A FIXED DIMENSION AND PERCEPTUALLY BASED DYNAMIC SINUSOIDAL MODEL OF SPEECH

Qiong Hu[1], *Yannis Stylianou*[2], *Korin Richmond*[1], *Ranniery Maia*[2], *Junichi Yamagishi*[1,3], *Javier Latorre*[2]

[1]The Centre for Speech Technology Research, University of Edinburgh, UK
[2]Toshiba Research Europe Ltd, Cambridge, UK
[3]National Institute of Informatics, Tokyo, Japan

Qiong.Hu@ed.ac.uk, yannis.stylianou@crl.toshiba.co.uk, korin@inf.ed.ac.uk
ranniery.maia@crl.toshiba.co.uk, jyamagis@inf.ed.ac.uk, javier.latorre@crl.toshiba.co.uk

ABSTRACT

This paper presents a fixed- and low-dimensional, perceptually based dynamic sinusoidal model of speech referred to as PDM (Perceptual Dynamic Model). To decrease and fix the number of sinusoidal components typically used in the standard sinusoidal model, we propose to use only one dynamic sinusoidal component per critical band. For each band, the sinusoid with the maximum spectral amplitude is selected and associated with the centre frequency of that critical band. The model is expanded at low frequencies by incorporating sinusoids at the boundaries of the corresponding bands while at the higher frequencies a modulated noise component is used. A listening test is conducted to compare speech reconstructed with PDM and state-of-the-art models of speech, where all models are constrained to use an equal number of parameters. The results show that PDM is clearly preferred in terms of quality over the other systems.

Index Terms— Sinusoidal Model, Critical band, Vocoder

1. INTRODUCTION

Vocoders have been usefully applied in a number of applications, such as low-bit rate speech coding, analysis, speech synthesis and for speech modification. Vocoders extract parameters from speech and then use them (or some modified form) to reconstruct speech. Current parametric synthesis methods are mainly based on the source-filter theory, where the source excitation is represented by a mixture of pulse train and white noise. More sophisticated models have been proposed to improve quality, such as STRAIGHT [1], which decomposes signals into spectral envelope, excitation and aperiodicity parameters. To reduce the number of parameters when using STRAIGHT, researchers [2] have suggested the use of an intermediate parameterisation, such as the Mel cepstrum [3] with critical band excitation (STRMCEP) to represent a given speech frame. An alternative category of vocoders represents speech as a sum of sinusoids [4]. This approach allows us to modify many speech characteristics, such as timbre and duration. Multiple sinusoidal models have been proposed, for example, the Harmonic pulse Noise Model (HNM) and time-varying models, such as the Quasi-Harmonic Model (QHM) [5] and the adaptive Quasi-Harmonic Model (aQHM) [6].

In [7], a set of source-filter vocoders is experimentally compared with various sinusoidal vocoders (e.g. adaptive Harmonic Model (aHM) [8] and Harmonic Model (HM)). Both objective error measures and preference listening tests show that aHM and HM are preferred to the source-filter vocoders in terms of quality. However, the number of parameters used in these sinusoidal vocoders is much higher than in the source-filter models, and moreover the varying number of parameters in each frame also constrains their further application [9]. Crucially, for example, both these factors make it difficult to use sinusoidal vocoders for statistical speech synthesis. To address the need for an analysis/synthesis method which can provide high quality with a fixed and low number of parameters, we propose a new perceptually based dynamic sinusoidal model (PDM). At this initial stage, we focus on evaluating the proposed model in a copy synthesis experiment. Listening test results show that in the same number of parameters, the suggested model is preferred to the standard sinusoidal model and STRMCEP.

This paper is organised as follows. Section 2 introduces the baseline sinusoidal model. In Section 3, we discuss how we address the issues mentioned above step by step, in order to develop a vocoder with the desired characteristics. In Section 4, results of experiments are presented to support our proposal for using both critical bands and dynamic features. Comparisons with state-of-the-art systems are also provided. Finally, we conclude our paper in Section 5.

