interspeech 2008

An Improved One-to-Many Eigenvoice Conversion System

Yamato Ohtani, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science,
Nara Institute of Science and Technology, Japan
{yamato-o, tomoki, sawatari, shikano}@is.naist.jp

Abstract

We have previously developed a one-to-many eigenvoice conversion (EVC) system enabling the conversion from a specific source speaker's voice into an arbitrary target speaker's voice. In this system, eigenvoice Gaussian mixture model (EV-GMM) is trained in advance with multiple parallel data sets composed of utterance pairs of the source and many pre-stored target speakers. The EV-GMM is effectively adapted to an arbitrary target speaker using a small amount of adaptation data. Although this system achieves the very flexible training of the conversion model, the quality of the converted speech is still not high enough. In order to alleviate this problem, we simultaneously apply the following promising techniques to the one-to-many EVC system: 1) STRAIGHT mixed excitation, 2) the conversion algorithm considering global variance, and 3) speaker adaptive training of the EV-GMM. Experimental results demonstrate that the proposed system causes remarkable improvements in the performance of EVC.

Index Terms: Speech synthesis, eigenvoice conversion, speaker adaptive training, mixed excitation, global variance

1. Introduction

Voice conversion (VC) [1] is a technique for converting input voice characteristics into other ones without changing linguistic information. As one of VC frameworks, a statistical method based on Gaussian mixture model (GMM) is used widely [2]. In this method, GMM is trained with a parallel data set consisting of utterance pairs of source and target speakers. When considering VC applications, this training framework causes many limitations, such as the necessity for many utterance pairs.

In order to make the training process more flexible, we proposed eigenvoice conversion (EVC) [3] in which the eigenvoice technique [4] is successfully applied to the GMM-based VC. As one of the EVC applications, we have developed a one-to-many EVC system [5] allowing the conversion from a specific source speaker's voice into an arbitrary target speaker's voice. In advance, eigenvoice GMM (EV-GMM) is trained with multiple parallel data sets consisting of the source speaker and many pre-stored target speakers. The resulting EV-GMM enables us to control the voice quality of the converted speech by manipulating a few free parameters, i.e. weights for eigenvectors. Moreover, we can estimate appropriate values of these parameters for given target speakers' voices without any linguistic information.

The conventional one-to-many EVC system has high flexibility for constructing the conversion model. However, the converted speech quality of the conventional one-to-many EVC system is still not high enough as shown in [5]. This is because the conventional system employs the following techniques:

- the simple excitation model based on switching a phase-manipulated pulse train and white noise signal [6], which is too simple to model the excitation signal appropriately;
- the spectral conversion algorithm not considering global variance (GV) [7], which often causes over-smoothed spectral parameters;
- the EV-GMM of which tied parameters are from the target speaker independent GMM (TI-GMM) [3], which causes the conversion model improperly to capture acoustic variations among many pre-stored target speakers.

These techniques often make the converted speech buzzing and muffled.

In this paper, we improve the one-to-many EVC system by applying all of the following promising techniques:

- STRAIGHT mixed excitation (ST-ME) [8] for VC [9],
- the spectral conversion algorithm considering GV [7],
- speaker adaptive training (SAT) [10] of the EV-GMM [11].

It has been reported that each of these techniques causes significant improvements of the converted speech quality. Therefore, it is worthwhile to investigate their effectiveness in the one-to-many EVC system. It is demonstrated from the results of experimental evaluations that a combination of these techniques causes dramatic quality improvements of the one-to-many EVC system.

The paper is organized as follows. In Section 2, we describe the conventional one-to-many EVC system. In Section 3, the proposed one-to-many EVC system is described. In Section 4, we describe experimental evaluations. Finally, we summarize this paper in Section 5.

2. Conventional One-to-Many EVC System

2.1. Eigenvoice Gaussian Mixture Model (EV-GMM)

We use $2D$-dimensional acoustic features, $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ (source speaker's) and $\boldsymbol{Y}_t^{(s)} = \left[\boldsymbol{y}_t^{(s)\top}, \Delta\boldsymbol{y}_t^{(s)\top}\right]^\top$ (the s^{th} target speaker's), consisting of D-dimensional static and dynamic features, where $\top$ denotes transposition of the vector. Joint probability density of time-aligned source and target features $\boldsymbol{Z}_t^{(s)} = \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)\top}\right]^\top$ determined by DTW is modeled with EV-GMM as follows:

$$P\left(\boldsymbol{Z}_t^{(s)}|\lambda^{(EV)}\right) = \sum_{i=1}^{M}\alpha_i\mathcal{N}\left(\boldsymbol{Z}_t^{(s)};\boldsymbol{\mu}_i^{(Z)},\boldsymbol{\Sigma}_i^{(ZZ)}\right), \quad (1)$$

$$\boldsymbol{\mu}_i^{(Z)} = \begin{bmatrix}\boldsymbol{\mu}_i^{(X)} \\ \boldsymbol{B}_i\boldsymbol{w}_s + \boldsymbol{b}_i^{(0)}\end{bmatrix}, \quad (2)$$

