ONE-TO-MANY AND MANY-TO-ONE VOICE CONVERSION BASED ON EIGENVOICES

Tomoki Toda, Yamato Ohtani, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology, Japan

tomoki@is.naist.jp

ABSTRACT

This paper describes two flexible frameworks of voice conversion (VC), i.e., one-to-many VC and many-to-one VC. One-to-many VC realizes the conversion from a user's voice as a source to arbitrary target speakers' ones and many-to-one VC realizes the conversion vice versa. We apply eigenvoice conversion (EVC) to both VC frameworks. Using multiple parallel data sets consisting of utterance-pairs of the user and multiple pre-stored speakers, an eigenvoice Gaussian mixture model (EV-GMM) is trained in advance. Unsupervised adaptation of the EV-GMM is available to construct the conversion model for arbitrary target speakers in one-to-many VC or arbitrary source speakers in many-to-one VC using only a small amount of their speech data. Results of various experimental evaluations demonstrate the effectiveness of the proposed VC frameworks.

Index Terms— Speech synthesis, voice conversion, eigenvoice, one-to-many, many-to-one

1. INTRODUCTION

Voice conversion (VC) is a technique for converting a certain speaker's voice into another speaker's voice [1]. One of typical VC frameworks is a statistical approach using a conversion model such as a Gaussian mixture model (GMM) [2] for representing joint probability density of source and target acoustics [3]. The conversion model is basically trained in advance using a parallel data set consisting of utterance-pairs of the source and the target speakers. It successfully converts any speech sample of the source speaker into that of the target speaker. It is no doubtful that VC is a useful technique and there are a lot of applications of using it. However, it is doubtful whether the VC framework requiring parallel data is acceptable for real users. It is more convenient to enable the user to convert own voices into the desired voices even if he doesn't obtain any speech samples of the target. VC from arbitrary speakers into the user also seems useful for generating various languages as if uttered by the user. In order to realize handy VC applications, it is essential to make the VC framework more flexible.

One promising approach for flexibly constructing the conversion model for the desired speaker-pair is to exploit voices of other speakers as a prior knowledge. Mouchtaris et al. [4] proposed a non-parallel training method based on maximum likelihood constrained adaptation of a GMM trained with an existing parallel data set of a different speaker-pair. Iwahashi and Sagisaka [5] proposed a conversion method based on speaker interpolation with multiple pre-stored speakers' voices. By integrating those two ideas such as adapting the conversion model and using many pre-stored speaker's voices into a unified VC framework, Toda et al. [6] proposed eigenvoice conversion (EVC) based on eigenvoices that was originally proposed as a speaker adaptation technique in speech recognition [7].

This paper describes one-to-many VC and many-to-one VC as flexible VC frameworks. One-to-many VC realizes the conversion from a source speaker's voice into arbitrary target speakers' ones and many-to-one VC realizes the conversion vice versa. The effectiveness of EVC in one-to-many VC has been reported in [6]. It is expected that EVC works in many-to-one VC as well. This paper applies EVC into not only one-to-many VC but also many-to-one VC. Various experimental evaluations are conducted for demonstrating the effectiveness of the proposed VC frameworks based on EVC.

The paper is organized as follows. **Section 2** describes frameworks of one-to-many VC and many-to-one VC. **Section 3** describes EVC. **Section 4** describes experimental evaluations. Finally, we summarize this paper in **Section 5**.

Thanks to MIC SCOPE-S and MEXT e-Society leading project for supporting this research in part.

2. ONE-TO-MANY VC AND MANY-TO-ONE VC

Frameworks of one-to-many VC and many-to-one VC consist of two main processes, i.e., 1) training and 2) adaptation and conversion. The training process employs multiple parallel data sets consisting of utterance-pairs of the source speaker, i.e., a user and many pre-stored target speakers in one-to-many VC or vice versa in many-to-one VC. Namely, those frameworks assume that the user utters a prepared sentence set only once. Voices of many pre-stored speakers uttering the same sentence set need to be recorded in advance. Around 50 phonetically balanced sentences would work properly as the sentence set. Those parallel data sets cause the conversion model capturing the correlation between the user's voice and many speakers' voices, which is effectively used as a prior knowledge in the model adaptation.

