An investigation of the application of dynamic sinusoidal models to statistical parametric speech synthesis

Qiong Hu[1], *Yannis Stylianou*[2], *Ranniery Maia*[2], *Korin Richmond*[1], *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, ranniery.maia@crl.toshiba.co.uk
korin@inf.ed.ac.uk, jyamagis@inf.ed.ac.uk, javier.latorre@crl.toshiba.co.uk

Abstract

This paper applies a dynamic sinusoidal synthesis model to statistical parametric speech synthesis (HTS). For this, we utilise regularised cepstral coefficients to represent both the static amplitude and dynamic slope of selected sinusoids for statistical modelling. During synthesis, a dynamic sinusoidal model is used to reconstruct speech. A preference test is conducted to compare the selection of different sinusoids for cepstral representation. Our results show that when integrated with HTS, a relatively small number of sinusoids selected according to a perceptual criterion can produce quality comparable to using all harmonics. A Mean Opinion Score (MOS) test shows that our proposed statistical system is preferred to one using mel-cepstra from pitch synchronous spectral analysis.

Index Terms: dynamic sinusoidal model, human perception, statistical parametric speech synthesis

1. Introduction

The prominence of hidden Markov model (HMM) based speech synthesis has grown rapidly over the past decade, driven by its advantages in terms of flexibility [1], statistical modelling [2] and small footprint [3]. However, compared with concatenative speech synthesis [4], the quality of speech generated by statistical speech synthesis (e.g. HTS) [5] is still not satisfactory. In the statistical parametric approach, acoustic features are first extracted from speech and modelled. Then, the trained models are used to generate novel parameter sequences, typically according to a maximum likelihood criterion, from which synthetic speech can be reconstructed. Thus, the parameterisation and reconstruction process can have a large impact on overall system performance. Current parametric synthesis methods used in HTS are mainly based on the source-filter theory, whereby the source excitation is represented as a combination of pulse train and white noise. A number of sophisticated source-filter vocoders [6, 7, 8, 9, 10] have been proposed to improve the quality of the generated speech.

Many acoustic signals, and in particular the human voice, can be modelled effectively as a sum of sinusoids. This principle forms the basis for an alternative type of vocoder. Multiple variants have been proposed, for example, the Harmonic [11], Quasi-Harmonic [12] and adaptive Quasi-Harmonic [13] models. In [14], to explore differing vocoder characteristics, multiple source filter vocoders [15, 8, 6] were experimentally compared with sinusoidal ones [13, 11, 16]. Both objective measures and listening tests showed that sinusoidal models were preferred in terms of quality. Unfortunately, however, the dimensionality of sinusoidal models (i.e. number of sinusoids) is higher than typical source-filter ones and varies from frame to frame. To address this problem, a perceptual dynamic sinusoidal model (PDM) [17] has been proposed to generate high quality speech with a fixed and low number of parameters.

Note, however, that all the comparisons in [14] were based on copy synthesis of natural speech without statistical modelling. Although sinusoidal models are widely found in speech coding, where they allow us to modify many speech characteristics such as timbre and duration for example, component sinusoids may be highly correlated with each other, and are also dependent upon pitch. This means they are not suited for direct integration within HTS. One approach to make the sinusoidal model more compatible with statistical modelling is to use an intermediate spectral parameterisation. In [18], the harmonics of a log-amplitude spectrum from Fourier analysis are used to calculate the regularized discrete cepstrum [11] to be used for modelling. The sinusoidal model is then used to reconstruct speech by using harmonics computed from the generated cepstral coefficients. In [16], a harmonic/stochastic waveform generator is presented. The complete spectral envelope is obtained by interpolating the amplitudes at each harmonic point. Then, mel-cepstral coefficients are computed from the interpolated spectral envelope. Both these papers show that sinusoidal models are a promising candidate for improving the overall quality of synthetic speech.

In addition, [17] has shown that incorporating the dynamic slope of sinusoids can greatly improve quality in copy synthesis. It is natural, therefore, to consider including this dynamic feature for statistical modelling too. This dynamic information, though, cannot be obtained by the traditional Fourier analysis used in [18]. Therefore, in order to integrate the proposed dynamic sinusoidal model into HTS, we propose to apply a least square error criterion to calculate the static amplitude and dynamic slope for each sinusoid. Both of these are subsequently transformed into a discrete cepstral representation for modelling. Then, least squared error is again used for calculating the cepstra with a regularisation term on a warped scale. Since intermediate parameters are used in HTS modelling instead of using the sinusoid parameters directly, information compression is not important. Hence, all harmonics can be used to compute cepstra and to resynthesise speech. In addition, however, in order to explore the degree of voice degradation by using a sparse representation of sinusoids compared with using all harmonics for calculating cepstra, a fixed- and low dimensional sinusoidal model based on a perceptual criterion [17] is also investigated. At this initial stage, only minimum phase resynthesis is employed.

