METHODS FOR APPLYING DYNAMIC SINUSOIDAL MODELS TO STATISTICAL PARAMETRIC SPEECH SYNTHESIS

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

[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

ABSTRACT

Sinusoidal vocoders can generate high quality speech, but they have not been extensively applied to statistical parametric speech synthesis. This paper presents two ways for using dynamic sinusoidal models for statistical speech synthesis, enabling the sinusoid parameters to be modelled in HMM-based synthesis. In the first method, features extracted from a fixed- and low-dimensional, perception-based dynamic sinusoidal model (PDM) are statistically modelled directly. In the second method, we convert both static amplitude and dynamic slope from all the harmonics of a signal, which we term the Harmonic Dynamic Model (HDM), to intermediate parameters (regularised cepstral coefficients) for modelling. During synthesis, HDM is then used to reconstruct speech. We have compared the voice quality of these two methods to the STRAIGHT cepstrum-based vocoder with mixed excitation in formal listening tests. Our results show that HDM with intermediate parameters can generate comparable quality as STRAIGHT, while PDM direct modelling seems promising in terms of producing good speech quality without resorting to intermediate parameters such as cepstra.

Index Terms— Sinusoidal model, Parametric statistical speech synthesis, Discrete cepstra, Quality

1. INTRODUCTION

The prominence of statistical parametric speech synthesis (SPSS) based on hidden Markov models (HMM) [1] has grown rapidly in recent years, driven by its recognised various advantages over concatenative speech synthesis [2]. However, the quality of the SPSS is still not satisfactory when compared to the best samples from unit-selection synthesis. The process of waveform parameterisation and reconstruction plays a crucial role in SPSS. Therefore, vocoder performance is a key factor which can influence and constrain overall voice quality [3]. Initially, the parametric vocoders used in HTS [4] were mainly based on source-filter theory, with simple periodic pulse-train or white-noise excitation, which generally give a buzzy quality to the generated speech. Subsequently, a range of high quality vocoders [5, 6, 7, 8, 9, 10] have been proposed to alleviate this problem. Most of these focus on using more sophisticated mixed excitation, using certain special trainable parameters for modelling. Specifically, for example, the STRAIGHT vocoder [5] cannot be integrated with HMMs directly because it has a prohibitively large number of parameters. Therefore, [6] proposed to convert those features into mel cepstral coefficients and band aperiodicities in order to use STRAIGHT with HTS.

An alternative category of vocoder which represents speech as a sum of sinusoids, referred to as sinusoidal vocoders, has been widely applied in speech coding, modification and conversion [11]. Multiple sinusoidal models (SM) have been proposed, for example, the Harmonic Model (HM) [11], the Harmonic plus Noise Model (HNM) [11] the Quasi-Harmonic Model (QHM) [12] and the adaptive Quasi-Harmonic Model (aQHM) [13]. In [14], multiple source-filter vocoders [6, 10, 9] were experimentally compared with sinusoidal ones [13, 11, 15]. Both objective measures and listening tests showed that sinusoidal models were preferred in terms of quality. However, the number of parameters used in these models is much higher than in the source-filter models, and this number also varies from frame to frame. These two factors make it difficult to use sinusoidal vocoders for SPSS. Similar to the integration of STRAIGHT with HTS, researchers have proposed to use intermediate parameters [16, 15] calculated from harmonic amplitudes from FFT analysis for statistical modelling, while HM or HNM is used for the analysis and synthesis stages.

In [17], we proposed a dynamic sinusoidal model (DSM) based on a SM with the addition of time-varying components. It has been shown these dynamic features are effective for improving voice quality [18], so it is natural to include them in statistical modelling. Similar to [16, 15], it is possible to convert the parameters from the proposed vocoder into other intermediate parameters for modelling. Since intermediate parameters are used in HTS instead of using sinusoidal features directly, information compression is not important. Hence, both static and dynamic features from all harmonics can be used for cepstrum computing and resynthesising speech.