Accepted after peer review of full paper

$$\boldsymbol{\Sigma}_i^{(ZZ)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(YX)} & \boldsymbol{\Sigma}_i^{(YY)} \end{bmatrix}, \quad (3)$$

where $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes Gaussian distribution with mean vector $\boldsymbol{\mu}$ and diagonal covariance matrix $\boldsymbol{\Sigma}$. In EV-GMM $\lambda^{(EV)}$, a target mean vector is modeled by the bias vector $\boldsymbol{b}_i^{(0)}$, representative vectors $\boldsymbol{B}_i = \left[\boldsymbol{b}_i^{(1)}, \boldsymbol{b}_i^{(2)}, \cdots, \boldsymbol{b}_i^{(J)}\right]$ and the weight vector $\boldsymbol{w}_s$. EV-GMM models arbitrary target speaker's individualities by setting $\boldsymbol{w}_s$ to appropriate values. The other parameters such as mixture component weights, source mean vectors, bias vectors, representative vectors and covariance matrices are tied for every target speaker.

2.2. Training of EV-GMM

We train the EV-GMM based on principal component analysis (PCA) as follows.

1. TI-GMM $\lambda^{(0)}$ is estimated with multiple parallel data sets consisting of utterance-pairs of the source speaker and multiple pre-stored target speakers.
2. The s^{th} target dependent GMM (TD-GMM) $\lambda^{(s)}$ is trained by updating only target mean vectors of $\lambda^{(0)}$.
3. Bias vector $\boldsymbol{b}_i^{(0)}$ and representative vectors $\boldsymbol{B}_i$ are extracted from the supervectors, which are 2DM-dimensional concatenated target mean vectors, by PCA.

2.3. Adaptation and Conversion

Figure 1 shows adaptation and conversion processes of the conventional one-to-many EVC system.

In the adaptation process, the EV-GMM is adapted to a desired target speaker by estimating $\boldsymbol{w}$ in the sense of maximum likelihood as described in [4]. In order to perform the unsupervised adaptation using only the target speaker's voice, we determine the optimal weight vector $\hat{\boldsymbol{w}}$ so that the likelihood of marginal distribution is maximized as follows [3]:

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \int P\left(\boldsymbol{X}, \boldsymbol{Y}^{(tar)} | \lambda^{(EV)}\right) d\boldsymbol{X}, \quad (4)$$

where, $\boldsymbol{Y}^{(tar)}$ is a time sequence of the given target features.

In the conversion process, we use the conversion method based on maximum likelihood estimation (MLE) considering dynamic features [7] for spectral features. Let a time sequence of the source features and that of the target features be $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ and $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$, respectively. Converted static feature vectors $\hat{\boldsymbol{y}} = \left[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top\right]^\top$ are obtained as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda^{(EV)}\right), \quad (5)$$
$$\text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y},$$

where $\boldsymbol{W}$ denotes the matrix to extend the static feature sequence to the static and dynamic feature sequence, and $\hat{\boldsymbol{m}}$ shows the optimum mixture sequence determined so that $P(\boldsymbol{m}|\boldsymbol{X}, \lambda^{(EV)})$ is maximized. Source speaker's F_0 is converted into that of the desired target speaker by simple linear conversion with mean and variance of each speaker.

In the synthesis process, a simple excitation is generated based on the converted F_0 by selecting the phase-manipulated pulse train for voiced segments or white noise for unvoiced segments. And then, the converted speech is synthesized with the generated excitation and the converted spectral sequence.

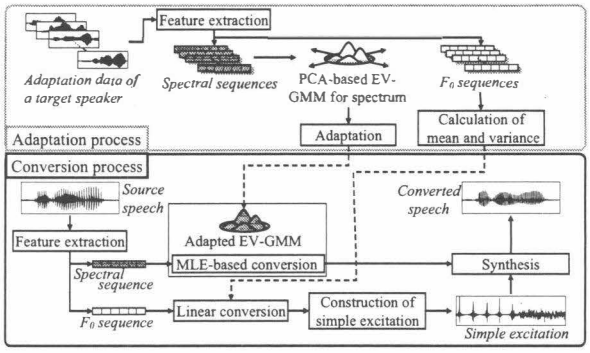

Figure 1: Adaptation and conversion process of conventional one-to-many EVC system.