The paper is organised as follows. Section 2 introduces the

dynamic sinusoidal model and parameter calculations. In Section 3, we discuss how to transfer the sinusoid amplitudes and slopes to a cepstral representation. In Section 4, results of experiments are presented and analysed to show the potential to use the dynamic sinusoidal model for statistical speech synthesis. Finally, we conclude our paper in Section 5.

2. Dynamic sinusoidal model

The general sinusoidal model (SM) (see 1) decomposes sounds into sums of sinusoids with parameters for amplitude A_k, frequency f_k and phase θ_k. Here a_k is a complex amplitude that contains both phase and amplitude information. $K(n)$ indicates the number of sinusoids in the nth frame.

$$s(n) = \sum_{k=-K(n)}^{K(n)} A_k e^{j\theta_k} e^{j2\pi f_k n} = \sum_{k=-K(n)}^{K(n)} a_k e^{j2\pi f_k n} \quad (1)$$

Based on SM, we add a time-varying term b_k for amplitude refinement [12], resulting in the dynamic sinusoidal model (DSM)

$$s(n) = \sum_{k=-K(n)}^{K(n)} (a_k + nb_k) e^{j2\pi f_k n} \quad (2)$$

where a_k and b_k represent the static complex amplitude and dynamic complex slope respectively. These complex numbers can be estimated by solving a least squares problem [12]. Parameters are computed for windowed frames by minimising the error between the speech model $s(n)$ and the original speech $h(n)$ as shown in (3). $w(n)$ is the analysis window for each frame and N is half the window length. Figure 1 shows the comparison of the natural signal (blue line) with the ones generated by SM and DSM after windowing one frame. We observe that the signal regenerated using DSM (green line) is closer to the natural signal than that of the SM one (red line).

$$\epsilon = \sum_{n=-N}^{N} w^2(n)(s(n) - h(n))^2 \quad (3)$$

When f_k are located at multiples of the fundamental frequency ($f_k = k * f_0$), the dynamic sinusoidal model becomes the harmonic dynamic model (HDM), and the number of sinusoids K varies in each frame depending on pitch. To fix and decrease the number of sinusoids per frame, a perceptual dynamic model (PDM) [17] has been proposed, where the distribution of sinusoids is more concentrated at lower frequencies and gradually becomes more sparse at higher frequencies based on the critical band criterion. For PDM, f_k represents the centre frequency of each critical band. a_k and b_k represent the amplitude and slope of one sinusoid, which has the highest amplitude in each critical band. Interpolation and modulation are also conducted to improve quality further. The main differences between HDM and PDM are summarised in Table 1.

3. Application of DSM for statistical parametric synthesis

3.1. Analysis

To integrate the dynamic model into the HTS framework, regularized discrete cepstra (RDC) [11] are utilized as an intermediate parameterisation for statistical modelling. The amplitudes of static and dynamic sinusoids are first calculated by minimizing (3). Then, we apply the regularized discrete cepstrum to

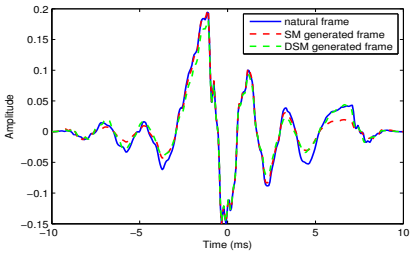

Figure 1: *natural frame (blue), generated (SM: red, DSM: green)*

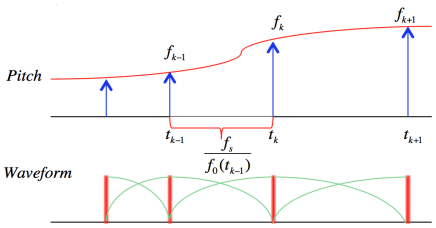

Figure 2: *Overlap-and-add speech synthesis*

parametrize the log amplitude for both static and dynamic sinusoids shown in (4) and (5).

$$log|A(f_k)| = c_0^a + \sum_{i=1}^{P_a} c_i^a cos(2\pi f_k i) \quad (4)$$

$$log|B(f_k)| = c_0^b + \sum_{i=1}^{P_b} c_i^b cos(2\pi f_k i) \quad (5)$$

where c^a, P_a and c^b, P_b represent the RDC and its dimension for both static amplitude and dynamic slope respectively. The cepstral coefficients can be calculated using a least squares error criterion (6) between natural spectrum S_k and estimated spectrum $A(f_k)$ with the regularisation term shown in (7). $R[A(f_k)]$ is applied mainly to ensure a smooth envelope. λ ($4e^{-4}$) is the regularization control parameter [18]. A regularization term is applied for slope computation as well.

