New Method for Rapid Vocal Tract Length Adaptation in HMM-based Speech Synthesis

Daniel Erro [1,2], *Agustín Alonso*[1], *Luis Serrano*[1], *Eva Navas*[1], *Inma Hernáez*[1]

[1] AHOLAB, University of the Basque Country (UPV/EHU), Bilbao, Spain
[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
derro@aholab.ehu.es

Abstract

We present a new method to rapidly adapt the models of a statistical synthesizer to the voice of a new speaker. We apply a relatively simple linear transform that consists of a vocal tract length normalization (VTLN) part and a long-term average cepstral correction part. Despite the logical limitations of this approach, we will show that it effectively reduces the gap between source and target voices with only one reference utterance and without phonetic segmentation. In addition, by using a minimum generation error criterion we avoid some of the problems that have been reported to arise when using a maximum likelihood criterion in VTLN.

Index Terms: statistical parametric speech synthesis, speaker adaptation, vocal tract length normalization

1. Introduction

The length of the vocal tract is one of the perceptually relevant characteristics of a speaker's voice, being known to correlate well with gender and/or age. Therefore, VTLN techniques are useful to make speech processing systems able to operate with a wide variety of voices. VTLN has been traditionally applied to compensate for vocal tract length mismatches between the pre-trained statistical models and the input voices in automatic speech recognition (ASR) [1]. Thus, the word error rate is reduced by 7-10% with respect to an equivalent nonadaptive ASR system.

From the speech generation side, VTLN has also been applied in voice conversion [2] and more recently in speaker-adaptive synthesis [3] to mimic the characteristics of specific target speakers. Frequency warping functions are used to transfer the vocal tract length of the target speaker to the generated speech by modifying either the signal (conversion) or the generative models (synthesis). In this context, VTLN has two main advantages with respect to other types of transformation: (i) the almost null degradation of the quality; (ii) the robustness of the method when few training data are available, which is due to the generally low dimension of the transformation function. These two advantages are often sufficient to justify the use of VTLN in speech generation even though the similarity between frequency-warped and target voices is obviously moderate.

In the particular case of speech synthesis based on hidden Markov models (HMMs), an extensive study was presented in [3] in which the main challenges arising when integrating VTLN in this framework were analyzed. Choosing the popular all-pass transform based on a bilinear function as basic frequency warping curve with only one parameter [4], several model adaptation strategies based on maximum likelihood (ML) criteria were examined. We would like to highlight some observations from the work presented in [3]: (a) the high dimension of the Mel-cepstral vectors typically used in synthesis hinders the adaptation process driven by likelihoods; (b) special attention has to be paid to Jacobian normalization during adaptation to avoid unstabilities; (c) during adaptation, a numerical algorithm is necessary to search for the maximum of an auxiliary function at each iteration of the expectation-maximization algorithm, which results in a doubly iterative procedure.

In our previous works on voice conversion, we showed the usefulness of the so called BLFW+AS (bilinear frequency warping plus amplitude scaling) method [5]. Fed with Mel-cepstral vectors, this method uses a GMM to partition the acoustic vector space of the source speaker into overlapping classes, each class being assigned specific frequency warping and amplitude scaling functions. Like the aforementioned VTLN-based speaker adaptation method, BLFW+AS uses bilinear frequency warping functions with one single parameter. It also uses additive cepstral terms as amplitude scaling functions that compensate for the differences between frequency-warped and target spectra. This paper reports the preliminary steps towards the design of a rapid speaker adaptation method inspired by BLFW+AS in the context of statistical parametric speech synthesis.

Interestingly, although BLFW+AS was designed to operate with multiple overlapping classes, the objective scores presented in [5] (and also those in [3]) suggested that a single frequency warping function followed by many class-dependent amplitude scaling terms performed almost equally well. Therefore, in order to facilitate the design of a BLFW+AS-based adaptation method, in this preliminary approach we have considered only the basic case with a unique transformation class, which means using the same transform in all the acoustic and phonetic contexts. This simplified method can be seen as VTLN followed by a sort of long-term average cepstral correction. In this document, emphasis will be placed on the estimation of the VTLN factor. Future extensions of this work will consider the use of many class-dependent amplitude scaling additive cepstral terms.

