Fusion of multiple parameterisations for DNN-based sinusoidal speech synthesis with multi-task learning

Qiong Hu[1], *Zhizheng Wu*[1], *Korin Richmond*[1], *Junichi Yamagishi*[1], *Yannis Stylianou*[2], *Ranniery Maia*[2]

[1]The Centre for Speech Technology Research,University of Edinburgh, UK
[2]Toshiba Research Europe Ltd, Cambridge, UK

Abstract

It has recently been shown that deep neural networks (DNN) can improve the quality of statistical parametric speech synthesis (SPSS) when using a source-filter vocoder. Our own previous work has furthermore shown that a dynamic sinusoidal model (DSM) is also highly suited to DNN-based SPSS, whereby sinusoids may either be used themselves as a "direct parameterisation" (DIR), or they may be encoded using an "intermediate spectral parameterisation" (INT). The approach in that work was effectively to replace a decision tree with a neural network. However, waveform parameterisation and synthesis steps that have been developed to suit HMMs may not fully exploit DNN capabilities. Here, in contrast, we investigate ways to combine INT and DIR at the levels of both DNN modelling and waveform generation. For DNN training, we propose to use multi-task learning to model cepstra (from INT) and log amplitudes (from DIR) as primary and secondary tasks. Our results show combining these improves modelling accuracy for both tasks. Next, during synthesis, instead of discarding parameters from the second task, a fusion method using harmonic amplitudes derived from both tasks is applied. Preference tests show the proposed method gives improved performance, and that this applies to synthesising both with and without global variance parameters.

Index Terms: Fusion vocoder, deep neural network, sinusoidal model, statistical speech synthesis[1]

1. Introduction

Statistical parametric speech synthesis (SPSS) [26] based on hidden Markov models (HMM) has many advantages, such as small footprint, flexibility and robustness. However, the quality of the speech it generates does not yet match that of concatenative speech synthesis. Of key importance [29] are: i) the statistical model used to learn the complex relationship between linguistic input and acoustic output, and ii) the parameterisation of the speech waveform. Ideally, the properties of the statistical model and speech vocoder should complement each other. In this paper, we propose techniques to improve SPSS voice quality by considering both of these factors.

In terms of statistical modelling, deep neural networks (DNN) have been shown to significantly outperform other methods [28, 17, 27]. However, most work so far has used vocoder parameterisations originally designed for HMM-based SPSS. In [1], a hybrid vocoder that switches between PSOLA and sinusoidal model was used for DNN based synthesis, but mel-cepstra and aperiodicity extracted from STRAIGHT [13] spectra were still used for analysis. Focus on source-filter vocoders [5, 13, 18] has persisted despite the fact that sinusoidal ones [6, 21, 4] have been shown to generate equal, if not higher, quality speech [8] with mixed phase. In [10], a dynamic sinusoidal model (DSM), which has additional time-varying components, was proposed. Two ways to use a DSM for SPSS have also been presented (DIR [9] and INT [10]). Results have shown that INT can give quality comparable to the state-of-the art vocoder, while DIR is promising in terms of producing good speech quality without resorting to intermediate parameters such as cepstra.

In [12], we hypothesised the main reason for reduced quality when using the DIR parameterisation was that it does not suit the HMM-based modelling in HTS. The strong correlation between sinusoidal components cannot be modelled by the diagonal covariance in a typical HMM system. In contrast, a DNN does not require acoustic features to be decorrelated [7]. To investigate this, a DSM was integrated with both HMM- and DNN-based synthesisers in [12]. DNNs were found to always outperform their HMM-based equivalent for both methods. In that work, only the decision tree was simply replaced by a single neural network, and speech was ultimately generated from either method individually. But in principle, although DIR and INT constitute different parameterisations of a DSM for use in a statistical model, there are potentially useful connections to be drawn between them:

1) Harmonic amplitudes are transformed to differing types of spectral feature for statistical modelling, but for synthesis harmonic amplitutes need to be recovered again. The harmonic dynamic model (HDM) vocoder is used in both cases.

2) For DIR, cepstra must still be calculated from generated sinusoids after statistical modelling in order to obtain minimum phase. Such cepstra are explicitly retained for modelling in the INT approach.

3) Although cepstra and log amplitudes may in principle be converted to each other by a known matrix (see Section 2.1), we have found the specific ways we derive these parameters mean they can contain complementary spectral information (discussed further in Section 3.1).