$$\varepsilon_a = -\sum_{k=1}^{L} ||20logS_k - logA(f_k)|| + \lambda R[logA(f_k)] \quad (6)$$

$$R[A(f_k)] = 2\pi \int_{-\pi}^{\pi} \left[\frac{d}{d\theta} log|A(\theta)|\right]^2 d\theta \quad (7)$$

L is the number of selected sinusoids for RDC calculation (Dimension of sinusoids: $f_s/2/f_0$ for HDM, and 30 for PDM. f_s: sampling frequency, f_0: pitch). Usually, sinusoids at harmonic frequencies are selected [18, 16] for calculating the cepstra. To improve perceptual quality, frequency warping [19] is used to emphasise accuracy of the spectral envelope at lower frequencies, where human perception is more sensitive. Examples of estimated amplitude envelopes on a Bark scale for both static amplitude and dynamic slope for harmonics are shown in Figure 3. As we see, after warping, though the lower frequency region is enlarged, most selected harmonics are wasted to compute the

Table 1: *Main differences between HDM and PDM (f_s: sampling frequency, f_0: pitch)*

System	sinusoidal frequency	estimated amplitude and phase	number of sinusoids
HDM	harmonics	corresponding sinusoids	$f_s/2/f_0$
PDM	critical band center	sinusoids which have the maximum amplitude in each band	30

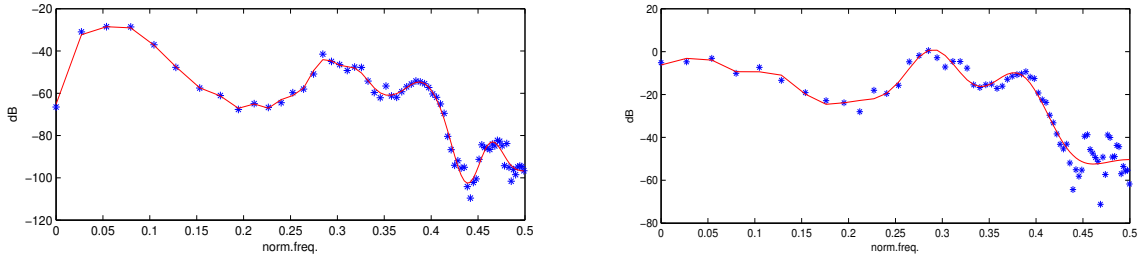

Figure 3: *Estimated log-amplitude envelope with a Bark scale for both static amplitude (left) and dynamic slope (right) from harmonics (blue stars: estimated harmonic amplitude calculated from (3), red lines: re-estimated envelope calculated from RDC)*

envelope of higher frequencies. But for human perception, sinusoids extracted at the higher frequencies tend to be less useful compared to the lower ones.

For the PDM, the sinusoids are selected according to the critical band criterion, where the distribution is more focused on the lower frequencies. Although only 30 sinusoids are used in PDM, which cannot be expected to achieve the same quality as using all harmonics, it may be that comparing HDM and PDM could potentially indicate how much quality the generated speech has lost by using this sparse sinusoidal representation and also the degradation after the statistical modelling. Therefore, we use both HDM and PDM to compute RDC, while using the same model (PDM) for synthesis (referred to as HarPDM and PDMPDM respectively in Table 2). Meanwhile, we also compare these two models for synthesis by keeping the analysis model the same (HarPDM and HarHar respectively).

3.2. Synthesis

For analysis, the speech signal is windowed every 5 ms to compute RDC. Since the residual phase and linear phase terms of the sinusoids are discarded after transforming to RDC, the pitch of each reconstructed frame will not vary if the signal is resynthesized every 5 ms with only the minimum phase (10)(11), which is related with the vocal tract. To put back pitch information, a pitch synchronous overlap-and-add method for synthesis (Figure 2) is used to relocate the center and the length of the synthesis window. For voiced frames, the new pitch marks are estimated at one pitch period distance from the other. Then, we center a window at these pitch marks, and the length of the window is set as pitch-dependent. Supposing pitch for frame $k-1$ is $f_0(t_{k-1})$ and sampling frequency is f_s, the pitch mark for the next frame k would become

$$t_k = t_{k-1} + \frac{f_s}{f_0(t_{k-1})} \tag{8}$$

For unvoiced frames, a dummy f_0 is applied and set as 100 Hz so the calculation is otherwise exactly the same as for voiced frames. Therefore, for synthesis, the dynamic sinusoidal model described in (2) becomes (9), where $|A_k|$, θ_k^a, $|B_k|$, and θ_k^b represent the amplitude and minimum phase for both sinusoidal amplitude and slope respectively. To improve quality, random phase is used for frequencies above 4 kHz.