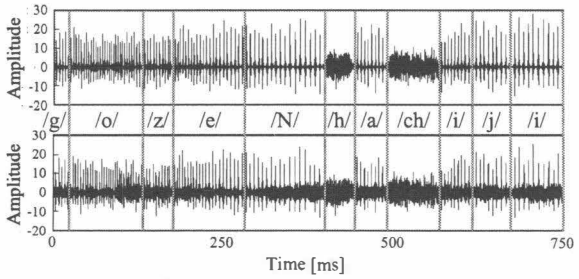

Figure 2: Example of excitation signals for a sentence fragment "g o z e N h a ch i j i": (Top) simple excitation and (Bottom) STRAIGHT mixed excitation.

3. Improved One-to-Many EVC System

In order to improve the converted speech quality of EVC, we apply STRAIGHT mixed excitation (ST-ME), global variance (GV) and speaker adaptive training (SAT) to the conventional system.

3.1. STRAIGHT Mixed Excitation (ST-ME)

STRAIGHT mixed excitation is defined as the frequency-dependent weighted sum of white noise and the phase-manipulated pulse train. The time-varying weight is determined based on an aperiodic component, which represents the degree of the noise component in each frequency bin [8], at each time frame. Figure 2 shows an example of excitation signals. The simple excitation employs the phase-manipulated pulse train for voiced segments or white noise for the unvoiced one. On the other hand, in ST-ME, the pulse and noise components are mixed based on aperiodic components for generating a more natural excitation signal.

We have proposed the conversion of aperiodic components based on a GMM to apply ST-ME to VC and have demonstrated its effectiveness in the conventional VC framework [9]. In this paper, aperiodic components are modeled by EV-GMM for applying ST-ME to the one-to-many EVC system.

3.2. MLE-based Conversion Considering GV

The GV is calculated as variances of the feature vectors over a time sequence. It has been reported that the conversion performance is dramatically improved by considering the GV of the converted feature vectors in the conversion process [7]. In this paper, the GV is calculated utterance by utterance.

In the proposed system, GV is modeled by an eigenvoice-

based single Gaussian distribution (EV-SG) as follows:

$$P\left(\boldsymbol{v}_{\boldsymbol{y}}^{(s)}|\lambda_v^{(EV)}\right) = \mathcal{N}\left(\boldsymbol{v}_{\boldsymbol{y}}^{(s)}; \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\right), \quad (6)$$

$$\boldsymbol{\mu}^{(v)} = \boldsymbol{B}^{(v)}\boldsymbol{w}_s^{(v)} + \boldsymbol{b}_0^{(v)}, \quad (7)$$

where $\boldsymbol{v}_{\boldsymbol{y}}^{(s)}$ denotes the GV of s^{th} target speaker and model parameter $\lambda_v^{(EV)}$ includes mean vector $\boldsymbol{\mu}^{(v)}$ and covariance matrix $\boldsymbol{\Sigma}^{(vv)}$. EV-SG is trained using all of GV vectors extracted from individual utterances of every pre-stored target speaker.

In the conversion process, the converted speech features are determined by maximizing with respect to $\boldsymbol{y}$ as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P\left(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda^{(EV)}\right)^{\omega} P\left(\boldsymbol{v}_{\boldsymbol{y}}|\lambda_v^{(EV)}\right), \quad (8)$$

where ω is a parameter for controlling the two likelihoods. In this paper, the value of ω is $\frac{1}{2T}$, which is set to the ratio of the number of dimensions between spectral features and GV.

3.3. Speaker Adaptive Training (SAT)

The acoustic variations among the pre-stored target speakers in the EV-GMM are effectively reduced by applying SAT. The resulting EV-GMM is called a canonical EV-GMM, which is trained by maximizing a total likelihood of the adapted models to individual pre-stored target speakers as follows:

$$\hat{\lambda}^{(EV)}(\hat{\boldsymbol{w}}_1^S) = \arg\max_{\lambda} \prod_{s=1}^{S}\prod_{t=1}^{T_s} P\left(\boldsymbol{Z}_t^{(s)}|\lambda^{(EV)}(\boldsymbol{w}_s)\right), \quad (9)$$

where $\lambda^{(EV)}(\boldsymbol{w}_s)$ denotes the adapted model for the s^{th} pre-stored target speaker with the weight vector $\boldsymbol{w}_s$. SAT estimates both canonical EV-GMM parameters $\hat{\lambda}^{(EV)}$ and a set of weight vectors $\hat{\boldsymbol{w}}_1^S = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_S)$ for individual pre-stored target speakers.