On that basis, we are considering to make full use of coefficients trained from both methods by combining them together. In this paper, we fuse the INT and DIR methods at both the modelling and synthesis stages. DNN-based SPSS is highly suited for this, not only because of its ability to model correlated features, but also because it imposes fewer restrictions on feature extraction pipelines. Multi-task learning (MTL) [2] has been proposed to improve the generalisation of a neural network

[1]This work is supported by Toshiba. Email: Qiong.Hu@ed.ac.uk

for tasks such as speech recognition [15, 16], spoken language understanding [24] and natural language processing [3]. Here, we apply MTL for learning spectral representations from both methods. Cepstra (from INT) and log amplitude (from DIR) are trained together as primary and secondary tasks. Normally, output features for the additional task would be discarded after training. But since coefficients trained from both the tasks can be used for generating speech, here we design a novel method to combine both parameterisations at waveform resynthesis time.

2. Methods for DSM parameterisation

In [9], a dynamic sinusoidal model with a time-varying term for amplitude refinement was introduced, under which speech is represented as a sum of static amplitudes a_k and their dynamic slopes b_k, with frequency f_k and phase θ_k:

$$s(n) = \sum_{k=1}^{K} (|a_k| + n|b_k|)cos(2\pi f_k n + \theta_k) \quad (1)$$

Static amplitude $A^{HDM} = [a_1, a_2, ..., a_K]^T$ and dynamic slope $B^{HDM} = [b_1, b_2, ..., b_K]^T$ are calculated using the least squares criterion (LSE) between the original and estimated speech. When sinusoids are located at frequencies of $f_k = k * f_0$ ($k = [1, 2, ..., K]$; K: number of harmonics per frame; f_0 :pitch), the DSM becomes the harmonic dynamic model (HDM).

However, the sinusoidal parameters at every harmonic frequency cannot be modelled directly [9]. Accordingly, two methods have been proposed to apply the DSM for SPSS. In the first method (INT), regularised discrete cepstra (RDC) computed from all harmonic amplitudes are employed as an intermediate parameterisation for statistical modelling. During synthesis, sinusoidal amplitude and minimum phase can be derived as: $log|a_k| = c_0^a + \sum_{i=1}^{P_a} c_i^a cos(2\pi f_k i)$; $\theta_k = -\sum_{i=1}^{P_a} c_i^a sin(2\pi f_k i)$, where c^a, P_a represent the RDC and its dimensionality for the static amplitudes respectively. Assuming W is a diagonal matrix representing the Hanning window and f_s is the sampling frequency, M =[1, $2cos(2\pi \frac{f_1 * 1}{f_s})$, ..., $2cos(2\pi \frac{f_K * 1}{f_s})$; ...; 1, $2cos(2\pi \frac{f_1 * K}{f_s})$; $2cos(2\pi \frac{f_K * K}{f_s})$], RDC ($C_a^{HDM} = [c_1^a, c_2^a, ..., c_{P_a}^a]$) for A^{HDM} can be estimated using LSE [20] between the natural and estimated spectra with a regularisation term R [21]: $C_a^{HDM} = (M^T W M + \lambda R)^{-1} M^T W log|A^{HDM}|$. By replacing $T = (M^T W M + \lambda R)^{-1} M^T W$, the RDC for static amplitude becomes (similar calculation for C_b^{HDM}):

$$C_a^{HDM} = T_a^{HDM} log|A^{HDM}| \quad (2)$$

For the DIR method, we model $log|A|$ and $log|B|$ explicitly. For this, we proposed in [11] a perceptual dynamic model (PDM) with fixed and low dimensionality to satisfy modelling constraints. The sinusoidal component which has the maximum spectral amplitude within each critical band is selected, and then its initial frequency is substituted by the critical band centre frequency ($a_m^{max} = a_m^i = max\{a_m^1, ..., a_m^i, ..., a_m^N\}; b_m^{max} = b_m^i$; N: number of harmonics in band m). The real static log amplitude of $log|A^{PDM}|$ ($A^{PDM} = [a_1^{max}, a_2^{max}, ..., a_M^{max}]$) and slope $log|B^{PDM}|$ ($B^{PDM} = [b_1^{max}, b_2^{max}, ..., b_M^{max}]$, where M is the number of bands) are modelled together with other acoustic features (pitch, voiced/unvoiced flag). During synthesis, HDM is used for generating speech, where amplitudes at each harmonic ($|A^{HDM}|$, $|B^{HDM}|$) are assigned the amplitude of the centre frequency of the critical band in which they lie. Figure 1 gives an overview of both methods for integrating the DSM into DNN-based speech synthesis (see [12] for more detail).