2. SINUSOIDAL MODEL

Many acoustic signals, and the human voice and music in particular, can be efficiently modelled as a sum of sinusoids. Furthermore, research in the field of psychoacoustics shows

it is reasonable to decompose sounds into sums of sinusoids [10]. The first, "standard" sinusoidal model (SM) [4] used non-harmonically related sinusoids with amplitude, phase and frequency parameters to represent speech. The number of sinusoids per frame could either be fixed or related to the value of pitch: $K(n) = \frac{F_s/2}{F_0(n)}$ (F_s: sampling frequency, F_0: time-varying pitch for harmonic models). Parameters θ_k, A_k and ω_k represent the phase, amplitude and frequency of the kth sinusoid respectively. As $A_k e^{j\theta_k}$ is invariant, it is possible to model speech as:

$$s(n) = \sum_{k=-K(n)}^{K(n)} A_k e^{j\theta_k} e^{j\omega_k n} = \sum_{k=-K(n)}^{K(n)} a_k e^{j\omega_k n} \quad (1)$$

Complex amplitudes a_k ($a_{-k} = \bar{a_k}$) can be estimated by peak picking or solving a least squares problem [11]. Following the latter approach, these parameters are computed for windowed frames by minimizing the error between the speech model s(n) and the original speech h(n) as shown in (2), where $w(n)$ is the analysis window for each frame and N is half of window length.

$$\epsilon = \sum_{n=-N}^{N} w^2(n)(s(n) - h(n))^2 \quad (2)$$

3. PERCEPTUALLY BASED DYNAMIC SINUSOIDAL MODEL (PDM)

Taking the basic model in Section 2, the approach we took to develop the PDM was first to decrease and fix the number of sinusoids in each frame according to knowledge of human perception. Specifically, for a wideband speech signal (0Hz ~ 8kHz), we assume 21 critical bands [12], and for each of these critical bands only one sinusoid is used. However, we found limiting the number of parameters in this way had some negative effects on speech quality which subsequently needed to be resolved. First, there is a general degradation in signal quality due to the parsimonious representation. Second, we found the resynthesised speech to have an attenuated, or "muffled", quality. Third, we observed a perceptual distortion which is best described as a "tube effect" (resynthesised speech sounds as though it has been spoken through a tube). In the rest of this section, we discuss the steps and issues involved in the development of the PDM in more depth.

3.1. Decreasing and fixing the number of parameters

From (1), we can see that the dimensionality of the sinusoidal components in each frame is high (i.e., with F_0=100, F_s=16k, 80 complex amplitudes would result), and it varies depending on F_0. In human perception, the range of sound sensitivity is broad: the auditory system is more sensitive to lower frequencies than to the higher frequencies. Furthermore, a range of

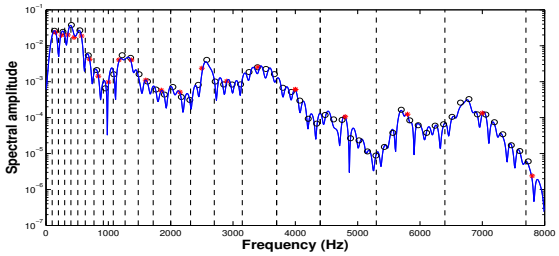

Fig. 1. *Speech magnitude spectrum (blue) along with the critical band boundaries (dashed lines). Estimated amplitudes at the centre of the critical bands (red stars) and harmonic amplitudes (black circles).*

frequencies may be perceived as the same, as they activate the same area on the basilar membrane [13]. In principle, therefore, we can ignore many redundant sinusoids and still retain the perceptually salient characteristics of a speech signal.