SAT is employed for training the EV-GMM for the spectral features, the EV-GMM for the aperiodic components, and the EV-SG for GV. EV-SG parameters are updated by directly maximizing the likelihood shown in Eq. (9). On the other hand, because the EV-GMM has hidden variables, its parameters are updated using the EM algorithm by maximizing the following auxiliary function:

$$Q\left(\lambda^{(EV)}(\boldsymbol{w}_1^S), \hat{\lambda}^{(EV)}(\hat{\boldsymbol{w}}_1^S)\right) = \sum_{s=1}^{S}\sum_{i=1}^{M} \bar{\gamma}_i^{(s)} \log P\left(\boldsymbol{Z}^{(s)}, m_i|\hat{\lambda}^{(EV)}(\hat{\boldsymbol{w}}_s)\right), \quad (10)$$

where,

$$\bar{\gamma}_i^{(s)} = \sum_{t=1}^{T_s} P(m_i|\boldsymbol{Z}_t^{(s)}, \lambda^{(EV)}(\boldsymbol{w}_s)). \quad (11)$$

3.4. Adaptation and Conversion

Figure 3 shows the proposed one-to-many EVC system. In the adaptation process, spectral features, aperiodic components, and GVs are extracted from the adaptation data. And then, each canonical model is adapted independently.

In the conversion process, spectral features are converted by MLE-based conversion considering GV. On the other hand, aperiodic components are converted by the conventional MLE-based conversion without GV because GV is not so effective for the aperiodic components.

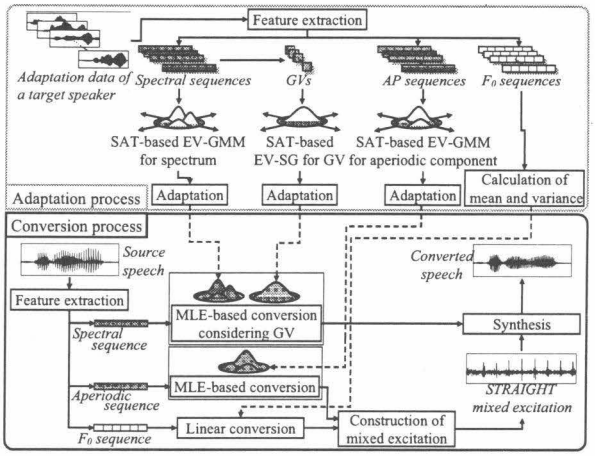

Figure 3: Adaptation and conversion process of proposed one-to-many EVC system.

4. Experimental Evaluation

We objectively and subjectively evaluate the performance of the proposed one-to-many EVC system compared with that of the conventional one.

4.1. Experimental Conditions

We used one male source speaker and 160 pre-stored target speakers composed of 80 male and 80 female speakers [12]. Each speaker uttered 50 phoneme-balanced sentences as shown in [3].

In the evaluations, we used 10 target speakers consisting of five male and five female speakers, which were not included in the pre-stored target speakers. We used 1 to 32 utterances for the adaptation, and 21 utterances for the evaluations.

We used 24-dimensional mel-cepstrum analyzed by STRAIGHT [6] as spectral features and aperiodic components that were averaged on five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6 and 6 to 8 kHz). The number of representative vectors was 159 for mel-cepstrum, 64 for aperiodic components and 4 for GV, respectively. The number of mixture components was 128 for spectral features and 64 for the aperiodic features, respectively.

4.2. Objective Evaluations

It has been reported that SAT causes significant improvements of the adaptation performance of the EV-GMM for the spectral feature [11]. In this paper, we further evaluated the effectiveness of SAT in the adaptation of the EV-GMM for the aperiodic components and that of the EV-SG for the GV. The likelihood of the adapted conversion model for the evaluation data was used as an evaluation measure.

Figure 4 shows log-scaled likelihoods of the PCA-/SAT-based EV-GMM for aperiodic components and those of the PCA-/SAT-based EV-SG for GV. We can see that SAT causes significant improvements of the adaptation performance of both the EV-GM for the aperiodic components and the EV-SG for the GV. Therefore, a combination of these techniques works reasonably well.