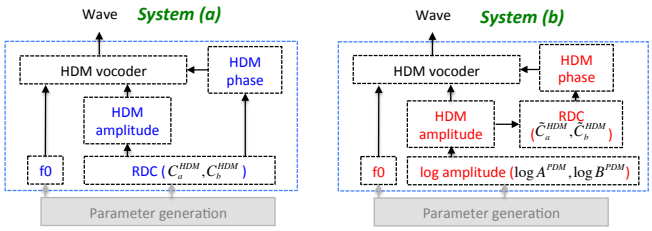

Figure 1: Standard DNN-based speech synthesis for INT (Standard-INT: system (a)) and DIR (Standard-DIR: system (b))

3. Proposed methods

3.1. DNN-based multi-task learning for the DSM

In MTL, extra target outputs associated with additional tasks are added to the original output for training the network. In [25], acoustic features and various secondary features were trained together to improve voice quality, demonstrating that the statistical model can be improved if the second task is chosen well. Specifically, the second task should be related with the primary task, with parameter sharing serving to improve the structure of the model. RDC and log amplitudes can be transformed to each other through matrix T easily (see (2)), so we can combine the INT and DIR methods together using MTL to refine the model. To identify which parameters are suitable for multi-task training, we have tested the potential parameter combinations to represent the DSM listed in Table 1.

Table 1: *Potential parameters for multi-task learning*

INT	DIR
$log\|A^{HDM}\|; log\|B^{HDM}\|$	$log\|A^{PDM}\|; log\|B^{PDM}\|$
$C_a^{HDM} = T_a^{HDM} log\|A^{HDM}\|$ $C_b^{HDM} = T_b^{HDM} log\|B^{HDM}\|$	$C_a^{PDM} = T_a^{PDM} log\|A^{PDM}\|$ $C_b^{PDM} = T_b^{PDM} log\|B^{PDM}\|$

As we can see, for INT, harmonic amplitudes (A^{HDM}, B^{HDM}) have varying dimensionality and cannot be used directly, so C_a^{HDM} and C_b^{HDM} derived from all harmonics are chosen as the first task. For DIR (column 2), in [10], perceptual preference tests show that RDC computed from all harmonics can generate better speech quality than the one (C_a^{PDM}, C_b^{PDM}) calculated from PDM. The steps involved in transforming A^{PDM} and B^{PDM} to cepstra can lead to the loss of spectral detail. Moreover, amplitudes from PDM can also serve as a complementary feature for modelling the spectral parameters, which may not be fully captured in C_a^{HDM} and C_b^{HDM}. Therefore, primary parameters C_a^{HDM} and C_b^{HDM} from INT are augmented to include a second task (A^{PDM} and B^{PDM}) from DIR together with pitch information for multi-task learning. The flow chart of the multi-task learning is shown in Figure 2. During synthesis, speech is resynthesised from "intermediate" and "direct" methods individually.