Regarding the estimation of the VTLN factor, the method we propose exhibits some remarkable differences with respect to the one described in [3]: (a) our method can deal with high-dimensional Mel-cesptral vectors because it is not based on ML criteria but on a different criterion similar to minimum generation error (MGE) [6]; (b) for the same reason, the Jacobian normalization related problem reported in [3] is avoided; (c) our solution is based on the iterative VTLN training algorithm presented in [7], which converges very rapidly and consistently with no irregular behaviors.

The remainder of this paper is structured as follows. Section 2 summarizes the algorithm that allows calculating a VTLN factor from two parallel sets of cepstral vectors. Then, for a better understanding of the MGE-based method we propose, section 3 will give a brief overview of standard parameter generation techniques used in HMM-based speech

synthesis. In sections 4 and 5 we will describe and evaluate the proposed method assuming one single recorded utterance for adaptation and its corresponding text.

2. Iterative estimation of VTLN factor from aligned vectors

All-pass transforms based on bilinear functions are one of the most popular choices in VTLN [8][4]. This section gives an introduction to this particular type of frequency warping function and shows how the optimal value of its unique parameter can be learnt from a set of aligned source and target vectors, $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_t\}$. Bilinear functions can be defined in the z domain in terms of one single parameter α:

$$z^{(\alpha)^{-1}} = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}}\ ,\quad z = e^{j\omega}\ ,\quad z^{(\alpha)} = e^{j\omega^{(\alpha)}}\ ,\quad |\alpha|<1 \tag{1}$$

The corresponding mapping between the original frequency scale and the warped one is given by

$$\omega^{(\alpha)} = \tan^{-1}\frac{\left(1-\alpha^2\right)\sin\omega}{\left(1+\alpha^2\right)\cos\omega - 2\alpha} \tag{2}$$

Previous research has shown that, given the VTLN factor α, the cepstral representation of a spectrum, $\mathbf{x}$, can be transformed into that of the corresponding frequency-warped spectrum, $\mathbf{x}^{(\alpha)}$. This cepstral transformation can be expressed as a linear operation [4][9]:

$$\mathbf{x}^{(\alpha)} = \mathbf{A}_\alpha \mathbf{x}\ ,\quad \mathbf{A}_\alpha = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots \\ 0 & 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots \\ 0 & -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3}$$

Given the strongly nonlinear dependence between $\mathbf{A}_\alpha$ and α, it is difficult to estimate the best value of α from aligned source and target training data. However, since $|\alpha| << 1$ when VTLN is performed on realistic human voices (in general, $|\alpha| < 0.1$), one can think of simplifying $\mathbf{A}_\alpha$ by neglecting the terms of the form α^n for $n > 1$, as originally proposed in [10]. This results in a more manageable transformation:

$$\mathbf{x}^{(\alpha)} \cong \mathbf{x} + \alpha \cdot \mathbf{d}(\mathbf{x}) \tag{4}$$

where $\mathbf{d}(\mathbf{x})$ is the vector whose i^{th} element is given by

$$\mathbf{d}(\mathbf{x})[i] = (i+1)\cdot\mathbf{x}[i+1] - (i-1)\cdot\mathbf{x}[i-1]\ ,\quad i = 0, 1, 2... \tag{5}$$

Under the assumption that (4) is accurate enough, given a set of T source and target parallel training vectors, $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_t\}$, it can be shown [7] that the VTLN factor than minimizes the error between warped and target vectors is

$$\alpha = \frac{\sum_{t=1}^{T}\mathbf{d}^{\mathsf{T}}(\mathbf{x}_t)\cdot(\mathbf{y}_t - \mathbf{x}_t)}{\sum_{t=1}^{T}\left\|\mathbf{d}(\mathbf{x}_t)\right\|^2} \tag{6}$$

In our previous works [7][5] we found expression (4) to be inaccurate for many voices, especially in cross-gender transformation (which is the case where accurate VTLN is most neeeded). Therefore, we proposed the following iterative algorithm to get the minimum-error value of α according to the full formulation (3):

