STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Alan W Black[†] *Heiga Zen*[‡] *Keiichi Tokuda*[‡]

[†]Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA
[‡]Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya, JAPAN
E-mail address: [†]awb@cs.cmu.edu, [‡]{zen,tokuda}@ics.nitech.ac.jp

ABSTRACT

This paper gives a general overview of techniques in **statistical parametric speech synthesis**. One of the instances of these techniques, called **HMM-based generation synthesis** (or simply **HMM-based synthesis**), has recently been shown to be very effective in generating acceptable speech synthesis. This paper also contrasts these techniques with the more conventional unit selection technology that has dominated speech synthesis over the last ten years. Advantages and disadvantages of statistical parametric synthesis are highlighted as well as identifying where we expect the key developments to appear in the immediate future.

Index Terms— Speech synthesis, hidden Markov models

1. BACKGROUND

With the increase in power and resources of computer technology, building natural sounding synthetic voices has progressed from a knowledge-based activity to a data-based one. Rather than handcrafting each phonetic unit and its applicable contexts, high-quality synthetic voices may be built from sufficiently diverse single speaker databases of natural speech. We can see a progression from fixed inventories, found in diphone systems [1] to the more general, but more resource consuming, techniques of unit selection synthesis where appropriate sub-word units are automatically selected from large data-bases of natural speech [2].

ATR ν-talk [3] was the first to show the effectiveness of automatic selection of appropriate units, then CHATR [2] generalized these techniques to multiple languages and an automatic training scheme. Unit selection techniques have risen to be the dominant synthesis technique. The quality of the output derives directly from the quality of the recordings, and it appears that the larger the database the better the coverage. Commercial systems have exploited these technique to bring us a new level of synthetic speech. However, although certainly successful, there is always the issue of spurious errors. When a desired sentence happens to require phonetic and prosody contexts that are under represented in a database, the quality of the synthesizer can be severely degraded. Even though this may be a rare event, a single bad join in an utterance can ruin the listeners flow.

It is not possible to guarantee that bad joins and/or inappropriate units do not occur, simply because of the vast number of possible combinations that could occur. However for particular applications it is often possible to almost always avoid them. Limited domain synthesizers [4], where the database is designed for the particular application, go a long way to making almost all the synthetic output near perfect.

However in spite of the desire for perfect synthesis all the time, there are limitations in the unit selection technique. No (or little) modification of the selected pieces of natural speech are carried out, thus limiting the output speech to the style of that in the original recordings.

With a desire for more control over the speech variation, larger databases containing examples of different styles are required. IBM's stylistic synthesis [5] is a good example but is limited by the amount of variations that can be recorded.

In direct contrast to this selecting of actual instances of speech from a database, **statistical parametric speech synthesis** has also grown in popularity over the last few years. Statistical parametric synthesis might be most simply described as generating the *average* of some set of similarly sounding speech segments. This contrasts directly with the desire in unit selection to keep the natural unmodified speech units, but using parametric models offers other benefits.

In both the Blizzard Challenge 2005 and 2006 ([6, 7]) where a common speech database is provided to participants to build a synthetic voice, the results from listening tests have shown that one of the instances of statistical parametric synthesis techniques called **HMM-based generation synthesis** (or even **HMM-based synthesis**) offers more preferred (through MOS tests) and more understandable (through WER scores) synthesis. Although even the proponents of statistical parametric synthesis feel that the best examples of unit selection are better than the best examples of statistical parametric synthesis, overall it appears that quality of statistical parametric synthesis has already reached a quality that can stand in its own right.

The quality issue really comes down to the fact that given a parametric representation it is necessary to reconstruct the speech from those parameters. The reconstruction process is still not ideal. Although modeling the spectral and prosody features is relatively well defined, models of the residual/excitation are still yet to be fully developed, though composite models like STRAIGHT [8] are proving to be useful.

The following section gives a more formal definition of unit selection techniques that will allow a easier contrast it to statistical parametric synthesis. Then statistical parametric speech synthesis is more formally defined, specifically based on the implementation on the HMM-based speech synthesis system (HTS) [9, 10]. The final sections discuss some of the advantages in a statistical parametric framework highlighting some of the existing a future directions.