Besides using intermediate parameters, an alternative approach is to select fixed sequences of parameters from the sinusoidal vocoder according to perceptual criteria in order to make its parameters suitable both for statistical modelling and for spectral representation. Following this approach, a new

This research is supported by Toshiba.

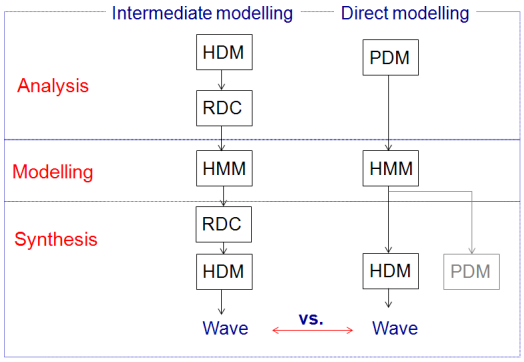

Fig. 1. *Comparison of methods for modelling DSM parameters (Left: Intermediate modelling; Right: Direct modelling)*

"perceptual dynamic sinusoidal model" (PDM) with fixed and low dimensionality based on critical bands was proposed in [18]. Although experiments have shown that using only a limited number of sinusoids can achieve good quality for copy-synthesis, the importance questions are 1) whether the sinusoidal parameters in PDM are capable for modelling; 2) how to synthesise using PDM with minimum phase. 3) how to fit the PDM with complex coefficient for statistical modelling. In this paper, we first propose a dynamic sinusoidal model with real-valued coefficient. Method of how to re-synthesise the signal from the sparse representations of sinusoids with minimum phase is presented and fully explained . Then we present a direct empirical evaluation section for both "direct" and "intermediate" approaches based on HMMs. A summary comparison of the two methods we aim to compare is shown in Fig. 1.

The rest of this paper is organised as follows. The new form of DSM is introduced in Section 2. Then, we discuss how to model sinusoidal parameters for HTS using both intermediate parameters and the sinusoidal parameters directly for HTS in Sections 3 and 4 respectively. In Section 5, our experiment design and listening tests are presented, along with analysis to show the potential of both methods for statistical speech synthesis. Finally, we discuss the outcome and conclude our paper in Section 6.

2. DYNAMIC SINUSOIDAL MODEL (DSM)

The general sinusoidal model (SM) decomposes sounds into sums of sinusoids with parameters for amplitude A_k, frequency f_k and phase θ_k such that

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

Here, a_k is a complex amplitude. $K(n)$ indicates the number of sinusoids in the nth frame. We extend SM to DSM by adding a time-varying term b_k for amplitude refinement:

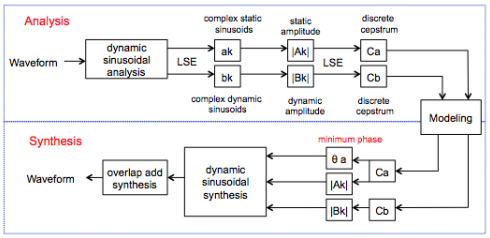

Fig. 2. *The flowchart of sinusoidal analysis (top) and synthesis (bottom) using intermediate parameters for modelling*

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

where a_k and b_k represent the static amplitude and dynamic slope respectively while f_k is the frequency in Hz. Parameters are computed for windowed frames by minimising the error $\varepsilon(n)$ between the speech model $s(n)$ and the original speech $h(n)$, which is calculated as

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

where $w(n)$ is the analysis window for each frame and N is half the window length. When f_k are located at multiples of the fundamental frequency ($f_k = f_0 * k$), the DSM becomes HDM. Note that b_k is the complex slope which reflects the variations of the amplitude and adjustments of its instantaneous frequency [11]. Therefore, DSM with real amplitude A_k, slope B_k and shared phase θ_k is proposed in (4) in this paper. The computation of parameters is similar to (3).