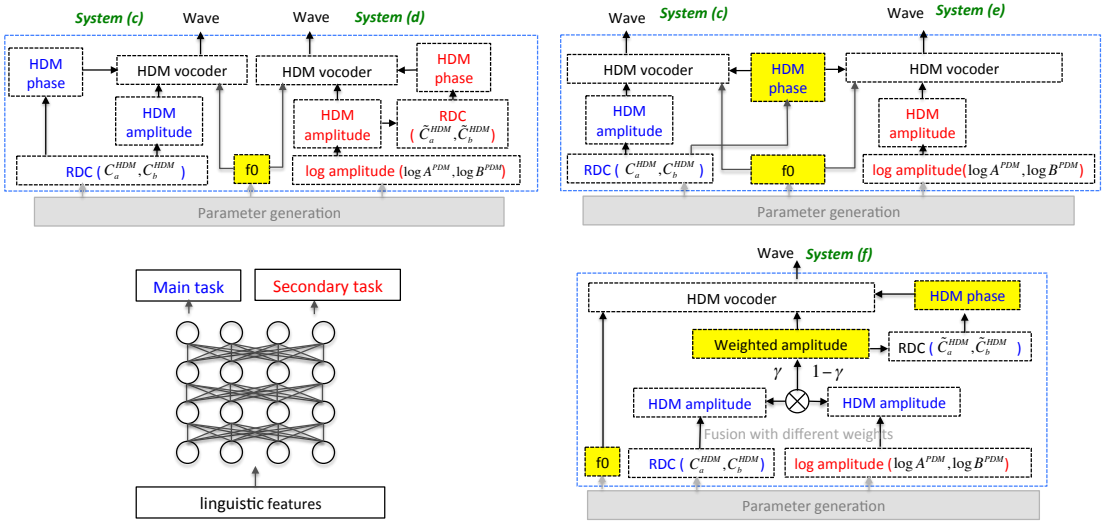

Figure 2: Left: Multi-task learning flowchart for INT (Multi-INT: system (c)) and DIR (Multi-DIR: system (d)); f0 (yellow) is shared by both systems; Right top: Fusion of phase for multi-task learning (Multi-DIR-Phase: system (e)); f_0 and phase are shared (yellow part) by the two systems; Right bottom: Fusion of amplitudes for multi-task learning (Multi-Fusion: system (f))

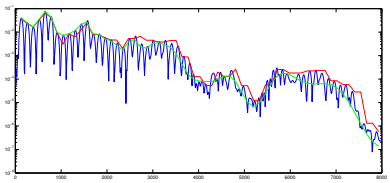

Figure 3: Spectral envelope derived from harmonic amplitude using INT (green) and DIR (red); natural speech FFT (blue).

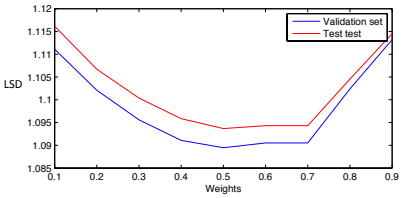

Figure 4: Log-spectral distance (LSD) when using the fusion method with different weightings of INT for both the validation set (blue) and testing set (red)

3.2. Fusion of the multiple parameterisations

Usually, after training the shared tasks, only outputs from the primary task are used and outputs from secondary tasks are discarded. The additional parameters are merely intended to improve optimisation of the trainable network parameters . But for the outputs of MTL in Figure 2, the generated parameters from both tasks can be used for synthesising speech. Although features from INT and DIR are combined for MTL, cepstra and log amplitudes are separated again after parameter generation and then transformed to harmonic amplitudes for synthesis individually. Therefore, there is no interactive combination of features themselves. However, from Figure 2, we can see both the INT and DIR methods use the HDM vocoder for synthesis, with the main difference being how to derive HDM amplitude and phase. In this section, we discuss how to combine the two methods during synthesis stage, focussing on these two aspects.

From systems (c) and (d) in Figure 2, we can see that in order to get HDM phase for synthesising (Section 2.1), RDCs (C_a, C_b) need to be computed first. For INT, C_a^{HDM} and C_b^{HDM} are extracted from all harmonics and explicitly modelled. Meanwhile, for DIR, the generated sparse amplitudes (A^{PDM}, B^{PDM}) need to be extended to harmonic amplitudes first and then transformed to RDC ($\widetilde{C}_a^{HDM}$, $\widetilde{C}_b^{HDM}$) using function (2). However, since more sinusoids are used to calculate C_a^{HDM}, C_b^{HDM} in INT than the $\widetilde{C}_a^{HDM}$ and $\widetilde{C}_b^{HDM}$ used in DIR, $\widetilde{\theta}^{HDM}$ in DIR may not be as accurate as θ^{HDM} derived from INT. Therefore, to test whether this inaccurate phase is the main cause for lower voice quality of DIR (system (d)), we "borrow" phase from INT to use for DIR with the aim of improving the performance of DIR subsequent to the multi-task learning (system (e) in Figure 2).

Figure 3 shows the spectral envelope derived from the harmonic amplitudes using each method. Although perceptual experiments have shown that system (a) can generate better quality than (b) [12], the error between the natural and estimated envelopes varies for different frequencies. Therefore, we propose to combine the two methods by minimising the fused log-spectral distance (LSD). If l^{INT} and l^{DIR} are the LSD for the entire frequency band ($f_s/2$) between generated speech and natural speech using INT and DIR respectively, we can minimise the following objective function by varying the weight γ:

$$l^{Fusion} = \gamma l^{INT} + (1-\gamma) l^{DIR} \tag{3}$$

The optimal weight can be identified by varying γ from 0 to 1 in increments of 0.1. The value which results in the lowest LSD (l^{Fusion}) on the development set will be selected. Figure 4 shows average l^{Fusion} with different weights for the development and test sets used in Section 4. We observe the same trend for both sets. Therefore, weight γ optimised with the development set is used during synthesis in the experiment. To further improve quality, we can extend γ to be a vector $\Upsilon = [\gamma_1, ..., \gamma_i ..., \gamma_M]$, and minimise the LSD for each band used in Section 2.1.