Table 2: *Systems with different analysis synthesis model*

System	Analysis model	Synthesis model
HarPDM	HDM	PDM
PDMPDM	PDM	PDM
HarHar	HDM	HDM

$$s(n) = \sum_{k=-K(n)}^{K(n)} (|A_k|e^{j\theta_k^a} + n|B_k|e^{j\theta_k^b})e^{j2\pi f_k n} \tag{9}$$

$$\theta^a(f_k) = -\sum_{i=1}^{P_a} c_i^a sin(2\pi f_k i) \tag{10}$$

$$\theta^b(f_k) = -\sum_{i=1}^{P_b} c_i^b sin(2\pi f_k i) \tag{11}$$

4. Experiment

4.1. System configuration

A standard open database *mngu0* [20] containing 2836 sentences, spoken by a male British speaker is utilized to train the statistical parametric speech synthesiser. The sampling frequency is 16 kHz. The HMM based speech synthesis toolkit [5] is used for training multi-stream models. HTS models the acoustic features generated from the vocoders by context-dependent 5-state left-to-right no-skip HSMMs [21]. During synthesis, the parameter generation algorithm [22] considering global variance [23] is used to obtain both spectral and excitation parameters. 50 sentences are randomly selected and excluded from the training set for testing. The pitch synchronous spectral analysis with 40 mel-cepstral coefficients is used as a baseline. At synthesis time, the generated cepstra are converted to spectra. Synthesis is then performed with simple excitation in the frequency domain followed by an overlap-and-add procedure. To maintain equivalent dimensionality, the observation vectors of the systems listed in Table 2 are constructed as

- stream 1: 28 warped RDC for sinusoidal static amplitude, its delta and delta-delta.

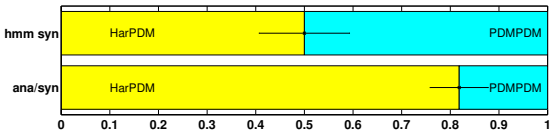

Figure 4: *Preference result comparing analysis models for both analysis/synthesis (bottom) and HMM synthesis (top)*

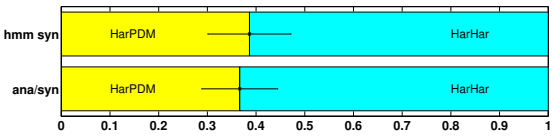

Figure 5: *Preference result comparing synthesis models for both analysis/synthesis (bottom) and HMM synthesis (top)*

- stream 2, 3, 4: log F_0, its delta and delta-delta
- stream 5: 12 warped RDC for sinusoidal dynamic slope, its delta and delta-delta.

Besides testing the statistically generated sentences, we also used a reference implementation of the same 50 sentences to create stimuli using copy synthesis for each model listed in Table 2. 33 subjects participated in the listening test. Several samples included in the test are available online at : http://homepages.inf.ed.ac.uk/s1164800/PDMcepDemo.html

4.2. Listening test

The aim of the first experiment is to compare speech generated using all harmonics on one hand against using sparse sinusoids based on the perceptual criterion in PDM for computing the RDC on the other. A preference test was conducted to compare HarPDM and PDMPDM in Table 2. Figure 4 shows that for analysis/synthesis, HarPDM is preferred to PDMPDM. But with the addition of statistical modelling, there is no statistically significant difference in preference between those two systems, which indicates that the sparse representation of sinusoids based on critical bands can generate comparable quality of speech even if many sinusoids at higher frequencies are not used to compute the RDC. Therefore, we can conclude although using all harmonics could generate higher quality than the sparse representation for analysis/synthesis, people cannot perceive the difference between these two systems after the statistical modelling of the intermediate parameters.

Similarly, a second preference test is conducted to compare these two models for synthesis when using all harmonics for RDC computation (HarPDM and HarHar in Table 2). The number of parameters used for HDM is greater than for PDM. Therefore, using HDM should generate speech with higher quality than the latter one from the same RDC. Results for both analysis/synthesis and HMM synthesis in Figure 5 support this assumption.