2. UNIT SELECTION SYNTHESIS

There seems to be two basic techniques in unit selection, though they are theoretically not very different. Hunt and Black presented a selection model [2], which actually existed previously in ATR ν-talk. The basic notion is that of a **target cost**, how well a candidate unit from the database matches the desired unit, and a **concatenation cost** which defines how well two selected units combine. Unit

selection requires the optimization of both these costs over the utterances.

The definition of target cost between a candidate unit u_i and a desired unit t_i is

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i), \quad (1)$$

where j indexes over all features (typically phonetic and prosodic contexts are used). Concatenation cost is defined as

$$C^c(u_{i-1}, u_i) = \sum_{k=1}^{q} w_k^c C_k^c(u_{i-1}, u_i). \quad (2)$$

Though in this case k may include spectral and acoustic features. Weights (w_j^t and w_k^c) have to be found for each feature, and actually implementations used a combination of trained and hand tuned weights.

The second direction, ([11] and similarly [12]) use a clustering method that allows the target cost to effectively be precalculated. Units of the same type are clustered into a decision tree that asks questions about features available at synthesis time (e.g. phonetic and prosody context).

All of these techniques depend on a **acoustic distance measure** which should be correlated with human perception.

These apparently unit selection specific issues are mentioned here because they have specific counterparts in statistical parametric synthesis.

3. STATISTICAL PARAMETRIC SYNTHESIS

3.1. Overview of a typical system

Figure 1 is a block diagram of a typical HMM-based speech synthesis system [9]. It consists of training and synthesis parts.

The training part is similar to those used in speech recognition systems. The main difference is that both spectrum (e.g., mel-cepstral coefficients [13] and their dynamic features) and excitation (e.g., $\log F_0$ and its dynamic features) parameters are extracted from a speech database and modeled by context-dependent HMMs (phonetic, linguistic, and prosodic contexts are taken into account). To model $\log F_0$ sequence which includes unvoiced regions properly, multi-space probability distributions [14] are used for the state output stream for $\log F_0$. Each HMM has state duration densities to model the temporal structure of speech [15]. As a result, the system models spectrum, excitation, and durations in a unified framework.

The synthesis part does the inverse operation of speech recognition. First, an arbitrarily given text corresponding an utterance to be synthesized is converted to a context-dependent label sequence and then the utterance HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Secondly, state durations of the HMM are determined based on the state duration probability density functions. Thirdly, the speech parameter generation algorithm (typically, case 1 in [16]) generates the sequence of mel-cepstral coefficients and $\log F_0$ values that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated mel-cepstral coefficients and F_0 values using the MLSA filter [17] with binary pulse or noise excitation.

3.2. Advantages and disadvantages

The biggest disadvantage of the HMM-based generation synthesis approach against the unit selection approach is the quality of synthesized speech. There seems to be three factors which degrade the quality: vocoder, modeling accuracy, and over-smoothing.

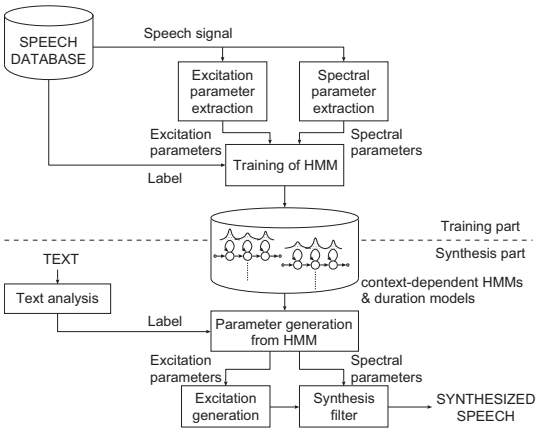

Fig. 1. Overview of a typical HMM-based speech synthesis system.

The synthesized speech by the HMM-based generation synthesis approach sounds *buzzy* since it is based on the vocoding technique. To alleviate this problem, a high quality vocoder such as multi-band excitation scheme [18–21] or STRAIGHT [8] have been integrated. Several groups have recently applied LSP-type parameters instead of mel-cepstral coefficients to the HMM-based generation synthesis approach [22, 23].

The basic system uses ML-estimated HMMs as its acoustic models. Because this system generates speech parameters from its acoustic models, model accuracy highly affects the quality of synthesized speech. To improve its modeling accuracy, a number of advanced acoustic models and training frameworks such as hidden semi-Markov models (HSMMs) [24], trajectory HMMs [25], buried Markov models [26], trended HMMs [27], stochastic Markov graphs [28], minimum generation error (MGE) criterion [29], and variational Bayesian approach [30] have been investigated.

