A Comparative Study of Spectral Transformation Techniques for Singing Voice Synthesis

S. W. Lee [1], *Zhizheng Wu* [2], *Minghui Dong* [1], *Xiaohai Tian* [2], *and Haizhou Li* [1,2]

[1]Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore
[2]School of Computer Engineering, Nanyang Technological University, Singapore

{swylee, mhdong, hli}@i2r.a-star.edu.sg, {wuzz, xhtian}@ntu.edu.sg

Abstract

Studies show that professional singing matches well the associated melody and typically exhibits spectra different from speech in resonance tuning and singing formant. Therefore, one of the important topics in speech-to-singing conversion is to characterize the spectral transformation between speech and singing. This paper extends two types of spectral transformation techniques, namely voice conversion and model adaptation, and examines their performance. For the first time, we carry out a comparative study over four singing voice synthesis techniques. The experiments on various data sizes reveal that maximum-likelihood Gaussian mixture model (ML-GMM) of voice conversion always delivers the best performance in terms of spectral estimation accuracy; while model adaptation generates the best singing quality in all cases. When a large dataset is available, both techniques achieve the highest similarity to target singing. With a small dataset, the highest similarity is obtained by ML-GMM. It is also found that the music context-dependent modeling in adaptation, in which detailed partition of transform space is involved, leads to pleasant singing spectra.

Index Terms: singing synthesis, speech-to-singing, voice conversion, adaptation, spectral transformation

1. Introduction

Singing voice synthesis has been a popular research topic in recent years [1], [2], [3], [4], [5], [6], enabling innovative services and applications, such as entertainment, music production and computer-assisted vocal training [7], [8], [9], [10].

Pleasant synthetic singing with distinctive vocal characteristics, such as individual timbre, styling in fundamental frequency (F0) etc., is appealing to the general public. This is especially the case for those who are not good at singing. Hence, generating singing voice with high level of quality, naturalness and impressive vocal characteristics is desirable.

Proper spectral transformation is an essential element of high-quality synthetic singing (Others are on F0 and rhythm, etc). Vocal studies indicated that singing formant, resonance tuning and vowel changes are always demonstrated by trained classical singers [11], [12], [13]. Based on the present vowel and F0, the spectral envelope of singing is transformed accordingly for efficient sound transmission [12], [13]. This paper focuses on spectral transformation for singing voice synthesis.

Among several popular approaches of singing voice synthesis, speech-to-singing (S2S) synthesis [14] converts a lyrics-reading speech input to a singing voice output by manipulating acoustic features, namely F0, spectrum and duration, with respect to a reference melody. As the vocal characteristics of an individual are rather captured in his/her speech input, S2S synthesis enables spectral transformation and provides an appropriate framework for generating personalized high-quality singing.

Voice conversion is another potential technique. It has been conventionally used to convert the voice of a source speaker to that of a target speaker [15], [16], [17], [18], [19], [20]. Methods have been proposed to modify the source voice's spectrum and F0 contour acoustically, so as to increase the similarity to target speaker, without knowing the speech content. Voice conversion seems to be suitable for speech-to-singing as it models the mappings between speech and singing. However, the output quality resulted from voice conversion is often degraded. Application in singing synthesis requires tailor-made singing-related algorithmic designs, so as to preserve the voice quality after voice conversion and maintain smooth transitions when moving from a singing segment to the next.

Hidden Markov model (HMM)-based text-to-speech (TTS) [21] sheds light on singing voice too. In HMM-based TTS, HMMs with dynamic features are used to model the vocal tract configurations of speech signals. Given an input text, the output speech is generated with the speech parameters estimated under optimization criteria. Context information such as phone identities in the neighborhood and word position are taken into account. To further approximate certain speaker properties, emotions or speech conditions, model adaptation is applied on these parameters [22], [23], [24], [25]. Saino *et al.* has used the basic HMM-based TTS approach for singing voice synthesis in [4]. Nevertheless, there is rarely a study on the feasibility and performance of using model adaptation algorithms for generating distinctive singing spectrum.