The conversion model is adapted to arbitrary target speakers in one-to-many VC or arbitrary source speakers in many-to-one VC using only their speech data without any linguistic restrictions. And then, VC is performed with the adapted conversion model. Therefore, we don't have to newly prepare a parallel data set between the user and the arbitrary speakers. Moreover, the amount of adaptation data is considerably reduced by exploiting the prior knowledge.

3. EIGENVOICE CONVERSION (EVC)

This section describes a framework of EVC in many-to-one VC. It is straightforward to apply EVC to one-to-many VC by replacing the source and the target each other as described in [6].

3.1. Eigenvoice GMM (EV-GMM)

We employ $2D$-dimensional acoustic features $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ (source speaker's) and $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ (target speaker's) consisting of D-dimensional static and dynamic features, where $\top$ denotes transposition of the vector. An EV-GMM represents the joint probability density as follows:

$$p(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}^{(EV)}) = \sum_{i=1}^{M} \alpha_i N(\boldsymbol{X}_t, \boldsymbol{Y}_t; \boldsymbol{\mu}_i^{(X,Y)}, \boldsymbol{\Sigma}_i^{(X,Y)}),$$

$$\boldsymbol{\mu}_i^{(X,Y)} = \begin{bmatrix} \boldsymbol{B}_i^{(X)}\boldsymbol{w} + \boldsymbol{b}_i^{(X)}(0) \\ \boldsymbol{\mu}_i^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_i^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(YX)} & \boldsymbol{\Sigma}_i^{(YY)} \end{bmatrix},$$

where $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ shows the normal distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The i^{th} mixture weight is α_i. The total number of mixtures is M. In many-to-one VC, the source mean vector for the i^{th} mixture is represented as a linear combination of a bias vector $\boldsymbol{b}_i^{(X)}(0)$ and representative vectors $\boldsymbol{B}_i^{(X)} = [\boldsymbol{b}_i^{(X)}(1), \cdots, \boldsymbol{b}_i^{(X)}(J)]$. The number of representative vectors is J. The source speaker individuality is controlled with only the J-dimensional weight vector $\boldsymbol{w} = [w(1), \cdots, w(J)]^\top$. This paper employs diagonal covariance matrices.

3.2. Training of EV-GMM

Firstly, a source independent GMM is trained using all of the multiple parallel data sets simultaneously. And then, each source dependent GMM is trained by updating only source mean vectors of the source independent GMM using each of the multiple parallel data sets. As a source dependent parameter, a supervector for each pre-stored source speaker is constructed by concatenating the source mean vectors of each of the source dependent GMMs. The bias and representative vectors, i.e., eigenvectors are determined with principal component analysis (PCA) for all source speakers' supervectors. Finally, the EV-GMM is constructed with the resulting bias and representative vectors and parameters of the source independent GMM.

It is essential to model phonemic features and speaker dependent features separately with the EV-GMM. In order to do it, the correspondence of each mixture into a phonemic space should be the same in the every source dependent GMM. Because the source dependent GMMs are trained while fixing probability density function on the target space as mentioned above, the correspondences between individual mixtures and the target phonemic spaces are kept consistent in all of the GMMs. Moreover, because the phonemic space of the target is aligned to the same phonemic space of the source due to parallel data, the every GMM has the consistent correspondences of individual mixtures into both the source and the target phonemic spaces. Consequently, a subspace representing speaker dependent features is constructed with the resulting supervectors of which variations capture not phonemic differences but differences of the source speaker individuality.

3.3. Unsupervised Adaptation of EV-GMM

The EV-GMM is adapted for arbitrary speakers by estimating the optimum weight vector for given their speech samples without any linguistic information. For example, in many-to-one VC, the weight vector is estimated so that a likelihood of the marginal distribution for a time sequence of the given source features $\boldsymbol{X}^{(tar)}$ is maximized [6] as follows:

$$\hat{\boldsymbol{w}} = \arg\max \int p(\boldsymbol{X}^{(tar)}, \boldsymbol{Y} | \boldsymbol{\lambda}^{(EV)}) d\boldsymbol{Y}.$$

Because the probability density is modeled with a GMM, EM algorithm is used for the estimation. This paper employs the speaker independent GMM for performing the first E-step process.