Finally, all three models based on HMM synthesis listed in Table 2 are compared with pitch synchronous analysis using mel-cepstra (baseline) by way of a Mean Opinion Score (MOS) test. Subjects are asked to rate the quality of speech on a one-to-five-point scale. As can be seen in Figure 6, all three sinusoid-based models are preferred to the baseline. Specifically, compared with HarPDM and PDMPDM, HarHar is preferred, which is consistent with the results of our previous preference test.

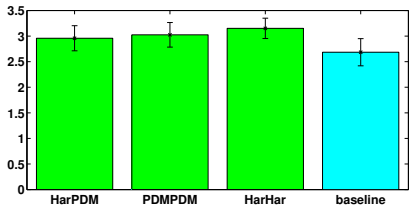

Figure 6: *MOS results for systems based on HMM synthesis*

5. Discussion

To separate the vocal tract filter from the effects of periodic excitation, the mel-cepstrum with pitch synchronous analysis is used as our baseline. Pitch marks are thus needed for the entire database, and the results are very much reliant on their accuracy. From the MOS test, we can see that all three of the proposed systems give better quality than the baseline, and crucially no pitch marks are currently used for them. This is a distinct advantage. Although the use of random phase above 4 kHz may also contribute to a better quality of sinusoidal model, results indicate the proposed approach represents a good candidate for statistical speech synthesis.

In this paper, we also investigate the degradation of voice quality by using a sparse representation of sinusoids (PDM) compared with utilizing all harmonics (HDM) for RDC calculation, as well as the interaction between statistical modelling. HDM demonstrates higher quality compared to PDM by using cepstra as intermediate parameters for analysis/synthesis, but this advantage from using all the harmonics is greatly diminished following the integration of statistical modelling, even though many more sinusoids are used in HDM for computing the RDC. Thus, while the number of sinusoids used in PDM is more limited, it seems this number of sinusoids is sufficient when their distribution is more dense at lower frequencies and more sparse at higher ones, which is compatible with human perception characteristics. Therefore, in the further work, it is also worth to apply the PDM features (low and fixed dimensionality) directly into the statistical models without applying the intermediate parameters.

6. Conclusions

This paper introduces a dynamic sinusoidal model and its implementation, and demonstrates it has been successfully incorporated within an HMM-based speech synthesizer. To apply the dynamic sinusoidal model in HTS, the static amplitudes and dynamic slopes of selected sinusoids are first calculated using least squares error. Then, we utilize regularized discrete cepstra (RDC) to represent both of them for statistical modelling. An MOS test shows that our proposed system is preferred to one using mel-cepstra from pitch synchronous spectral analysis. Furthermore, comparison of different analysis variants (PDM and HDM) for computing RDC is also investigated. Our results show that when integrated with HTS, a relatively small number of sinusoids selected according to a perceptual criterion can produce quality comparable to using all harmonics.

7. Acknowledgements

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

8. References

[1] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP*. IEEE, 2001, vol. 2, pp. 805–808.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," pp. 2347–2350, 1999.

[3] A. Gutkin, X. Gonzalvo, S. Breuer, and P. Taylor, "Quantized HMMs for low footprint text-to-speech synthesis.," in *Proc. Interspeech*, 2010, pp. 837–840.

[4] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*. IEEE, 1996, vol. 1, pp. 373–376.

[5] 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. Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.

[6] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 968–981, 2012.

[7] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling," in *Proc. 6th ISCA Speech Synthesis Workshop*, 2007.

[8] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, 2011.

[9] H. Zen, T. Toda, and T. Keiichi, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," *IEICE Transactions on Information and Systems*, vol. 91, no. 6, pp. 1764–1773, 2008.

[10] J.P. Cabral, K. Richmond, J. Yamagishi, and S. Renals, "Glottal spectral separation for speech synthesis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 2, pp. 195–208, April 2014.

[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] 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.

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

[14] 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.

[15] 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.

[16] 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," *Proc. FALA*, pp. 29–32, 2010.

[17] Q. Hu, Y. Stylianou, K. Richmond, R. Maia, J. Yamagishi, and J. Latorre, "A fixed dimension and perceptually based dynamic sinusoidal model of speech," in *Proc. ICASSP*, Florence, Italy, May 2014.

[18] S. Shechtman and A. Sorin, "Sinusoidal model parameterization for HMM-based TTS system.," in *Proc. Interspeech*, 2010, pp. 805–808.

[19] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U.K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *Journal of the Audio Engineering Society*, vol. 48, no. 11, pp. 1011–1031, 2000.

[20] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus.," in *Proc. Interspeech*, 2011, pp. 1505–1508.

[21] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis.," in *Proc. Interspeech*, 2004.

[22] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*. IEEE, 2000, vol. 3, pp. 1315–1318.

[23] T. Tomoki and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.