difference of quality between genders is also clearly reduced when using HMPD. Additionally, with a quality similar to that of STRAIGHT, this test shows that PDD allows to model both voiced and unvoiced segments the same way, without the need of any voicing decision. This reduces risks of voicing misestimation and complexity of statistical modelling, as shown in the following test.

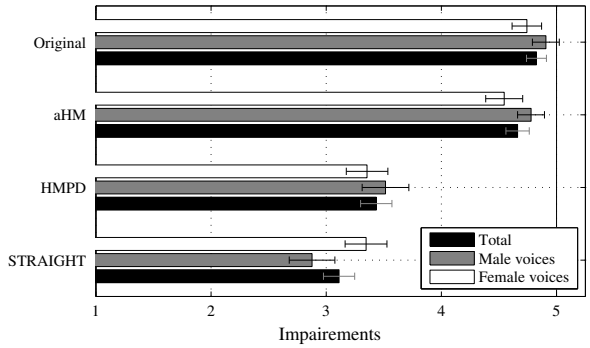

Figure 2: Assessment of quality by 43 listeners of resynthesis methods, with the 95% confidence intervals.

3.2. Quality of statistical parametric speech synthesis

To assess the quality of HMPD in statistical parametric speech synthesis, we built on the HTS HMM-based synthesis system[35](v2.1.1). We trained models for 4 different voice databases: two Spanish (1.2k male utterances [36], 2k female utterances [37]) and two English (2.8k UK male utterances [38], 1.1k US female utterances [39], all with $f_s = 16$kHz). In order to obtain a fair comparison, the STRAIGHT's $f_0(t_i)$ values were used for both HMPD and STRAIGHT. In unvoiced segments, the $f_0(t_i)$ values of HMPD were simply obtained by linear interpolation of the non-zero $f_0(t_i)$ values. The support of $\sigma_i(f)$ being $[0, \infty)$, like $|A_i(f)|$, $\sigma_i(f)$ was compressed, through frequency warping, in order to obtain a feature based on Mel-Frequency Cepstral Coefficients (MFCC), like $|A_i(f)|$. Three streams were considered in HTS: log-f_0, MFCC of $|A_i(f)|$ (order 39) and MFCC of $\sigma_i(f)$ (order 12). For HMPD, as it requires no voicing decision, continuous HMMs with one Gaussian mixture per state were used to model log-f_0, as proposed by [30]. We conducted a preference test where the HMPD-based system was compared to that of STRAIGHT [40]. Using a web-based interface listeners gave their preferences between comparison pairs given by the methods, using a 3-points scale [34]. For each voice, each listener gave his/her preference for one random synthesized utterance among 10. Fig. 3 shows the mean preference scores. Basically, the results show no clear differences between the two systems. The order selection for the compression of PDD might explain why the improvement shown in the previous test does not appear in this one. Future works will address this question. However, HMPD's quality is comparable to that of the state of the art, while simplifying the speech representation by discarding the voicing decision. About the trends, listeners seem to prefer the male voices of HMPD and the female voices of STRAIGHT. This can be explained by HMPD which does not allow a proper reconstruction of noise when a perceptual band lies in a gap between two harmonics. STRAIGHT alleviates this issue by using a wideband noise [14]. Forthcoming works will also address this issue.

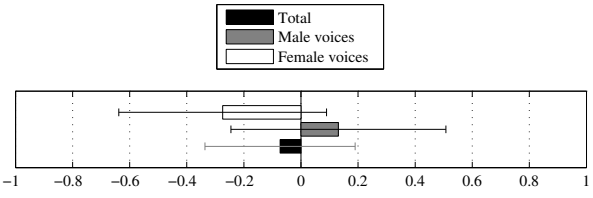

Figure 3: Preferences for 31 listeners about synthesized voices, with the 95% confidence intervals (a positive preference corresponds to a preference for the suggested HMPD).

4. Conclusions

In this paper, a randomization feature based on the Phase Distortion's standard-Deviation (PDD) has been suggested for representation of the noisiness in analysis/synthesis using a harmonic model. The resulting vocoder has been called Harmonic Model + Phase Distortion (HMPD) vocoder. This feature 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, in average, slightly better than that of the state-of-the-art STRAIGHT vocoder. More precisely, it provides similar quality for female voices and better quality for male voices. A second test has shown that the quality of HMPD in HMM-based speech synthesis is similar to that of STRAIGHT. Therefore, HMPD basically reduces the complexity of the signal representation, while keeping the perceived quality.

5. Acknowledgements

The work of G. Degottex was funded by the Swiss National Science Foundation (SNSF) (grant PBSKP2_140021) and the Foundation for Research and Technology-Hellas (FORTH).

6. References

[1] A. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.

[2] T. Quatieri and R. McAulay, "Speech transformations based on a sinusoidal representation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, vol. 10, 1985, pp. 489–492.