- Step 1: initialize α as 0.
- Step 2: for the current α, calculate a set of warped vectors $\{\mathbf{x}_n^{(\alpha)}\}$, $\mathbf{x}_n^{(\alpha)} = \mathbf{A}_\alpha \mathbf{x}_n$, where the warping matrix $\mathbf{A}_\alpha$ is given by expression (3).
- Step 3: calculate the incremental warping factor $\Delta\alpha$ that is necessary to make the vectors $\{\mathbf{x}_n^{(\alpha)}\}$ closer to the target vectors $\{\mathbf{y}_n\}$. This is done by solving the approximate expression (6) for $\{\mathbf{x}_n^{(\alpha)}\}$ instead of $\{\mathbf{x}_n\}$.
- Step 4: accumulate $\Delta\alpha$ into the current α. This can be done via the following expression [8]:

$$\alpha^{(updated)} = \frac{\alpha + \Delta\alpha}{1 + \alpha\cdot\Delta\alpha} \tag{7}$$

- Step 5: if the last update of α was insignificant (in other words, if $|\Delta\alpha|$ was lower than a threshold), exit. Otherwise, go back to step 2.

3. ML parameter generation from HMMs

For a better understanding of the method to be proposed next, this section explains briefly how the speech parameter generation algorithm of a standard statistical synthesizer [11][12] works. Although usually referred to as HMM-based synthesis, statistical parametric speech synthesis is actually based on context dependent hidden semi Markov models (CD-HSMMs), where the duration of each state is explicitly modeled through normal distributions instead of depending on state transition probabilities. During training, CD-HSMMs are used to model the correspondence between the phonetic, linguistic and prosodic context labels and the observed acoustic parameters (together with their 1st and 2nd-order derivatives over time). During synthesis, once the context labels are extracted from the input text, the system's engine determines the sequence of CD-HSMM states that corresponds to that text and also the duration of each state (either using statistics or specifications by the user). Let us refer to the state index at frame t as m_t. The goal is finding the most probable sequence of acoustic vectors $\{\mathbf{y}_t\}_{t=1\ldots T}$ given the sequence of mean vectors $\{\boldsymbol{\mu}_{m_t}\}_{t=1\ldots T}$ and covariance matrices $\{\boldsymbol{\Sigma}_{m_t}\}_{t=1\ldots T}$. To make the problem mathematically tractable, the output sequence is expressed as a supervector:

$$\mathbf{y} = \left[\mathbf{y}_1^{\mathsf{T}}\quad \mathbf{y}_2^{\mathsf{T}}\quad \ldots\quad \mathbf{y}_T^{\mathsf{T}}\right]^{\mathsf{T}} \tag{8}$$

Defining $\mathbf{W}$ as the matrix that appends dynamic features to the vectors in $\mathbf{y}$ and omitting the derivation (interested readers should refer to [11][12] for details), the most probable $\mathbf{y}$ is

$$\mathbf{y} = \left(\mathbf{W}^{\mathsf{T}}\mathbf{D}\mathbf{W}\right)^{-1}\mathbf{W}^{\mathsf{T}}\mathbf{D}\boldsymbol{\mu} \tag{9}$$

where

$$\boldsymbol{\mu} = \left[\boldsymbol{\mu}_{m_1}^{\mathsf{T}}\quad \boldsymbol{\mu}_{m_2}^{\mathsf{T}}\quad \ldots\quad \boldsymbol{\mu}_{m_T}^{\mathsf{T}}\right]^{\mathsf{T}} \tag{10}$$

and $\mathbf{D}$ is a block-diagonal matrix given by

$$\mathbf{D} = \text{diag}\left\{\boldsymbol{\Sigma}_{m_1}^{-1}, \boldsymbol{\Sigma}_{m_2}^{-1}, \ldots, \boldsymbol{\Sigma}_{m_T}^{-1}\right\} \tag{11}$$

Although the synthesis engine of modern synthesizers includes a global variance enhancement algorithm [13], this is not crucial for our MGE-based adaptation method to perform correctly.