Traditionally, voice conversion and model adaptation are used in different scenarios. Speech content is usually known in model adaptation, but not in voice conversion. Parallel recordings are commonly used as training materials for voice conversion, but not for model adaptation. Personalized singing voice synthesis is a new application such that recordings of speech and singing, together with the music score, can be utilized. *This paper extends the above two types of techniques for the generation of singing spectrum, together with the speech-to-singing technique [14], and compares their performance. This is (to our knowledge) the first paper presenting such comparison for singing voice synthesis.* In particular, we aim to answer the following questions: Given the same training amount, which technique generates the best singing voice? The vocal study by Joliveau *et al.* [12] stated that vowel becomes indistinguishable after resonance tuning. This implies only a small amount of spectral models are needed for singing synthesis. Is it the case? For the same piece of music, singing voices vary a lot, maybe on F0 contour, spectrum and so on, compared to the speech signals reading the same lyrics. What is the sufficient amount of

data to generate proper singing spectra?

The experiments in this paper lay the foundation for many innovative applications. Given some speech and singing recordings of a professional singer, the spectral transformation between speech and singing domains is learnt. This resultant spectral transformation can be used to impersonate the professional singing from someone's speech. With speech and singing recordings collected from multiple professional singers, the singer-independent spectral transformation exhibited generally by all of them can be even learnt by extending the above with speaker-adaptive training (SAT) [24].

2. Extension of transformation techniques

In the following, we will briefly describe the spectral transformations used and highlight the extension we made for singing synthesis. Tandem-STRAIGHT [26] is used as our analysis-reconstruction framework. Singing voice is synthesized segment by segment, where each segment contains a line of lyrics.

2.1. Voice conversion

2.1.1. Maximum-likelihood Gaussian mixture model

Gaussian mixture model (GMM)-based voice conversion remains popular, for its good similarity between converted and target voices, and the probabilistic and flexible framework. We adopt the ML-GMM method with dynamic feature constraint [19] as one of the techniques examined. The voice conversion is done acoustically, without any linguistic nor music content like phone, music note, tempo, etc.

ML-GMM consists of offline training and runtime conversion. During offline training, an GMM jointly models aligned features of the source speech and target singing (with dynamic coefficients) under maximum likelihood criterion. This GMM represents a soft partition of the acoustic space. 34-th order mel generalized cepstral (MGC) coefficients (c0 to c34) are used. At runtime, given this GMM and the source feature trajectory, the converted feature trajectory (defining the output spectral component) is found by maximizing its likelihood function [19].

We prepare parallel speech-singing training data with a two-stage alignment process, tailor-made for this cross-domain alignment. In the first stage, the speech utterance of a speech-singing pair is forced-aligned with a phone-level speech recognizer and the lyrics information. Forced alignment on the associated singing utterance is performed as well. With the phone boundaries in the forced-alignment results, the start and end times of individual phones are found. In the second stage, for each phone in this speech-singing pair, its spectral segments of speech and singing are extracted according to the start and end times. These two spectral segments are then aligned by dynamic time warping. The resultant alignment is used to constitute the sets of aligned feature vectors.

2.1.2. Weighted frequency warping

A variant of the weighted frequency warping (WFW) proposed by Erro *et al.* [20] is adopted as another transformation technique here. This WFW technique combines the typical GMM approach with frequency warping transformation, showing a good balance between speaker similarity and speech quality. Low-order line spectral frequencies (LSFs) are used. After fitting a joint GMM (m mixtures) on the aligned features of speech and singing as in ML-GMM, piecewise linear frequency warping functions are then defined for each GMM mean vectors [20]. During conversion, for each input spectral frame, these frequency warping functions are weighted by the relative probabilities that this input frame belongs to individual GMM components. The resultant function is finally used to warp the input speech spectrum to singing counterpart. We do not employ the energy correction filter as in [20], so as to preserve the output singing quality as much as possible.

We adopt Tandem-STRAIGHT instead of harmonic plus stochastic model (HSM) [20]. In HSM, voiced speech is decomposed into a sum of harmonic components (harmonic frequencies, magnitudes and phases). We know that voicing often switches between speech and singing. The decoupled extraction of spectrum and F0 in Tandem-STRAIGHT allows us to flexibly manipulate these two components and voicing. This essentially avoids the modification on F0 and phase in HSM [20], where spectrum, F0 and phase modifications are possible in voiced-to-voiced scenarios only.