In order to distinguish the smallest frequency difference that a listener could perceive, we adopted a perceptual sinusoidal model (PM) based on critical bands [14] in order to decrease and fix the number of parameters. The whole frequency band is divided into 21 critical bands [12]. Instead of using all harmonic components, only 21 sinusoids at the frequencies of critical band centres are used to represent speech, as illustrated in Fig. 1. The PM function is defined as (3). ω_k^c and a_k^c represent the frequency of the critical centre and corresponding estimated complex amplitude. In order to demonstrate the effectiveness of critical bands, as part of the evaluation presented in Section 4, we have compared it with equivalent systems using linear and Mel frequency scales (LM and MM respectively). An informal pilot listening test conducted during the development of PDM indicated that using only one sinusoidal component in each critical band was preferred to using linear and Mel frequency scales.

$$s_{cen}(n) = \sum_{k=-21}^{21} a_k^c e^{j\omega_k^c n} \quad (3)$$

3.2. Integrating dynamic features for sinusoids

However, the test also indicated that the quality of the reconstructed speech was not satisfactory. To address this problem, we have introduced dynamic features for each sinusoid, similar to the method of [5]. The new model is referred to as the *perceptual dynamic sinusoidal model* (PDM):

$$s_{cen}(n) = \sum_{k=-21}^{21} (a_k^c + n b_k^c) e^{j\omega_k^c n} \quad (4)$$

where a_k^c and b_k^c represent the static amplitude and dynamic slope respectively while ω_k^c is the centre frequency of each critical band. The parameters are computed in a similar

way as (2). Hence, PDM has twice as many parameters compared with PM. Although the slope parameter performs a different role to a static amplitude, we want to further compare the quality of samples generated from PDM with the ones from PM with an equal number of parameters. So, by dividing every original critical band into half, another version of PM with doubled critical band frequencies ($s_{cen}(n) = \sum_{k=-42}^{42} \tilde{a}_k^c e^{j\tilde{\omega}_k^c n}$) is implemented. Comparisons between PM and PDM will be presented in Section 4.

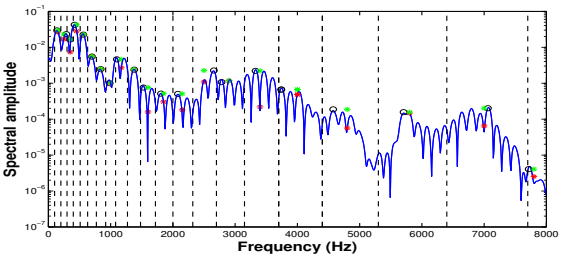

Fig. 2. *Speech magnitude spectrum (blue) along with the critical bands boundaries (dashed lines). Estimated amplitudes at the centre of the critical bands (red stars), and maximum amplitudes in each band (black circles). Green stars denote the sinusoids with the maximum amplitude per critical band as moved at the central frequency of each critical band.*

3.3. Maximum band energy

In Sections 3.1 and 3.2, we have proposed a fixed- and low-dimensional perceptional sinusoidal model to represent speech, based on 21 critical bands with dynamic features. However, such a parameterisation sounds muffled. In Fig. 2, the sinusoid corresponding to the centre of the critical bands are shown with red crosses, while the sinusoid with the maximum amplitude in each band is shown with a black circle. From this example, it is easily seen that the critical band centre sinusoids frequently have a lower amplitude, which may lead to loss of the energy of the signal.

Here, instead of using the critical centre component, for each band, we propose to compute the sinusoidal component which has the maximum spectral amplitude (black circles), and then substitute the initial frequency of the sinusoid with the centre frequency of the critical band (green stars). Peak picking is used to identify which sinusoid has the highest amplitude in each band. Doing this, most of the energy of the signal is modeled. The new suggested system is defined in (5), where a_k^{max} and b_k^{max} represent the static amplitude and dynamic slope for the sinusoid with the maximum spectral amplitude in each critical band, and w_k^c is the centre frequency of critical band.

$$s_{max}(n) = \sum_{k=-21}^{21} (a_k^{max} + n b_k^{max}) e^{j\omega_k^c n} \qquad (5)$$

3.4. Perceived distortion ("tube effect")

The muffled sound is much improved in the form of PDM described in Section 3.3. However, in Fig. 2, we can see there are only 4 sinusoidal components above 4kHz. Due to this decreased number of sinusoidal components for the higher frequency range, we have found that the generated samples sound as they have been spoken from a tube (the "tube effect") with some frequencies being removed completely. This is especially critical for fricative sounds. As the critical bands become very sparse in the higher frequency range, more sinusoidal components are required to compensate the loss of quality in these bands.