$$s(n) = \sum_{k=1}^{L} (|A_k| + n|B_k|) cos(2\pi f_k n + \theta_k) \quad (4)$$

3. APPLYING DSM TO STATISTICAL PARAMETRIC SYNTHESIS

3.1. Intermediate parameters modelling

To integrate the dynamic model into the HTS framework, regularised discrete cepstra (RDC) [11] are used as an intermediate parameterisation for statistical modelling. The whole process is shown in the flowchart in Fig. 2. In [16], Fourier analysis is applied to calculate the harmonics of a log-amplitude spectrum, which is further converted to RDC for modelling. However, the dynamic amplitude cannot be obtained by the traditional Fourier analysis. Therefore, the least square error criterion (LSE) in Section 2 is used to calculate the static amplitude and dynamic amplitude for each sinusoid. All harmonics are utilised during analysis and synthesis. Then, we apply the RDC to parametrize the log amplitude for static sinusoids such that

$$log|A(f_k)| = c_0^a + \sum_{i=1}^{P_a} c_i^a cos(2\pi f_k i) \quad (5)$$

where c^a, P_a represent the RDC and its dimensionality for the static amplitudes respectively. The cepstral coefficients can be calculated using a least squares error criterion between the natural spectrum S_k and the estimated spectrum $A(f_k)$ with a regularisation term [17]. The computation of RDC for dynamic amplitudes is the same as (5). For analysis, the frame shift is set to 5ms with a four period-long window. For synthesis, the pitch synchronous overlap-and-add method is applied. Real amplitude from static and dynamic sinusoids can be computed from (5), and minimum phase can be derived as

$$\theta(f_k) = -\sum_{i=1}^{P_a} c_i^a sin(2\pi f_k i) \quad (6)$$

3.2. Direct modelling of sinusoidal parameters

Typically, mel-cepstra or line spectral pairs (LSP) are used as parameter vectors to represent spectra. If, in contrast, we wished to avoid these intermediate features, parameters extracted from a sinusoidal vocoder are subject to the following concerns for HTS modelling:

- Speech should be parameterised into fixed-dimensional parameter sequences, but in DSM, the number of sinusoids varies in each frame.
- Increasing the number of observation parameters can enhance performance from HMMs. However, using too many parameters results in data sparsity the models to the training data. But from (2), we can see that the dimensionality of the sinusoidal components in each frame is high (i.e., with F_0=100Hz, F_s=16kHz, 80 sinusoids would result)
- For a typical HMM-based speech synthesis system, diagonal covariance matrices are used, imposing the assumption that individual components in each vector are uncorrelated. However, for harmonics, parameters are highly correlated.

Thus, parameters from classical DSM cannot be directly modelled by HTS. Accordingly, we previously proposed a new type of sinusoidal vocoder referred to as the *perceptual dynamic sinusoidal model* [18] with a fixed number of sinusoids to meet all the requirements mentioned above. The sinusoidal component which has the maximum spectral amplitude at each critical band is selected to represent speech, and then its initial frequency is substituted by the critical band centre.

Although PDM can achieve good quality and meet all the above requirements for a vocoder, it still cannot be directly integrated into HTS. In [18], both a_k and b_k are complex values, containing both amplitude and phase. Amplitude parameters can be directly modelled by HTS. But the phase which is contained in both the static and dynamic sinusoids cannot be modelled, as the distribution of the sinusoids is too sparse to achieve correct phase unwrapping. Therefore, a PDM using (4) with minimum phase is proposed for sinusoidal analysis:

$$s(n) = \sum_{k=1}^{L}((|A_k^{max}| + n|B_k^{max}|)cos(2\pi f_k^{cen} n + \theta_k^{max})) \quad (7)$$

where f_k^{cen} represents each critical band centre. $|A_k^{max}|$, $|B_k^{max}|$ and θ_k^{max} are the static, dynamic amplitude and phase at the sinusoids which have the maximum spectral amplitude in each band. L is the number of selected critical bands. Then, the real log static amplitude $|A_k^{max}|$ and slope $|B_k^{max}|$ are modelled in individual streams to represent spectrum parameters.