2.2. Model adaptation in HMM-based TTS framework

Our model adaptation technique for singing synthesis is based on the procedure given in [27], [28], but with detailed implementations specific to singing voice. A set of speech models is first built, then adapted to singing.

Using the same set of timing labels as the first stage of the alignment process in voice conversion, monophone models are initialized. Full-context Hidden Semi Markov phone models (HSMMs) with duration modeling are subsequently built. Five left-to-right single-Gaussian emitting states and diagonal covariance are used. The spectral component is represented by the same 34-th order MGC coefficients, together with the log F0 and 5-band aperiodicity. Dynamic coefficients are used. Note that this modeling enables learning of the joint distributions of spectrum, F0 and aperiodicity, which is essential to tonal languages and singing voice (Various music vocal studies show that singing spectrum for the same vowel changes with F0).

Although speech utterance is different from singing voice that there is no music specifications imposed in typical read speech, we explicitly add such information in full-context labels to indirectly link the corresponding speech and singing models together, and enable detailed division in the singing model space built later (This singing model space division will be refined during clustering). MIDI files corresponding to the singing data are used for context labeling. Specifically, our full-context phone labels contain the following linguistic and music information: (1) phone identity (of previous, current and the next), (2) note identity (associated with the previous, current and next phone), (3) note interval relative to the current note in the unit of semitones (associated with the previous and next phone), (4) note duration (associated with the previous, current and the next phone), tempo class of the respective song, number of words in the current line of lyrics, initial identity (of previous, current and the next phone), final identity (of previous, current and the next phone). We work on singing synthesis for Mandarin Chinese songs here, where a Mandarin syllable consists of an optional initial and a final.

Adaptation is then started for the above full-context speech models. We do not implement any SAT here, since all of the speech and singing data in our experiments below are from the same speaker. For data with multiple speakers in the future, SAT may be used. Constrained structural maximum a posteriori linear regression (CSMAPLR) adaptation with structural maximum a posteriori (MAP) criterion [27] is performed.

Synthesized singing should have the rhythm specified by

Table 1: *Comparisons of the four techniques.*

property	voice conversion (ML-GMM & WFW)	adaptation	S2S
statistical or rule-based?	statistical, GMM-based	statistical, HSMM-based	rule-based
identical (global) transform?	transform space is partitioned into m portions ML-GMM: resultant transform is linear, weighted by these m mean vectors, acting on source feature WFW: resultant warping is a weighted function of these m mean vectors, acting on source spectrum	transform space is partitioned into a large no. of portions and only 1 mean vector will be selected	almost identical with little difference in scales for various consonants
intrinsic dynamics in spectra of adjacent singing frames preserved?	ML-GMM: Yes WFW: No	Yes	No
spoken content to be known?	No	Yes	Yes
rhythm in score to be known?	almost no (except for rhythm adjustment)	Yes	Yes
pitch in score to be known?	No	Yes	Yes
power adjustment?	No	Yes	Yes
automatic?	Yes	Yes	No

the music score. Consequently, for a given segment, we constitute the full-context labels and estimate the timing information of individual phones with the corresponding music score. It is found by maximizing the product of all the associated state duration probabilities within each note and scaling to the target note duration. Finally, with this phone timing information, the coefficients of spectrum, F0, aperiodicity are found by the parameter generation algorithm in [21].

2.3. Spectral transformation in speech-to-singing

This S2S synthesis technique [14] is solely designed for personalized singing voice synthesis, manipulating the F0, spectrum and aperiodicity of a lyrics-reading speech input. Specifically, the spectral component retrieved from Tandem-STRAIGHT is transformed in two steps, lengthening and boost to singing formant. First, individual syllables in speech input are manually located and associated to the respective notes in music score. Within each syllable, a 40 msec boundary region between consonant and vowel is marked [14]. The consonant portion is lengthened according to the type of consonant. For the vowel portion, it is extended to match the remaining duration in the respective note, while keeping the boundary region intact. Singing formant is then added to the speech spectrum by multiplying a bandpass filter centering at the peak of speech spectral envelope nearest to 3 kHz. The dip of aperiodicity is emphasized in the same way. These resultant spectrum and aperiodicity are finally combined with singing F0 in Tandem-STRAIGHT to produce the synthesized singing. More implementation details can be found in [14].