In the basic system, the speech parameter generation algorithm (typically case 1 described by Tokuda et al. [16]) is used to generate spectral and excitation parameters from HMMs. By taking account of constraints between the static and dynamic features, it can generate smooth speech parameter trajectories. However, the generated spectral and excitation parameters are often over-smoothed. Synthesized speech using over-smoothed spectral parameters sounds muffled. To reduce this effect and enhance the speech quality, post-filtering [18, 22], a conditional speech parameter generation algorithm [31], or a speech parameter generation algorithm considering global variance [32] have been used.

Advantages of the HMM-based generation synthesis approach are

1) its voice characteristics can be easily modified,

2) it can be applied to various languages with little modification,

3) a variety of speaking styles or emotional speech can be synthesized using the small amount of speech data,

4) techniques developed in ASR can be easily applied,

5) its footprint is relatively small.

The voice characteristics in 1) can be changed by transforming HMM parameters appropriately because the system generates

speech waveforms from the HMMs themselves. For example, either a speaker adaptation [33, 34], a speaker interpolation [35], or an eigenvoice technique [36] was applied to this system, and it was shown that the system could modify voice characteristics. Multilingual support in 2) can be easily realized because in this system only contextual factors are dependent on each language. Japanese [9], Mandarin [37, 38], Korean [39], English [40], German [41], Portuguese [42, 43], Swedish [44], Finnish [45, 46], Slovenian [47], Croatian [48], Arabic [19], Farsi [49], and Polyglot [50] systems have already been developed by various groups. Speaking styles and emotional voices in 3) can be constructed by re-estimating existing average voice models with only a few utterances using adaptation techniques [51–53]. As for 4), we can employ a number of useful technologies developed for the HMM-based speech recognition. For example, structured precision matrix models, which can approximate full covariance models well using the small number of parameters, have successfully been applied to the system [23]. Small footprints in 5) can be realized by storing statistics of HMMs rather than multi-templates of speech units. For example, footprints of the Nitech's Blizzard Challenge 2005 voices were less than 2 MBytes with no compression [54] .

4. RELATION AND HYBRID APPROACHES

4.1. Relation between two approaches

Some of clustering-based unit selection approaches uses HMM-based state clustering [11]. In this case, the structure is very similar to that of the HMM-based generation synthesis approach. The essential difference between the clustering-based unit-selection approach and the HMM-based generation synthesis approach is that each cluster in the generation approach is represented by statistics of the cluster instead of multi-templates of speech units.

In the HMM-based generation synthesis approach, distributions for spectrum, F_0, and duration are clustered independently. Accordingly, it has different decision trees for each of spectrum, F_0, and duration. On the other hand, unit selection systems often use regression trees (or CART) for prosody prediction. The decision trees for F_0 and duration in the HMM-based generation synthesis approach are essentially equivalent to the regression trees in the unit selection systems. However, in the unit selection systems, leaves of one of trees must have speech waveforms: other trees are used to calculate target costs, to prune waveform candidates, or to give features for constructing the trees for speech waveforms.

It is noted that in the HMM-based generation synthesis approach, likelihoods of static feature parameters and dynamic feature parameters corresponds to the target costs and concatenation costs, respectively. It is easy to understand, if we approximate each state output distribution by a discrete distribution or instances of frame samples in the cluster: when the dynamic feature is calculated as the difference between neighboring static features, the ML-based generation results in a frame-wise DP search like unit selection. Thus HMM-based parameter generation can be viewed as an *analogue version* of unit selection.

4.2. Hybrid approaches

As a natural consequence of the above viewpoints, there are also hybrid approaches.

Some of these approaches use spectrum parameters, F_0 values, and durations (or a part of them) generated from HMM to calculate acoustic target costs for unit selection [55–58]. Similarly, HMM likelihoods are used as "costs" for unit selection [59, 60]. Among of these approaches, [57] and [60] use frame-sized units, and [61] use generated longer trajectories to provide "costs" for unit selection. Another type of hybrid approaches uses statistical models as a probabilistic smoother for unit selection [62, 63]. Unifying unit selection and HMM-based generation synthesis is also investigated [64].

In the future, we may converge at an optimal form of corpus-based speech synthesis fusing generation and selection approaches.