In one-to-many VC, EVC realizes the converted speech with various voice characteristics by manually manipulating the weight vector. It is possible to realize the weight vector modifying various speech acoustics simultaneously by using supervectors including target dependent parameters of the conversion models for the individual speech acoustics.

3.4. Conversion with EV-GMM

The spectral conversion is straightforwardly performed with the adapted EV-GMM. This paper employs the conversion method based on maximum likelihood estimation considering dynamic features [8].

4. EXPERIMENTAL EVALUATIONS

4.1. Experimental conditions

In order to train EV-GMMs in one-to-many VC and many-to-one VC, we used 160 speakers consisting of 80 male and 80 female speakers as the pre-stored speakers. These speakers were included in Japanese Newspaper Article Sentences (JNAS) database [9]. Each of them uttered a set of phonetically balanced 50 sentences. We used a male speaker not included in JNAS as the source speaker in one-to-many VC or the target speaker in many-to-one VC, who uttered the same sentence sets as uttered by the pre-stored speakers. More detail conditions are described in [6].

In order to evaluate the performance of unsupervised adaptation of the EV-GMM, we compared EVC with the conventional VC. We used ten test speakers consisting of five male and five female speakers who were not included in the pre-stored speakers. Those speakers uttered 53 sentences that were also not included in the pre-stored data sets. The number of adaptation sentences was varied from 1 to 32. The remaining 21 sentences were used for evaluations. The conventional VC trained GMMs for the conversion between individual speaker-pairs using their parallel training data sets.

In order to demonstrate the effectiveness of the manual weight control in one-to-many VC, we investigated changes of the converted parameters when varying the weight setting. For controlling various speech acoustics, supervectors were constructed by concatenating not only target mean vectors of GMMs for the spectral conversion but also various parameters such as target mean vectors of GMMs for the aperiodic conversion, a mean vector of global variance (GV) of spectral features [8], and parameters for the F_0 conversion.

We used mel-cepstrum as a spectral feature. The first through 24^{th} mel-cepstral coefficients were extracted from 16 kHz sampling speech data. The STRAIGHT analysis method [10] was employed for the spectral extraction. A simple linear conversion with means and standard deviations of log-scaled F_0 of the source and the target speakers was employed in the F_0 conversion.

4.2. Objective evaluations

4.2.1. Unsupervised adaptation in one-to-many VC

Figure 1 shows mel-cepstral distortion when varying the number of target adaptation sentences and the number of representative vectors of the EV-GMM. When the number of representative vectors is small, an increase of the number of adaptation sentences doesn't cause any improvements of the spectral conversion-accuracy due to the small number of free parameters to be adapted. The conversion-accuracy is improved by increasing the number of representative vectors. Although it might be possible that the conversion-accuracy is degraded due to the over-training when using a large number of representative vectors and a small number of adaptation sentences, such a trend is not observed even if using all of representative vectors, i.e., 159 vectors. Therefore, all of representative vectors were used in the following experiments. When using a larger number of pre-stored speakers, it seems necessary to determine the number of representative vectors appropriately according to the amount of adaptation data because the number of available representative vectors also increases.

Figure 2 shows mel-cepstral distortion as a function of the number of mixtures when varying the number of target adaptation sentences (or training sentence-pairs in the conventional VC). It also shows a result of the conversion with the target independent GMM ("TI-GMM"). In conventional VC, the conversion-accuracy is improved by increasing the number of mixtures because the joint prob-

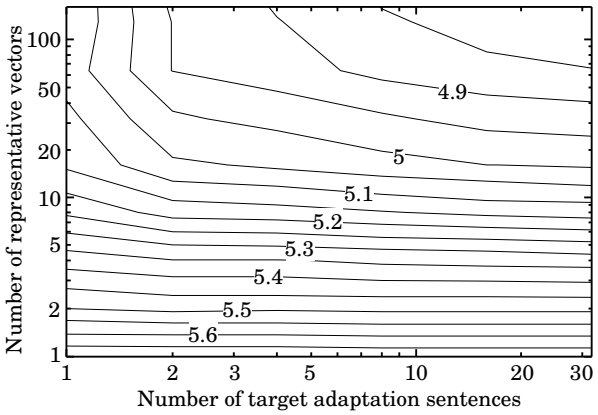


Fig. 1. Mel-cepstral distortion on each combination of the number of representative vectors and the number of target adaptation sentences in one-to-many VC. The number of mixtures is set to 128.

ability density is accurately represented with more complex models. However, by further increasing it, the conversion-accuracy starts to be degraded due to the over-training. Consequently, as the amount of training data is larger, the optimum number of mixtures increases and the conversion-accuracy is improved.