4.3. Subjective Evaluations

We conducted a preference test on speech quality and an XAB test on conversion accuracy for speaker individuality. In our

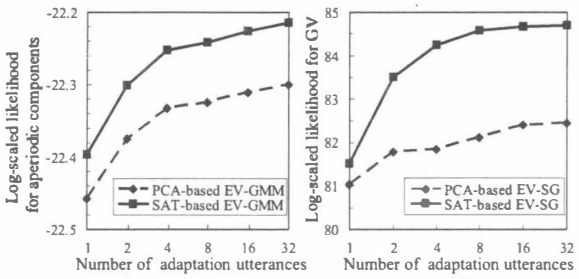

Figure 4: Log-scaled likelihood as a function of the number of adaptation utterances: (a) EV-GMM for aperiodic components and (b) EV-SG for GV.

preliminary experiments, it was shown that a combination of ST-ME and GV causes significant improvements of the conversion performance as similarly as reported in HMM-based speech synthesis [13]. In order to further demonstrate the effectiveness of a combination of these two techniques and SAT, we evaluated the following three types of converted speech:

- converted speech of the conventional one-to-many EVC system;
- converted speech of the proposed system using ST-ME, MLE-based conversion considering GV and PCA-based GMM (ST-ME+GV);
- converted speech of the proposed system using ST-ME, MLE-based conversion considering GV and SAT-based GMM (ST-ME+GV+SAT).

In the preference test, a pair of two different types of the converted speech was presented to listeners, and then they were asked which voice sounded better. In the XAB test, a pair of two different types of the converted speech were presented to them after presenting the target speech as a reference. And then, they were asked which voice sounded more similar to the reference. Each listener evaluated every pair-combination of all types of the converted speech. The number of listeners was seven.

Figure 5 shows the results. In the speech quality test, the proposed system with ST-ME and GV significantly outperforms the conventional system. This is because the buzzing sound was reduced by ST-ME and the over-smoothing effect was alleviated GV. Moreover, additional significant improvements of the converted speech quality were caused by further introducing SAT to the proposed system. In the evaluation of the conversion accuracy for speaker individuality, the proposed system also causes remarkable improvements although the additional improvement caused by SAT is marginal.

These results demonstrate that the proposed system causes dramatic improvements in the performance of the one-to-many EVC system.

5. Conclusions

To improve the converted speech quality of the one-to-many eigenvoice conversion system, we applied STRAIGHT mixed excitation, conversion algorithm considering global variance and speaker adaptive training of the eigenvoice Gaussian mixture model. Results of objective and subjective evaluations demonstrated that the proposed system considerably outperforms the conventional one.

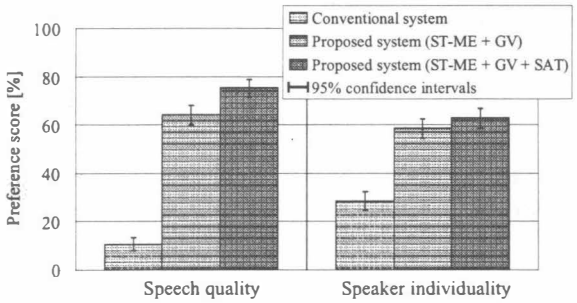

Figure 5: Results of subjective evaluations.

6. Acknowledgements

This work was supported in part by MEXT Grant-in-Aid for Young Scientists (A).

7. References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn. (E)*, Vol. 11, No. 2, pp. 71–76, 1990.

[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, 1998.

[3] T. Toda, Y. Ohtani, K. Shikano, "Eigenvoice conversion based on Gaussian mixture model", *Proc. ICSLP*, pp. 2446–2449, September, 2006.

[4] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space" *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 6, pp. 695–707, 2000.

[5] T. Toda, Y. Ohtani, K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices", *Proc. ICASSP*, April, 2007.

[6] H. Kawahara, I. Masuda-Katsuse and A.de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and instantaneous frequency-based F_0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.

[7] T. Toda, A.W. Black, K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, November, 2007.

[8] H. Kawahara, Jo Estill and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *MAVEBA 2001*, Sept. 13-15, Firentze Italy, 2001.

[9] Y. Ohtani, T. Toda, H. Saruwatari, K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," *Proc. ICSLP*, pp. 2266–2269, September, 2006.

[10] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, "A compact model for speaker adaptive training" *Proc. ICSLP*, vol.2, pp. 1137–1140, 1996.

[11] Y. Ohtani, T. Toda, H. Saruwatari, K. Shikano, "Speaker adaptive training for one-to-many eigenvoice conversion based on Gaussian mixture model," *Proc. Interspeech 2007*, pp. 1981–1084, August, 2007.

[12] JNAS: Japanese newspaper article sentences. *http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html*

[13] H. Zen, T. Toda, M. Nakamura and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans*, vol.E90-D, no.1, Jan 2007