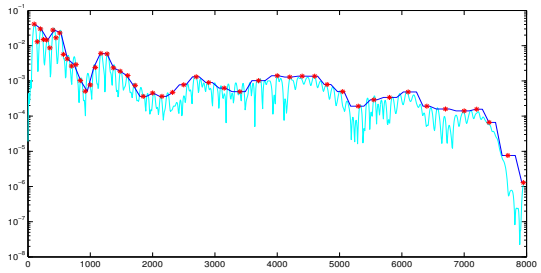

Fig. 3. *Amplitude envelope from PDM (Cyan line: natural spectrum calculated from FFT ; Red point: selected sinusoids $|A_k^{max}|$ at each critical band; Blue line: envelope of the harmonics $|A_k^{har}|$ recovered from $|A_k^{max}|$;)*

From [18], we know that the critical bands become very sparse at higher frequencies. So we increase the number of bands for HTS training, but still very few sinusoids are distributed in this region. In [17], a listening test shows that based on HMM-based synthesis, although HDM and PDM perform almost the same during analysis, using HDM is significantly preferred to using PDM at the synthesis stage. Therefore, after the generation of static and dynamic amplitudes from HTS, instead of using PDM with interpolation, HDM is used to synthesise speech, where amplitudes at each harmonic ($|A_k^{har}|$ and $|B_k^{har}|$) are recovered from the sinusoids of each critical band by putting its value equal to the one at the band center (8). The recovered envelope of all harmonics is shown in Fig. 3.

$$s(n) = \sum_{i=1}^{N}((|A_i^{har}| + n|B_i^{har}|)cos(2\pi i f_0 n + \theta_i^{har})) \quad (8)$$

N is the number harmonics in each frame: $N = f_s/2/f_0$ (f_s: sampling frequency, f_0: time-varying pitch for harmonic models, $A_i^{har} = A_k^{max}(f_{k-1}^{cen} < f_i^{har} \leq f_k^{cen})$; $B_i^{har} = B_k^{max}$). For the phase, θ_i^{har} at each harmonic is derived from the discrete cepstra using (6).

4. EVALUATION

A standard open database related to [19] containing 2992 sentences, spoken by a male British English speaker was

Table 1. *Stream configuration for the three systems tested. Streams include respective delta and delta-delta features.*

	STR (STRAIGHT)	**INT** (Intermediate modelling of parameters from HDM)	**DIR** (Direct modelling of parameters from PDM)
Stream1	50 mel-cepstral coefficients	40 warped RDCs for static amplitude	50 sinusoidal log amplitudes
Stream2,3,4	$\log F_0$ (+ separate Δ and $\Delta\Delta$)	$\log F_0$ (+ separate Δ and $\Delta\Delta$)	$\log F_0$ (+ separate Δ and $\Delta\Delta$)
Stream5	25 aperiodicities (dB)	40 warped RDCs for dynamic amplitude	50 sinusoidal log slope