Table 2: *Objective results comparing DNN-based synthesis with and without multi-task learning.*

	INT					DIR				
	C_a^{HDM}	C_b^{HDM}	f_0	V/UV	LSD	$log\|A^{PDM}\|$	$log\|B^{PDM}\|$	f_0	V/UV	LSD
Standard	2.53	5.09	9.51	4.27%	1.13	5.50	8.85	9.41	4.14%	1.15
Multitask	2.40	5.06	9.55	4.13%	1.12	5.35	9.12	9.55	4.13%	1.13

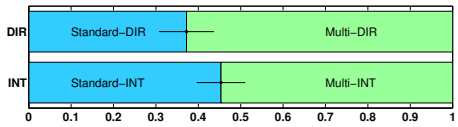

Figure 5: Preference test to demonstrate the effect of multi-task learning for DIR (top) and INT (bottom)

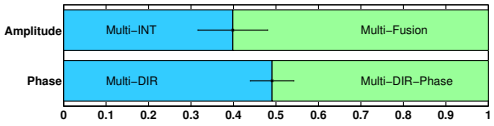

Figure 6: Preference test to investigate the effectiveness of fusion of amplitudes (top) and phase (bottom)

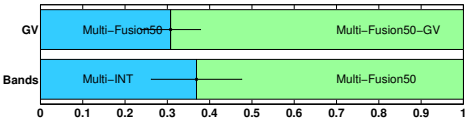

Figure 7: Preference test to investigate the effectiveness of using GV (top) and multiple fusion band weights (bottom)

4. Experiment

Speech data [19] from a British male professional speaker was used to train DNN-based SPSS systems (see [12] for details), either using the HDM with RDC as intermediate parameters (INT) or PDM direct modelling (DIR). The database [19] consists of 2400 utterances for training, 70 for development, and 72 utterances for testing, recorded with a sample rate of 16kHz. The tanh and linear activation functions were used for DNN hidden and output layers respectively. Stochastic gradient descent optimisation was used, with a mini-batch size of 256 and a maximum number of epochs set to 25. Six hidden layers were used, with 1024 units per layer. A momentum term of 0.3 and learning rate of 0.002 were set for the first 10 epochs, and then momentum was increased to 0.9 with a learning rate halved at each epoch. The learning rate for the final two layers was set to half that of other layers. For synthesis, the maximum likelihood parameter generation algorithm (MLPG) [23] was used to obtain both spectral coefficients and excitation. 25 native English subjects participated, listening in sound-treated perceptual booths with headphones. Generated samples are available online[2].

First, we tested the effect of multi-task learning. Objective measures calculated for the test set were compared, as shown in Table 2. Since incompatible spectral parameters were used for analysis and synthesis, LSD computed from the synthesised waveform was compared. In addition, cepstrum error was measured using Mel-cepstral distortion (MCD) [14] for INT, while RMS error of sinusoid log amplitude was used for DIR. We can see that most error rates were improved by multi-task training. Figure 5 shows preference test results between Standard-INT (a) and Multi-INT (c), and Standard-DIR (b) and Multi-DIR (d). We can see that for both methods, systems with MTL were preferred compared with the non-MTL equivalents. This indicates features derived using INT and DIR complement each other and so refine the acoustic model. Specifically, we find increased performance is especially evident for DIR.

To evaluate the fusion of phase, a preference test was conducted to compare Multi-DIR (d) and Multi-DIR-Phase (e) with phase "borrowed" from Multi-INT. Figure 6 shows there was no clear preference between these two systems. From this we conclude that phase ($\widetilde{\theta}^{HDM}$) recovered from the sparse amplitudes is no worse than that computed from RDC using all harmonics. To test the effectiveness of fusing harmonic sinusoid amplitudes, we compared systems using function (3) for one band (Multi-Fusion) and multiple bands (Multi-Fusion50) respectively with Multi-INT (c), which gave the best quality in all our systems so far. Figure 6 and 7 show that for both one band and multiple bands, systems using the fusion method can give better performance than the system using only MTL. Finally, in [12] and [9], listening test results showed that while using global variance (GV) [22] was greatly preferred for the INT method, this strong preference dropped for DIR in both HMM and DNN cases. As the fusion method trained from the MTL is a combination of features from both, another preference test was conducted to explore whether GV is still effective for the fusion case. We compared systems with and without GV for the fusion method using multiple bands. The strong preference in Figure 7 shows GV still works for the proposed method.