[3] ——, "Phase coherence in speech reconstruction for enhancement and coding applications," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1989, pp. 207–210.

[4] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[5] Y. Stylianou, "Harmonic plus noise models for speech combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, TelecomParis, France, 1996.

[6] J. Jensen and J. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 7, pp. 731–740, 2001.

[7] Y. Hu and P. C. Loizou, "On the importance of preserving the harmonics and neighboring partials prior to vocoder processing: Implications for cochlear implants," *Journal of Acoustic Society of America*, vol. 127, no. 1, pp. 427–434, 2010.

[8] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 21, no. 10, pp. 2085–2095, 2013.

[9] G. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou, "Time-scale modifications based on a full-band adaptive harmonic model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[10] E. Banos, D. Erro, A. Bonafonte, and A. Moreno, "Flexible harmonic/stochastic modelling for HMM-based speech synthesis," in *Proc. V Jornadas en Tecnologias del Habla*, 2008.

[11] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, 2014.

[12] J. Bonada, "Voice processing and synthesis by performance sampling and spectral models," Ph.D. dissertation, Universitat Pompeu Fabra, Spain, 2008.

[13] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+noise model," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1993, pp. 550–553.

[14] H. Kawahara, J. Estill, and O. Fujimura, "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.

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[16] V. Hansen and E. R. Madsen, "On aural phase detection: Part 1," *J. Audio Eng. Soc*, vol. 22, no. 1, pp. 10–14, 1974.

[17] ——, "On aural phase detection: Part 2," *J. Audio Eng. Soc*, vol. 22, no. 10, pp. 783–788, 1974.

[18] S. P. Lipshitz, M. Pocock, and J. Vanderkooy, "On the audibility of midrange phase distortion in audio systems," *J. Audio Eng. Soc*, vol. 30, no. 9, pp. 580–595, 1982.

[19] H. Banno, K. Takeda, and F. Itakura, "The effect of group delay spectrum on timbre," *Acoustical Science and Technology*, vol. 23, no. 1, pp. 1–9, 2002.

[20] I. Saratxaga, I. Hernaez, M. Pucher, and I. Sainz, "Perceptual Importance of the Phase Related Information in Speech," in *Proc. Interspeech*. ISCA, 2012.

[21] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *Signal Processing Letters, IEEE*, vol. 20, no. 12, pp. 1235–1239, Dec 2013.

[22] Y. Agiomyrgiannakis and Y. Stylianou, "Wrapped gaussian mixture models for modeling and high-rate quantization of phase data of speech," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 17, no. 4, pp. 775–786, 2009.

[23] G. Degottex, A. Roebel, and X. Rodet, "Function of phase-distortion for glottal model estimation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 4608–4611.

[24] ——, "Phase minimization for glottal model estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1080–1090, 2011.

[25] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, "Simple representation of signal phase for harmonic speech models," *Electronics Letters*, vol. 45, no. 7, pp. 381–383, 2009.

[26] J. Latorre, M. J. F. Gales, S. Buchholz, K. Knill, M. Tamurd, Y. Ohtani, and M. Akamine, "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)*, 2011, pp. 4724–4727.

[27] G. Richard and C. d'Alessandro, "Analysis/synthesis and modification of the speech aperiodic component," *Speech Communication*, vol. 19, no. 3, pp. 221–244, 1996.

[28] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Comm.*, vol. 55, no. 2, pp. 278–294, 2013.

[29] K. Tokuda, T. Masuko, N. Myizaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Transactions on Information and Systems*, vol. E85-D, pp. 455–464, 2002.

[30] K. Yu and S. Young, "Continuous f0 modeling for HMM-based statistical parametric speech synthesis," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 5, pp. 1071–1079, 2011.

[31] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Prentice-Hall, 2nd edition, 1978.

[32] Y. Stylianou, "Removing linear phase mismatches in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 232–239, 2001.

[33] N. I. Fisher, *Statistical Analysis of Circular Data*. Cambridge University Press, Oct. 1995.

[34] T. I. R. Assembly, "ITU-R BS.1284-1: En-general methods for the subjective assessment of sound quality," ITU, Tech. Rep., 2003.

[35] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. ISCA Workshop on Speech Synthesis (SSW), http://hts.sp.nitech.ac.jp*, 2007.

[36] E. Rodriguez-Banga and C. Garcia-Mateo, "Documentation of the uvigo_esda spanish database," Univ. de Vigo, Tech. Rep., 2010.

[37] I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, and I. Odriozola, "Versatile speech databases for high quality synthesis for basque," in *Proc. of LREC'12*. European Language Resources Association (ELRA), 2012.

[38] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[39] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Proc. ISCA Speech Synthesis Workshop*, 2003, pp. 223–224, http://www.festvox.org/cmu_arctic.

[40] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the nitech HMM-based speech synthesis system for the blizzard challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.