2.4. Comparisons of the four techniques

To examine the principles of the above techniques, we compare and highlight their differences in Table 1. In summary, voice conversion and adaptation are automatic and statistical techniques, while S2S requires manual annotation on syllable timing. Adaptation requires the most context information.

3. Experiments

We report both the objective and subjective evaluations of the four spectral transformation techniques below. These are, in particular, relevant for the impersonation application stated in at the end of Section 1. Several indices were used to evaluate their performance on singing voice synthesis, namely (1) cepstral distance of transformed spectra; (2) quality of synthesized singing; (3) similarity to target singing. A collection of solo singing recordings from a male professional singer was used. There were altogether 50 Mandarin Chinese pop songs. Each song lasted about four minutes, totaling 194 min 33 sec. There were corresponding lyrics-reading speech recordings and MIDI files. These constituted 1848 singing segments (and their respective speech segments) for training and 54 segments for testing. These testing segments were unseen from training. For fair comparison across different techniques, the reference singing F0 contours and aperiodicity are used for reconstruction.

3.1. Cepstral distance

The transformation accuracy was examined first by looking at the cepstral distance between the transformed spectra and the target counterpart. The measurements are given in Table 2. For voice conversion, there were several systems built by varying the amount of parallel training segments used and m. For adaptation, the number of adaptation segments was varied. Small training sets are always subsets of large sets.

Table 2: *Cepstral distance (mean [standard derivation]).*

technique(m)	no. of segments				
	50	100	250	500	1848
ML-GMM(16)	5.01 [0.43]	5.08 [0.51]	4.99 [0.38]	4.84 [0.41]	4.79 [0.38]
ML-GMM(32)	5.05 [0.44]	5.03 [0.50]	4.93 [0.42]	4.84 [0.40]	4.97 [0.43]
ML-GMM(64)	5.12 [0.48]	5.06 [0.57]	4.83 [0.43]	4.77 [0.47]	4.74 [0.37]
WFW(16)	7.04 [0.56]	7.05 [0.54]	7.19 [0.57]	6.97 [0.51]	6.96 [0.54]
WFW(32)	7.18 [0.57]	7.21 [0.54]	7.10 [0.59]	7.20 [0.61]	7.04 [0.55]
WFW(64)	7.05 [0.52]	6.99 [0.54]	7.03 [0.58]	7.09 [0.57]	6.96 [0.57]
adaptation	5.98 [0.6]	5.95 [0.6]	5.38 [0.46]	5.37 [0.46]	5.15 [0.45]
S2S	7.37 [0.69]				

Among the four techniques, ML-GMM achieves the lowest cepstral distance; adaptation is ranked as the second. Spectra transformed by WFW or S2S are typically far away from the target spectra. If the number of segments increases from 50 to 1848, the cepstral distances from all systems often decreased. Nevertheless, the trend ML-GMM $<$ adaptation $<$ WFW $<$

S2S remains the same.

3.2. Quality of synthesized singing

In the following subjective listening tests, the best system among each technique in Section 3.1 was tested. *We studied on two cases: little data (50 segments) and large data (1848 segments).* In the first listening test, listeners were asked to compare and rate the singing quality of the various systems by mean opinion score (MOS). Possible MOS ranged from 1 (bad) to 5 (excellent). For large data case, ML-GMM_64m_1848t, WFW_64m_1848t, A_1848t and S2S were compared (A system with name αm_βt means the number of mixtures and the number of segments are α and β respectively). For little data case, ML-GMM_16m_50t, WFW_16m_50t, A_50t and S2S were tested. There were 10 testing segments, randomly taken from the testing set. Listeners could play the stimuli as many times as they wished. A total of 17 listeners participated.