4. Adaptation based on MGE criterion

The algorithms shown in the previous sections provide the necessary tools to build a MGE-based VTLN method in the context of statistical parametric speech synthesis. The idea is (i) to generate a synthetic copy of the utterances available for adaptation and then (ii) to calculate the VTLN factor α that produces the lowest error between warped synthetic utterances and adaptation utterances. For simplicity, we will assume a single adaptation utterance given by the acoustic vector set $\{\mathbf{x}_t\}_{t=1\ldots T}$ and its corresponding text, from which the synthesis engine can determine the sequence of CD-HSMM states to be used. For clarity, we will assign an index to each state in order of appearance: {1, 2, … , M}. Since the iterative algorithm in section 3 requires a set of aligned vectors as input, the state durations must match those of the target utterance. For a more versatile adaptation, it is interesting to perform time-alignment automatically even when a segmentation of that reference utterance is not available. Therefore, in this work we use the synthesis models to obtain such segmentation via forced alignment. A Viterbi search is carried out to establish the correspondence between frames {1…T} and states {1…M} by determining the sequence {m_1 … m_T} that fulfils the continuity and left-to-right conditions ($m_1 = 1$; $m_T = M$; $m_{t+1} = m_t$ or $m_t + 1$ for all t) and maximizes the following log-likelihood function:

$$C\{m_1, m_2 \ldots m_T\} = \sum_{t=1}^{T} \log N(\mathbf{X}_t; \boldsymbol{\mu}_{m_t}, \boldsymbol{\Sigma}_{m_t}) + \\ + \sum_{t=1}^{T-1} \log p(m_{t+1}/m_t, d_t) \tag{12}$$

where $\mathbf{X}_t$ is the result of appending dynamic features to $\mathbf{x}_t$, d_t is the duration of state m_t until frame t (it can be obtained recursively: $d_t = 1$ if $t = 1$ or $m_t \neq m_{t-1}$; $d_t = d_{t-1} + 1$ elsewhere), N denotes the normal distribution, and

$$p(m_{t+1}/m_t, d_t) = \begin{cases} \theta_{m_t, d_t} & , \; m_{t+1} = m_t \\ 1 - \theta_{m_t, d_t} & , \; m_{t+1} \neq m_t \end{cases} \tag{13}$$

$$\theta_{m_t, d_t} = \int_{d_t}^{\infty} N(\lambda; \mu_{m_t}^{(d)}, \sigma_{m_t}^{(d)^2}) d\lambda$$

Note that (13) means calculating the probability that the duration of the current state is greater than it was at frame t according to the duration means $\mu^{(d)}$ and variances $\sigma^{(d)2}$ learnt during CD-HSMMs training.

Once the state durations are known, ML parameter generation (9)–(11) can be applied to obtain a synthetic version of the adaptation material, $\{\mathbf{y}_t\}_{t=1\ldots T}$, $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_t\}$ being completely parallel. In these conditions, it would be straightforward to obtain the necessary VTLN factor α by means of the algorithm in section 2. The main problem of this approach is that high vocal tract length contrasts between the source (synthetic) voice and the target voice may result in inaccurate state durations and therefore inaccurate α. To avoid it, the following iterative algorithm is applied to jointly optimize α and the durations:

- Step 1: initialize α as 0.
- Step 2: for the current α, calculate a set of frequency-warped adaptation vectors $\{\mathbf{x}_t^{(\alpha)}\}$, $\mathbf{x}_t^{(\alpha)} = \mathbf{A}_\alpha \mathbf{x}_t$, where the involved matrix is given by expression (3).
- Step 3: use the forced alignment method described above to determine the state durations using $\{\mathbf{x}_t^{(\alpha)}\}$ (not $\{\mathbf{x}_t\}$) as reference; then generate $\{\mathbf{y}_t\}$ through expressions (9)–(11).
- Step 4: calculate a new α using the iterative method in section 2, taking the adaptation vectors $\{\mathbf{x}_t\}$ as source and the synthetic vectors $\{\mathbf{y}_t\}$ as target (although this is the opposite direction to the desired one, it simplifies the calculations substantially). Note that the current $\{\mathbf{y}_t\}$ depends on the current durations, which in turn depend on the current α (we avoid more specific notation for clarity).
- Step 5: if the last update of α was insignificant, multiply α by -1 (this means inverting the warping function, thus making it suitable to transform the synthetic voice into the target voice) and exit. Otherwise, go back to step 2.