5. Conclusion

This paper has presented a novel approach to employing sinusoidal vocoders in DNN-based SPSS. The approach combines methods in which sinusoids are either used as a direct parameterisation (DIR) or following conversion to an intermediate spectral parameterisation (INT). Exploiting DNN capabilities, these two methods are fused together at the statistical modelling and synthesis levels. For statistical training, multi-task learning which models cepstra (from INT) and log amplitudes (from DIR) are trained together as primary and secondary tasks. Objective results and preference test show that both tasks contribute to improving modelling accuracy. For synthesis, instead of discarding parameters from the second task, a fusion method using harmonic amplitudes derived from both tasks is applied. Preference tests show the proposed method gives further improved performance, and that this applies to synthesising both with and without global variance parameters. In the future, other objective functions can be selected and optimised for the fusion method to further improve the perceptual result.

[2]http://homepages.inf.ed.ac.uk/s1164800/Fusion15Demo.html

6. References

[1] Y. Agiomyrgiannakis. Vocaine the vocoder and applications in speech synthesis. In *Proc. ICASSP*, 2015.

[2] R. Caruna. Multitask learning: A knowledge-based source of inductive bias. In *Machine Learning: Proceedings of the Tenth International Conference*, 1993.

[3] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[4] 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*, 21(10):2085–2095, 2013.

[5] 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*, 20(3):968–981, 2012.

[6] 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*, 8(2):184–194, 2014.

[7] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. Sainath. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[8] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre. An experimental comparison of multiple vocoder types. In *Pro. 8th SSW*, 2013.

[9] Q. Hu, Y. Stylianou, R. Maia, K. Richmond, and J. Yamagishi. Methods for applying dynamic sinusoidal models to statistical parametric speech synthesis. In *Proc. ICASSP*, 2015.

[10] Q. Hu, Y. Stylianou, R. Maia, K. Richmond, J. Yamagishi, and J. Latorre. An investigation of the application of dynamic sinusoidal models to statistical parametric speech synthesis. In *Proc. Interspeech*, 2014.

[11] 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*, 2014.

[12] Q. Hu, Z. Wu, K. Richmond, J. Yamagishi, Y. Stylianou, R. Maia, S. King, and M. Akamine. Sinusoidal speech synthesis using deep neural networks. *manuscript*, 2015.

[13] 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*, 27(3):187–207, 1999.

[14] J. Kominek, T. Schultz, and A. Black. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *Pro. SLTU*, 2008.

[15] Y. Lu, F. Lu, S. Sehgal, S. Gupta, J. Du, C. Tham, P. Green, and V. Wan. Multitask learning in connectionist speech recognition. In *Proceedings of the Australian International Conference on Speech Science and Technology*, 2004.

[16] S. Parveen and P. Green. Multitask learning in connectionist robust asr using recurrent neural networks. In *INTERSPEECH*, 2003.

[17] Y. Qian, Y. Fan, W. Hu, and F. Soong. On the training aspects of deep neural network for parametric tts synthesis. In *Proc. ICASSP*, 2014.

[18] 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*, 19(1):153–165, 2011.

[19] K. Richmond, P. Hoole, and S. King. *Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus.*, 2011. http://dx.doi.org/10.7488/ds/140.

[20] S. Shechtman and A. Sorin. Sinusoidal model parameterization for HMM-based TTS system. In *Proc. Interspeech*, 2010.

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

[22] T. Toda and K. Tokuda. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, 90(5):816–824, 2007.

[23] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, 2000.

[24] G. Tur. Multitask learning for spoken language understanding. In *ICASSP*, 2006.

[25] Z. Wu, C. Botinhao, O. Watts, and S. King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *Proc. ICASSP*, 2015.

[26] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. Eurospeech*, 1999.

[27] H. Zen and A. Senior. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *Proc. ICASSP*, 2014.

[28] H. Zen, A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. In *Proc. ICASSP*, 2013.

[29] H. Zen, K. Tokuda, and A. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.