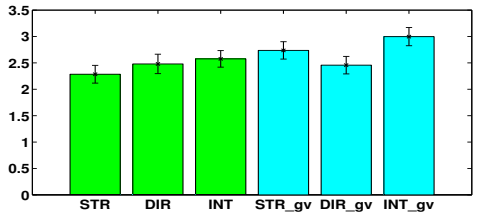

Fig. 4. *MOS results with (blue) and without (green) GV.*

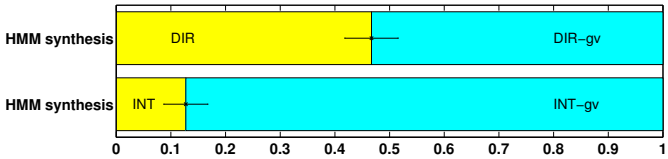

Fig. 5. *Preference test for the performance of GV for both proposed systems*

utilised for our experiments. The sampling frequency is 16 kHz. The HTS HMM-based speech synthesis system [4] was used for training the multi-stream models. Acoustic features were modelled by context-dependent 5-state HSMMs [20]. During synthesis, the parameter generation algorithm [21] both with and without global variance (GV) [22] was used to get both spectral coefficients and excitation. To help gauge system quality, the STRAIGHT cepstrum-based vocoder with mixed excitation [6] was used as a baseline. Each observation vector for the three systems was constructed as detailed in Table 1. Several samples are available on the webpage http://homepages.inf.ed.ac.uk/s1164800/PDMHDMDemo.html

To evaluate quality, 50 testing sentences (chosen randomly and excluded from the training set) were synthesised by the three systems listed in Table 1, using configurations both with and without GV. 30 native English subjects participated in the listening test, conducted in sound-treated perceptual testing booths with headphones. A mean opinion score (MOS) test was used to measure overall quality. Subjects were asked to rate the quality of speech on a one-to-five-point scale. From Fig. 4, we can see for the condition without GV, STR, DIR and INT can generate comparable quality based on HMM synthesis. With the addition of GV modelling, while STR and INT are greatly improved and the performance of INT seems preferred than STR (not statistically significant), there is no quality improvement for DIR.

In order to further confirm the effect of including GV on both proposed systems, another preference test was conducted. The same 30 native listeners participated in this test to give their preference in term of quality. Fig. 5 shows that while INT with GV is strongly preferred, there is no difference between the DIR with GV and the one without GV. Therefore, we can conclude that GV does not improve performance when applied to sinusoidal parameters directly.

5. DISCUSSION

For this male voice, our results show that using intermediate parameters for sinusoidal statistical speech synthesis can achieve comparable quality compared to the state-of-the-art vocoder, which is consistent with [23]. Discrete cepstra converted from the dynamic slope are also trained in the system, which helps improve quality. Noting that the complexity and computation cost of HDM are also less. In our second proposed approach, sinusoidal parameters are trained directly in HTS. Although it can generate relatively good quality speech, we have found classical GV doesn't improve its performance. This is similar to findings with LSPs. We believe that, since sinusoidal parameters are closely tied to the frequency domain, similar to LSPs, our future work should investigate poster-filtering, GV modelling in the frequency domain or minimum generation error as alternatives. Moreover, since information loss can occur during transfer to an intermediate parametrization, and sinusoidal features are more physically meaningful and related with perception, we argue the direct modelling approach still holds significant interest. In future work, different system configurations and more speakers should also be tested.

6. CONCLUSION

This paper focuses on how to apply DSM into a statistical parametric synthesis system. Two strategies for modelling sinusoidal parameters have been compared: converting to an intermediate parametrization or using sinusoidal parameters for training directly. Whereas our previous work focused on copy synthesis, this paper proposed a new representation of sinusoidal parameters and successfully implemented it in TTS by modelling sinusoidal features directly. DSM with real-valued amplitude and slope is proposed. Depending on each approach, different sinusoidal models (HDM/PDM) have been applied during analysis and synthesis. The implementations of HDM from PDM at synthesis stage have also been presented. Our experiments have shown that HDM using intermediate parameters can achieve better quality than the direct sinusoidal feature modelling. Nevertheless, the direct modelling approach still also seems a promising alternative which merits further investigation.

7. REFERENCES

[1] 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.

[2] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, 1996.

[3] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[4] 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. 6th SSW*, 2007.

[5] 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.

[6] 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.

[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 SSW*, 2007.

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

[9] 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.

[10] 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.

[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 *Pro. 8th SSW*, 2013.

[15] 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," 2010.

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

[17] 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.

[18] 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.

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

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

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

[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*, vol. 90, no. 5, pp. 816–824, 2007.

[23] 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*, vol. 8, no. 2, pp. 184–194, 2014.