5. CONCLUSION

We can see that statistical parametric speech synthesis offers a wide range of techniques to improve spoken output. Its more complex models, when compared to standard unit selection, allow for general solutions, without necessarily requiring recording speech in all phonetic and prosodic contexts. The pure unit selection view requires very large databases to cover examples of all desired prosodic, phonetic and stylistic variation. In contrast statistical parametric synthesis allows for models to be combined and adapted thus not requiring instances of all possible combinations of contexts.

6. ACKNOWLEDGMENTS

This work was partly supported by the MEXT e-Society project. This work was also partly supported by the US National Science Foundation under grant number 0415021 "SPICE: Speech Processing Interactive Creation and Evaluation Toolkit for new Languages." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] J. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English Speech: A Dynamic Approach*, Springer Verlag, 1993.

[2] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP*, 1996, pp. 373–376.

[3] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "ATR ν-TALK speech synthesis system," in *ICSLP*, 1992, pp. 483–486.

[4] A. Black and K. Lenzo, "Limited domain synthesis," in *ICSLP*, 2000, pp. 411–414.

[5] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to <AHEM/> expressive speech synthesis authors," in *ISCA SSW5*, 2004.

[6] C. Bennett, "Large scale evaluation of corpus-based synthesizers: Results and lessons from the Blizzard Challenge 2005," in *Interspeech*, 2005, pp. 105–108.

[7] C. Bennett and A. Black, "Blizzard Challenge 2006," in *Blizzard Challenge Workshop*, 2006.

[8] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "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, pp. 187–207, 1999.

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

[10] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.W. Black, and T. Nose, "The HMM-based speech synthesis system (HTS)," `http://hts.ics.nitech.ac.jp/`.

[11] R. Donovan and P. Woodland, "Improvements in an HMM-based speech synthesiser," in *Eurospeech*, 1995, pp. 573–576.

[12] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Eurospeech*, 1997, pp. 601–604.

[13] T. Fukada, K. Tokuda, Kobayashi T., and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *ICASSP*, 1992, pp. 137–140.

[14] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *ICSLP*, 1998, pp. 29–32.

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

[17] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *ICASSP 83*, 1983, pp. 93–96.

[18] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Eurospeech*, 2001, pp. 2263–2266.

[19] O. Abdel-Hamid, S. Abdou, and M. Rashwan, "Improving Arabic HMM based speech synthesis quality," in *Interspeech*, 2006, pp. 1332–1335.

[20] C. Hemptinne, *Integration of the harmonic plus noise model into the hidden Markov model-based speech synthesis system*, Master thesis, IDIAP Research Institute, 2006.

[21] S.-J. Kim and M.-S. Hahn, "Two-band excitation for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 378–381, 2007.

[22] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.

[23] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," in *Blizzard Challenge Workshop*, 2006.

[24] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Interspeech*, 2004, pp. 1185–1180.

[25] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in *ISCA SSW5*, 2004.

[26] I. Bulyko, M. Ostendorf, and J. Bilmes, "Robust splicing costs and efficient search with BMM models for concatenative speech synthesis," in *ICASSP*, 2002, pp. 461–464.

[27] J. Dines and S. Sridharan, "Trainable speech synthesis with trended hidden Markov models," in *ICASSP*, 2001, pp. 833–837.

[28] M. Eichner, M. Wolff, S. Ohnewald, and R. Hoffman, "Speech synthesis using stochastic Markov graphs," in *ICASSP*, 2001, pp. 829–832.

[29] Y.-J. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *ICASSP*, 2006, pp. 89–92.

[30] Y. Nankaku, H. Zen, K. Tokuda, T. Kitamura, and T. Masuko, "A Bayesian approach to HMM-based speech synthesis," in *Tech. rep. of IEICE*, 2003, vol. 103, pp. 19–24.

[31] T. Masuko, K. Tokuda, and T. Kobayashi, "A study on conditional parameter generation from HMM based on maximum likelihood criterion," in *Autumn Meeting of ASJ*, 2003, pp. 209–210.

[32] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *Eurospeech*, 2001, pp. 2801–2804.

[33] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *ICASSP*, 1997, pp. 1611–1614.

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

[35] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Eurospeech*, 1997, pp. 2523–2526.

[36] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *ICSLP*, 2002, pp. 1269–1272.

[37] H. Zen, J. Lu, J. Ni, K. Tokuda, and H. Kawai, "HMM-based prosody modeling and synthesis for Japanese and Chinese speech synthesis," Tech. Rep. TR-SLT-0032, ATR-SLT, 2003.