Under the assumption that the state durations and the resulting set $\{\mathbf{y}_t\}$ have converged together with the VTLN factor α, an additive cepstral correction term is calculated as a complement for VTLN:

$$\mathbf{b} = \frac{1}{T} \sum_{t=1}^{T} (\mathbf{x}_t - \mathbf{A}_\alpha \mathbf{y}_t) \tag{14}$$

The use of this additive term can be seen as as sort of long-term average spectrum normalization. The final adaptation of the mean vectors $\{\boldsymbol{\mu}_m\}$ and covariance matrices $\{\boldsymbol{\Sigma}_m\}$ at every state of the trained CD-HSMMs is carried out as follows:

$$\hat{\boldsymbol{\mu}}_m = \left[(\mathbf{A}_\alpha \boldsymbol{\mu}_m^{(s)} + \mathbf{b})^\intercal \quad (\mathbf{A}_\alpha \boldsymbol{\mu}_m^{(\Delta)})^\intercal \quad (\mathbf{A}_\alpha \boldsymbol{\mu}_m^{(\Delta\Delta)})^\intercal \right]^\intercal \\ \hat{\boldsymbol{\Sigma}}_m = \mathrm{diag}\{\mathbf{A}_\alpha \boldsymbol{\Sigma}_m^{(s)} \mathbf{A}_\alpha^\intercal, \; \mathbf{A}_\alpha \boldsymbol{\Sigma}_m^{(\Delta)} \mathbf{A}_\alpha^\intercal, \; \mathbf{A}_\alpha \boldsymbol{\Sigma}_m^{(\Delta\Delta)} \mathbf{A}_\alpha^\intercal\} \tag{15}$$

where diag{…} denotes a block-diagonal matrix and (s), (Δ) and (ΔΔ) denote the sub-parts of the vectors/matrices related to static features, their 1st derivatives and their 2nd derivatives, respectively.

Interestingly, we found the described method to perform better when the involved vectors are weighted by the local probability of voicing when calculating both α and $\mathbf{b}$ through expressions (6) and (14), respectively. This prevents long silences and unvoiced segments from biasing the results too much. In HMM-based speech synthesis, the probability of voicing at each state can be easily extracted from the weights of the multi-space distributions (MSD) used to model/generate the log f_0 contour [14].

Finally, an average pitch modification factor is calculated by generating a synthetic log f_0 contour according to the last instance of the state durations and comparing it with that of the adaptation utterance. In this case, adaptation is performed by summing the appropriate constant value to the static part of the mean vectors of the CD-MSD-HSMMs trained from log f_0.

5. Preliminary evaluation

As discussed in [3], evaluating VTLN is not an easy task because the behavior of the method depends on the specific voices involved in the test. In addition, performing VTLN followed by long-term average cepstral correction implies modifying just a few characteristics of the source voice. Therefore, for arbitrary input voices and large amounts of data the proposed adaptation method cannot compete with more sophisticated methods such as the well known CSMAPLR [15]. On the other hand, we would also like to emphasize that these are just the preliminary steps towards the design of a better method inspired by the BLFW+AS voice conversion

method [5]. Taking all this into account, we have conducted a relatively simple perceptual test to show that (i) our method can effectively reduce the gap between the voice of a synthesizer and a given target voice and (ii) it can do it rapidly using only one reference utterance and its corresponding text.

The text-to-speech (TTS) synthesis system used in our experiments, AhoTTS [16], includes a statistical engine based on HTS [17] and a Mel-cepstral vocoder based on a harmonics plus noise model [18]. The default voice of the system was trained from 2k utterances recorded from a female speaker in Castilian Spanish and digitized at 16 kHz sampling frequency. In previous informal listening tests we had found this voice to be quite suitable for VTLN-based transformations, even towards male voices. We recorded one short utterance (9 words) from 11 different non-professional speakers (5 female plus 6 male speakers). Using them as target, we applied our adaptation method to transform the models of AhoTTS's default voice and then we synthesized speech in all of these voices. Next, 15 volunteer listeners (half of them were speech processing experts) rated the following aspects on a 5-point scale: similarity between the default synthetic voice and the target natural voice, similarity between the adapted synthetic voice and the target natural voice, and relative quality of the adapted voice with respect to the default synthetic voice. As usual, the score indicating the lowest similarity/quality is 1 and the highest score is 5.