Fig. 1 shows the box plots of the MOS result. On each box, the central mark is the median. The edges are the 25th and 75th percentiles. Outliers are indicated by '+'. The experiment results suggested that for large data case, the singing quality achieved by adaptation (A_1848t) is significantly better than others (with 95% confidence intervals). S2S is ranked the second. The two voice conversion systems (ML-GMM_64m_1848t and WFW_64m_1848t) performed more or less the same. For little data case, adaptation (A_50t) and S2S achieve similar singing quality and outperform the remaining two voice conversion techniques. Measurements in lower quartile of S2S have slightly higher MOS than adaptation. WFW (WFW_16m_50t) is ranked the third and significantly better than ML-GMM (ML-GMM_16m_50t) with 95% confidence intervals. This indicates the frequency warping acting on source spectrum brings quality improvement over ML-GMM for little data case. This improvement is not prominent for the large data case.

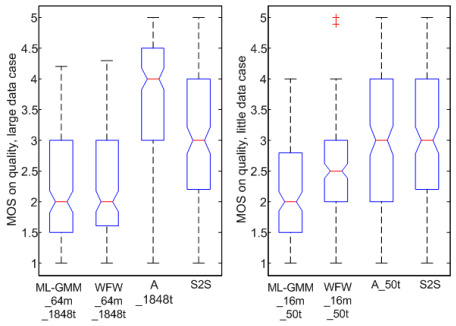

Figure 1: *Results on singing quality on (left) large data case and (right) little data case.*

3.3. Similarity to target singing

The similarity to target singing from various systems was measured in the second listening test. Same systems from the first listening test were evaluated on the little and large data cases. The recorded singing with Tandem-STRAIGHT analysis and reconstruction (no other modification) acted as the target singing. Pairs of a converted singing and the corresponding target singing were presented to listeners at random order. Listeners were asked to determine how similar the vocal characteristics of the converted singing to the target counterpart, without paying attention the quality. The similarity is on a 1-to-5 MOS scale (1 representing "extremely different" and 5 representing"extremely similar"). Listeners could play the stimuli as many times as they wished. There were five testing segments for each system. A total of 19 listeners participated. Fig. 2 shows the box plots of the MOS result.

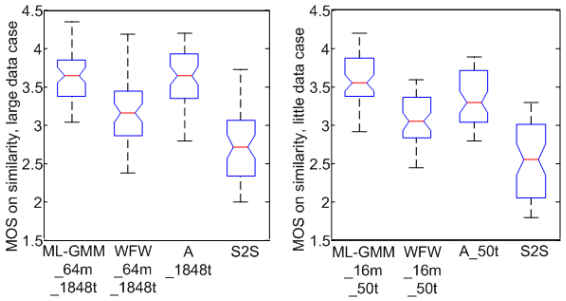

Figure 2: *Results on similarity on (left) large data case and (right) little data case.*

For large data case, output singing generated from ML-GMM and adaptation are found to achieve the highest similarity, followed by the singing generated from WFW. The similarity achieved by S2S is much lower. For little data case, highest similarity is from the singing generated by ML-GMM, which is significantly higher than adaptation with 95% confidence intervals. WFW and S2S are ranked as the third and the last place respectively.

Taking all these results into consideration, we found that: *Given the same large amount of singing data, adaptation is the technique that offers the best spectral transformation, in terms of distance measure, quality and similarity*. Concerning the number of singing segments, measurements of cepstral distance are more or less the same for system A_250t and A_500t, while A_1848t has much lower distance measure. Our preliminary listening test showed that the distortions in outputs from A_250t or A_500t are not found in the outputs from A_1848t. *All of these indicate that this amount of singing data is essential, leading to significantly high-quality singing*. If less singing segments are used, the quality of singing is still alright with little distortion.

The four techniques have very different model sizes, ranging from nearly global transform for S2S, dozens of models for ML-GMM and WFW, to thousands of transforms for adaptation. For large data case, output quality from voice conversion and S2S is far below the one from adaptation. For little data case, adaptation offers similar quality as S2S, but with higher similarity to the target singing. In our preliminary listening tests, given a fixed number of segments, we found that the output quality remains roughly the same even the number of mixtures used in ML-GMM or WFW increases. *The outstanding performance of adaptation probably indicates that a large number of context-dependent models (detailed division of transform space) are needed for satisfactory spectral transformation.*

4. Conclusions

Singing has high variability in spectral evolution, pitch, for instance. Converting an input speech to singing voice enables impersonation and personalized singing synthesis for laymen. This paper focuses on the spectral transformation from speech to singing. We extend two types of state-of-the-art techniques for singing synthesis and examine their performance with other alternatives. Experiments indicate that the extended transformation with model adaptation on large data offers the best quality and similarity, where music context-specific transformation contributes to the outstanding performance.