[38] Y. Qian, F. Soong, Y. Chen, and M. Chu, "An HMM-based Mandarin Chinese text-to-speech system," in *ISCSLP*, 2006.

[39] S.-J. Kim, J.-J. Kim, and M.-S. Hahn, "Implementation and evaluation of an HMM-based Korean speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E89-D, pp. 1116–1119, 2006.

[40] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Speech Synthesis Workshop*, 2002.

[41] C. Weiss, R. Maia, K. Tokuda, and W. Hess, "Low resource HMM-based speech synthesis applied to German," in *ESSP*, 2005.

[42] R. Maia, H. Zen, K. Tokuda, T. Kitamura, and F. Resende Jr., "Towards the development of a Brazilian Portuguese text-to-speech system based on HMM," in *Eurospeech*, 2003, pp. 2465–2468.

[43] M. Barros, R. Maia, K. Tokuda, D. Freitas, and F. Resende Jr., "HMM-based European Portuguese speech synthesis," in *Interspeech*, 2005, pp. 2581–2584.

[44] A. Lundgren, *An HMM-based text-to-speech system applied to Swedish*, Master thesis, Royal Institute of Technology (KTH), 2005.

[45] T. Ojala, *Auditory quality evaluation of present Finnish text-to-speech systems*, Master thesis, Helsinki University of Technology, 2006.

[46] M. Vainio, A. Suni, and P. Sirjola, "Developing a Finnish concept-to-speech system," in *2nd Baltic conference on HLT*, 2005, pp. 201–206.

[47] B. Vesnicer and F. Mihelic, "Evaluation of the Slovenian HMM-based speech synthesis system," in *TSD*, 2004, pp. 513–520.

[48] S. Martincic-Ipsic and I. Ipsic, "Croatian HMM-based speech synthesis," *Journal of Computing and Information Technology*, vol. 14, no. 4, pp. 307–313, 2006.

[49] M. Homayounpour and S. Mehdi, "Farsi speech synthesis using hidden Markov model and decision trees," *The CSI Journal on Computer Science and Engineering*, vol. 2, no. 1&3 (a), 2004.

[50] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *ICASSP*, 2005, vol. 1, pp. 1–4.

[51] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Interspeech*, 2003, pp. 2461–2464.

[52] J. Yamagishi, *Average-Voice-Based Speech Synthesis*, Ph.D. thesis, Tokyo Institute of Technology, 2006.

[53] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," *IEICE Trans. Inf. & Syst.*, vol. E89-D, no. 3, pp. 1092–1099, 2006.

[54] H. Zen, T. Toda, M. Nakamura, and T. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.

[55] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," in *ISCA SSW5*, 2004.

[56] S. Rouibia and Rosec, "Unit selection for speech synthesis based on a new acoustic target cost," in *Interspeech*, 2005, pp. 2565–2568.

[57] T. Hirai and S. Tenpaku, "Using 5 ms segments in concatenative speech synthesis," in *ISCA SSW5*, 2004.

[58] J.-H. Yang, Z.-W. Zhao, Y. Jiang, G.-P. Hu, and X.-R. Wu, "Multi-tier non-uniform unit selection for corpus-based speech synthesis," in *Blizzard Challenge Workshop*, 2006.

[59] N. Mizutani, K. Tokuda, and T. Kitamura, "Concatenative speech synthesis based on HMM," in *Autumn meeting of ASJ*, 2002, pp. 241–242.

[60] Z. Ling and R. Wang, "HMM-based unit selection using frame sized speech segments," in *Interspeech*, 2006, pp. 2034–2037.

[61] J. Kominek and A. Black, "The Blizzard Challenge 2006 CMU entry introducing hybrid trajectory-selection synthesis," in *Blizzard Challenge Workshop*, 2006.

[62] M. Plumpe, A. Acero, H. Hon, and X. Huang, "HMM-based smoothing for concatenative speech synthesis," in *ICSLP*, 1998, pp. 2751–2754.

[63] J. Wouters and M. Macon, "Unit fusion for concatenative speech synthesis," in *ICSLP*, 2000, pp. 302–305.

[64] P. Taylor, "Unifying unit selection and hidden Markov model speech synthesis," in *Interspeech*, 2006, pp. 1758–1761.