One-to-many VC is a conversion process from a single input feature into multiple output features. It is reasonable that the target independent GMM doesn't work because that model just converts the source features into the average features among pre-stored target speakers.

EVC works much better than the conventional VC when using the small amount of target adaptation data because information of pre-stored target speakers is effectively used as a prior knowledge. The conversion-accuracy is improved by increasing the number of mixtures. The over-training effect is not observed even if using 512 mixtures because the amount of training data of the EV-GMM is enough large. Although a larger amount of adaptation data also causes improvements of the conversion-accuracy, those improvements are not so large when increasing over two adaptation sentences because of the limited number of adapted parameters.

4.2.2. Unsupervised adaptation in many-to-one VC

Figure 3 shows mel-cepstral distortion as a function of the number of mixtures in many-to-one VC when varying the number of source adaptation sentences (or training sentence-pairs). It also shows a result of the conversion with the source independent GMM ("SI-GMM"). The conventional VC has the same tendencies as shown in one-to-many VC.

We can observe completely different results between many-to-one VC and one-to-many VC in the speaker independent GMMs, i.e., TI-GMM and SI-GMM. The source independent GMM works as the conversion model in many-to-one VC. Many-to-one VC is the conversion process from multiple input features into a single output feature. Therefore, the conversion reasonably works if an input feature space including characteristics of various speakers is modeled precisely. In fact, the conversion-accuracy is improved by increasing the number of mixtures because a more complex model is effectively used for representing feature spaces of various source speakers. However, different speakers have different acoustics for the same phonemes in general. Therefore, even if acoustic features are similar, they don't always capture the same phonemic features. The source independent GMM may cause the conversion between different phonemic spaces.

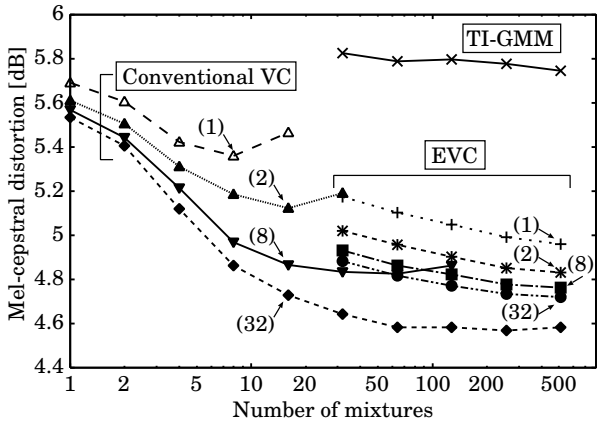


Fig. 2. Mel-cepstral distortion as a function of the number of mixtures in one-to-many VC. We show mean distortions over 10 target speakers. The number in each bracket shows the number of target adaptation sentences (or training sentence-pairs).