Table 1. *Results of the perceptual test: MOS and 95% confidence interval.*

	Source-target Sim.	Adapted-target Sim.	Quality of adapted
Intra-gender	1.40 ± 0.16	2.78 ± 0.25	4.12 ± 0.25
Cross-gender	1.03 ± 0.04	2.63 ± 0.20	3.74 ± 0.22
Total average	1.20 ± 0.08	2.69 ± 0.16	3.94 ± 0.17

The mean opinion scores (MOSs) summarized in Table 1 indicate that, despite the evident differences between source and target voices (~1.2 similarity MOS on a 1-to-5 scale) and the low amount of training material, the proposed method makes the adapted voice significantly closer to the target (~2.7 similarity MOS). The quality loss due to the adaptation process is not particularly high (~4 relative quality MOS). Given the differences between intra-gender and cross-gender cases, we believe that this apparent 1-point quality gap is partially related to the naturalness of the voice that results from this particular type of adaptation rather than to the appearance of artifacts. Overall, taking into account the nature of the method and despite the absence of baseline methods in the listening test, these MOSs reveal that our research goes in the right direction and also that the method proposed in this preliminary work still needs to be improved in order to achieve more satisfactory similarity MOSs.

6. Conclusions

We have presented a new method to adapt the voice of an HMM-based synthesizer using a single unlabeled utterance from the target speaker and its corresponding text. Despite using a relatively simple transformation consisting of VTLN, long-term average cepstral correction and pitch shifting, the method succeeds at reducing the distance between adapted and target voices significantly. Future works will aim at improving the similarity scores achieved by the method through the use of class-dependent additive cepstral terms.

7. Acknowledgements

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (SpeechTech4All, TEC2012-38939-C03-03), the Basque Government (Ber2tek, IE12-333) and Euroregion Aquitaine-Euskadi (Iparrahotsa, 2012-004).

8. References

[1] P. Zhan, A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition", CMU Computer Science Tech. Rep., 1997.

[2] D. Sündermann, H. Ney, "VTLN-based voice conversion", Proc. IEEE Symp. Signal Process. Inf. Technol., pp. 556–559, 2003.

[3] L. Saheer, J. Dines, P. Garner, "Vocal Tract Length Normalization for Statistical Parametric Speech Synthesis", IEEE Trans. Audio, Speech and Lang. Process., vol. 20(7), pp. 2134-2148, 2012.

[4] J. McDonough, W. Byrne, "Speaker adaptation with all-pass transforms", Proc. ICASSP, pp. 757–760, 1999.

[5] D. Erro, E. Navas, I. Hernaez, "Parametric Voice Conversion Based on Bilinear Frequency Warping Plus Amplitude Scaling", IEEE Trans. Audio, Speech and Lang. Process., vol. 21(3), pp. 556-566, 2012.

[6] L. Qin, Y.J. Wu, Z.H. Ling, R.H. Wang, L.R. Dai, "Minimum generation error linear regression based model adaptation for HMM-based speech synthesis", Proc. ICASSP, pp. 3953-3956, 2008.

[7] D. Erro, E. Navas, I. Hernaez, "Iterative MMSE Estimation of Vocal Tract Length Normalization Factors for Voice Transformation", Proc. Interspeech, pp. 86-89, 2012.

[8] A. Acero, "Acoustical and environmental robustness for automatic speech recognition", Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1990.

[9] M. Pitz, H. Ney, "Vocal tract normalization equals linear transformation in cepstral space", IEEE Trans. Speech Audio Process., vol. 13(5), pp. 930–944, 2005.

[10] T. Emori, K. Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation", Proc. Eurospeech, pp. 1649–1652, 2001.

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

[12] H. Zen, K. Tokuda, A. W. Black, "Statistical parametric speech synthesis", Speech Commun., vol. 51(11), pp. 1039-1064, 2009.

[13] T. Toda, K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis", IEICE Trans. Info. & Syst., vol. E90-D(5), pp. 816-814, 2007.

[14] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling", Proc. ICASSP, pp. 229-232, 1999.

[15] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," IEEE Trans. Audio, Speech, Lang. Process., vol. 17(1), pp. 66–83, 2009.

[16] I. Sainz, D. Erro, E. Navas, I. Hernaez, J. Sanchez, I. Saratxaga, I. Odriozola, I. Luengo, "Aholab Speech Synthesizers for Albayzin 2010", Proc. FALA, pp. 343-347, 2010.

[17] Online: "HMM-based Speech Synthesis System (HTS)", http://hts.sp.nitech.ac.jp/

[18] D. Erro, I. Sainz, E. Navas, I. Hernaez, "Improved HNM-based vocoder for statistical synthesizers", Proc. InterSpeech, pp. 1809-1812, 2011.