Based on the fact that human auditory system is not very selective at high frequencies as in the low frequencies, a time and frequency domain modulated noise $s_H(n)$, covering the high frequencies, is added to the model. For this purpose, a random sinusoidal signal is obtained with amplitudes obtained at every 100 Hz through interpolation of the amplitudes estimated at the high frequency bands (i.e., $a_k(max)$, $k = 18, ..., 21$), and with random phase. No dynamic features are used for this random signal. This signal is further modulated over time by the time-domain envelope (estimated through the Hilbert Transform) from the sinusoidal signal made by the highest 4 sinusoidal components in (5).

At low frequencies, a strong sinusoidal component at the center of a critical band will mask all the other sinusoidal components in that band. The masking threshold is highest at each critical band center and lowest at the boundaries. Therefore, the masking effect will not be as strong at the boundaries of the critical bands [14]. This implies the sinusoids at the the critical band boundaries can potentially affect perception. Accordingly, we chose to add another 9 sinusoidal components at the lower critical band boundaries.

$$s_L(n) = \sum_{k=-9}^{9} (a_k^{bo} + n b_k^{bo}) e^{j\omega_k^{bo} n}; w_k \leq 4kHz \qquad (6)$$

where a_k^{bo}, b_k^{bo} and ω_k^{bo} represent static amplitudes, dynamic slopes and frequencies for 9 sinusoids at the critical boundaries. Finally, the suggested PDM is composed by the sum of the above 3 components:

$$s(n) = s_{max}(n) + s_L(n) + s_H(n) \qquad (7)$$

4. EVALUATION

Phonetically balanced speech data from 3 male and 4 female English speakers was selected for testing. Five neutral speaking sentences were selected for each speaker, with 16kHz sampling rate. We used a reference implementation of each of the models to create stimuli using copy synthesis. The frame shift was set to 5ms with a window length of 20ms for all the methods. Several generated samples are available online at

Table 1. *Parameters and dimensions used in the 3 systems*

Name	Model	Dimensionality: Parameters
PDM	Perceptual dynamic sinusoidal model	120: (30 static + 30 slope)*(real + imaginary)
STRMCEP	STRAIGHT Mel cepstral with band excitation	123: 100 Mel cepstrum + 22 aperiodicity + F0
SM	Sinusoidal model with 40 maximum sinusoids	120: 40 frequency + 40 amplitude + 40 phase

Table 2. *Objective quality for LM, MM and PM*

System	Frequency	PESQ
LM	Linear band	2.5961
MM	Mel band	2.8595
PM	Critical band	3.2183

http://homepages.inf.ed.ac.uk/s1164800/PDM.html

The first experiment aims to compare the use of critical bands (PM) with Mel frequency (MM) and linear frequency scales (LM). Perceptual Evaluation of Speech Quality (PESQ) [15] is calculated as an objective error measure. The average values of all 35 sentences of the seven speakers are listed in Table 2. The increased PESQ value of PM shows that the sinusoidal model based on critical bands produces higher quality than those based on Mel and linear frequency scales. This was also confirmed with informal listening tests.