5. References

[1] *Synthesis of Singing Challenge (Special Session), Proc. Interspeech*, Aug. 2007.

[2] M. Akagi, "Rule-based voice conversion derived from expressive speech perception model: How do computers sing a song joyfully?" in *Proc. ISCSLP*. Tutorial 01, Nov. 2010.

[3] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, pp. 67–79, 2007.

[4] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-based singing voice synthesis system," in *Proc. Interspeech*, Sep. 2006, pp. 2274–2277.

[5] S. W. Lee, S. T. Ang, M. Dong, and H. Li, "Generalized F0 modeling with absolute and relative pitch features for singing voice synthesis," in *Proc. ICASSP*, Mar. 2012, pp. 429–432.

[6] S. W. Lee and M. Dong, "Singing voice synthesis: Singer-dependent vibrato modeling and coherent processing of spectral envelope," in *Proc. Interspeech*, Aug. 2011, pp. 2001–2004.

[7] H. Kenmochi and H. Ohshita, "VOCALID – Commercial singing synthesizer based on sample concatenation," in *Proc. Interspeech*, Aug. 2007.

[8] P. Kirn, "iPhone Day: LaDiDa's Reverse Karaoke Composes Accompaniment to Singing [Online]," Mar. 2014, available: http://createdigitalmusic.com/2009/10/iphone-day-ladidas-reverse-karaoke-composes-accompaniment-to-singing/.

[9] "An app with speech-to-singing utility. NDP 2013 Mobile App [Online]," Mar. 2014, available: https://itunes.apple.com/sg/app/ndp-2013-mobile-app/id524388683?mt=8.

[10] M. Goto, T. Nakano, S. Kajita, Y. Matsusaka, S. Nakaoka, and K. Yokoi, "Vocalistener and vocawatcher: Imitating a human singer by using signal processing," in *Proc. ICASSP*, Mar. 2012, pp. 5393–5396.

[11] J. Wolfe, M. Garnier, and J. Smith, "Vocal tract resonances in speech, singing and playing music instruments," *Human Frontier Science Program Journal*, vol. 3, pp. 6–23, 2009.

[12] E. Joliveau, J. Smith, and J. Wolfe, "Tuning of vocal tract resonance by sopranos," *Nature*, vol. 427, p. 116, Jan. 2004.

[13] J. Sundberg, "The acoustics of the singing voice," *Scientific American*, vol. 236, pp. 82–91, Mar. 1977.

[14] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2007, pp. 215–218.

[15] E. Moulines and Y. Sagisaka, "Voice conversion: State of the art and perspective," *Special Iss. Speech Commun.*, vol. 16, no. 2, 1995.

[16] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech & Audio Proc.*, vol. 6, pp. 131–142, Mar. 1998.

[17] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Proc. ICASSP*, May 2001, pp. 841–844.

[18] A. B. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Science & Engineering, Oct. 2001.

[19] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, & Lang. Proc.*, vol. 15, pp. 2222–2235, Nov. 2007.

[20] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. Audio, Speech, & Lang. Proc.*, vol. 18, pp. 922–931, Jul. 2010.

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

[22] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP*, May 2011, pp. 805–808.

[23] J. Yamagishi, T. Masuko, and T. Kobayashi, "HMM-based expressive speech synthesis – Towards TTS with arbitrary speaking styles and emotions," in *Proc. Special Workshop in Maui (SWIM)*, Jan. 2004.

[24] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, pp. 533–543, Feb. 2007.

[25] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio, Speech, & Lang. Proc.*, vol. 20, pp. 2505–2517, Sep. 2012.

[26] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," in *Proc. ICASSP*, Mar. 2008, pp. 3933–3936.

[27] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, & Lang. Proc.*, vol. 17, pp. 66–83, Jan. 2009.

[28] H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A. W. Black, and K. Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)," in *Proc. APSIPA ASC*, Oct. 2009, pp. 121–130.