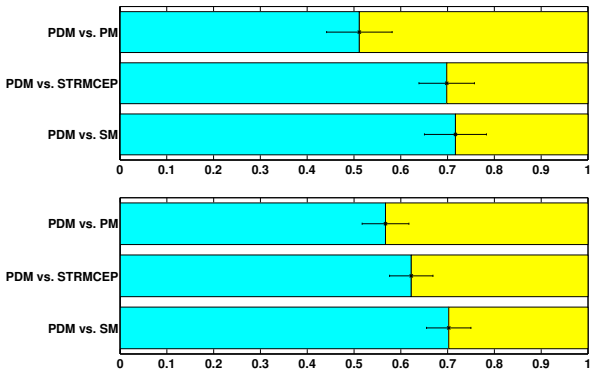

Fig. 3. *Preference result with 95% confidence interval (Top: online test; Bottom: lab-based test)*

Next, we are interested in how the suggested PDM performs compared to other state-of-the-art models, and specifically when the same number of parameters are used with each model. As STRMCEP and the standard sinusoidal model are the two popular models which give high quality of reconstructed speech, a preference listening test was conducted to compare these three models. Details concerning the parameters used in each model is given in Table 1.

Two groups of subjects were tested separately: 24 listeners participated in a pilot web-based experiment ("online"), and then 30 native English speakers took the test in sound-treated perceptual testing booths ("lab-based"). In the listening test, we also compared PDM and PM with the same number of parameters in order to investigate the effectiveness of the dynamic features. From Fig. 3, we see that the online and lab-based results are consistent with each other. Little or no preference is shown between PDM and PM, though PDM uses only half the number of critical bands compared to PM. It also shows that with an equal number of parameters, PDM is clearly preferred compared with the other two state-of-the-art systems. Regarding PDM and PM, we notice that in a well-controlled environment (i.e. sound booths, headphones), PDM is preferred over PM. Moreover, the slope features estimated from the signal offer a natural way to model the dynamic aspects of speech. Therefore, we ultimately favour PDM over PM.

5. CONCLUSION

This paper has presented a perceptual dynamic sinusoidal model based on critical bands (PDM) for representing speech. Initially, only one sinusoidal component is used in each critical band, and objective results show that this parametrization is more effective than using Mel and linear frequency scales. For each band, the sinusoid with the maximum spectrum amplitude is selected and its frequency is associated with the centre frequency of the critical band. Dynamic features (complex slopes) are further integrated, and are found to improve quality in the same way as doubling the number of critical bands in PM. Frequency and time-domain envelope modulation of a noise component at higher frequencies and adding sinusoidal components at the critical boundaries for lower frequencies are also considered in an effort to remove what we refer to as a "tube effect". Compared with STRMCEP and standard SM, our listening test shows PDM is preferred in terms of the quality of the reconstructed signal over the other models when using the same number of parameters. In the future, we plan to reduce the number of parameters of the model further and to apply it in the context of parametric speech synthesis.

6. ACKNOWLEDGEMENTS

This research is supported by Toshiba. The authors thank all the participants of the listening test.

7. REFERENCES

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

[2] H. Zen, T. Tomoki, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE transactions on information and systems*, vol. 90, no. 1, pp. 325–333, 2007.

[3] S. Imai, "Cepstral analysis synthesis on the Mel frequency scale," in *Proc. ICASSP*. IEEE, 1983, vol. 8, pp. 93–96.

[4] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[5] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM–FM signal decomposition with application to speech analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 290–300, 2011.

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

[7] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, "An experimental comparison of multiple vocoder types," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, August 2013, pp. 155–160.

[8] G. Degottex and Y. Stylianou, "A full-band adaptive harmonic representation of speech.," in *Proc. Interspeech*, 2012.

[9] D. Erro, I. Sainz, I. Saratxaga, E. Navas, and I. Hernáez, "MFCC+ F0 extraction and waveform reconstruction using HNM: preliminary results in an HMM-based synthesizer," in *Proc. FALA*, 2010, pp. 29–32.

[10] T. Hirvonen and A. Mouchtaris, "Top-down strategies in parameter selection of sinusoidal modeling of audio," in *Proc. ICASSP*, 2010, pp. 273–276.

[11] Y. Stylianou, *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*, Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications, 1996.

[12] J. O. Smith III and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[13] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[14] P. Noll, "Wideband speech and audio coding," *Communications Magazine, IEEE*, vol. 31, no. 11, pp. 34–44, 1993.

[15] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part ii: psychoacoustic model," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, 2002.