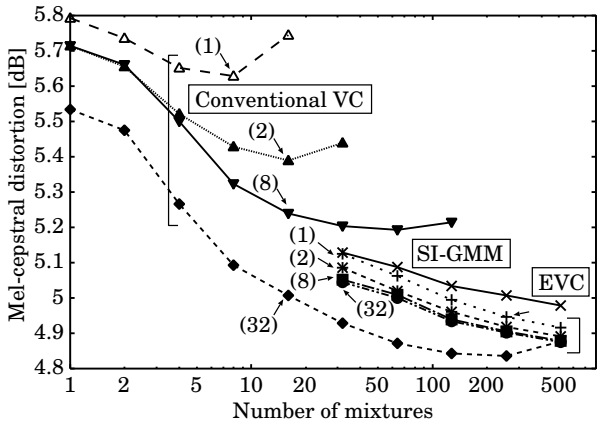
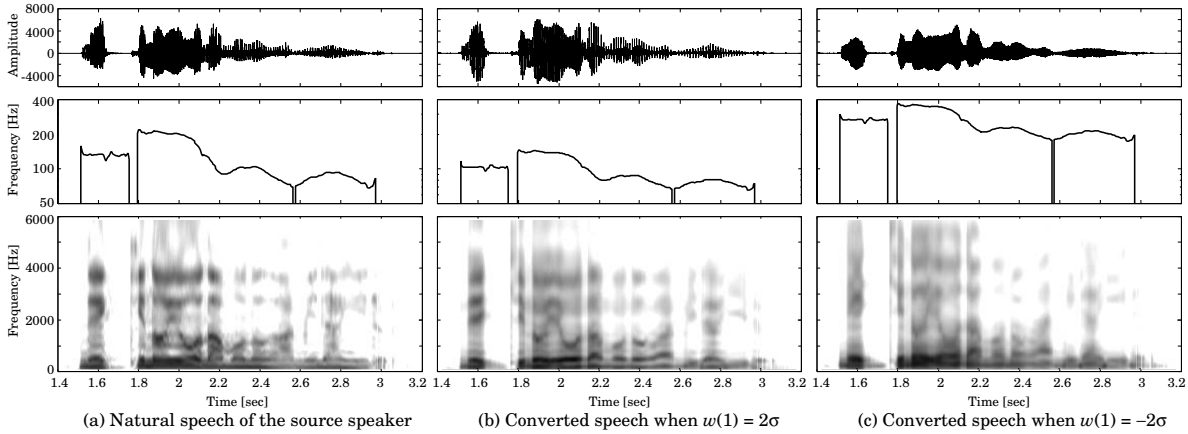


Fig. 3. Mel-cepstral distortion as a function of the number of mixtures in many-to-one VC. We show mean distortions over 10 source speakers. The number in each bracket shows the number of source adaptation sentences (or training sentence-pairs).

On the other hand, the EV-GMM separately models phonemic features and speaker dependent features on the acoustic space with the subspace efficiently representing only the speaker individuality. Because it is adapted to an arbitrary source speaker while keeping a correspondence of phonemic spaces between the source and the target features, the conversion-accuracy is better than that of the source independent GMM. As described in one-to-many VC, EVC in many-to-one VC also works much better than the conventional VC when using a small amount of the source adaptation data.

4.2.3. Manual weight control in one-to-many VC

Figure 4 shows an example of acoustics of an input voice and converted voices when modifying only the first coefficient of the weight vector while keeping the others zero. Every acoustic feature effectively changes by manipulating only the single parameter. Specifically, an increase of the first weight coefficient causes a higher F_0 contour and formant shifts toward higher frequencies. Results of an informal listening test showed that the first representative vector seems to capture gender information of the target speakers. This trend has also been found in HMM-based speech synthesis based on eigenvoices [11].



(a) Natural speech of the source speaker (b) Converted speech when $w(1) = 2\sigma$ (c) Converted speech when $w(1) = -2\sigma$

Fig. 4. An example of waveforms, F_0 contours, and spectrograms of individual voices at a sentence fragment "n e Q k i n o y oo n a m o n o g a m i n a g i r u." Only the first weight coefficient $w(1)$ is manually varied in one-to-many VC, where σ is the first principal component.

4.3. Subjective evaluations

We conducted an opinion test and an XAB test for evaluating the performance of EVC compared with the conventional VC in one-to-many VC. In the opinion test, listeners evaluated speech quality of the converted voices using a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). In the XAB test (X: target speech, A and B: converted voices with EVC and the conventional VC), listeners were asked which converted speech sounded more similar to the target speech. The number of listeners was five. The number of mixtures of the EV-GMM was set to 512. On the other hand, the number of mixtures in the conventional VC was set to the optimum value for each target speaker and the each number of training sentences in the sense of the spectral conversion-accuracy.

The result of the opinion test is shown in **Fig. 5**. EVC outperforms the conventional VC when using the small amount of target adaptation data. In the conventional VC, speech quality is obviously improved by increasing the amount of training data. EVC successfully synthesizes the converted speech with equal quality to that of the conventional VC even if using 32 target sentences. Note that speech quality of the converted voices is insufficient because we didn't use the conversion method considering GV [8] in this test. It is expected to make MOS around 1.0 larger by considering GV as reported in [8].

The result of the preference test is also shown in **Fig. 5**. It is observed that EVC outperforms the conventional VC when using 2 target sentences. Even if using 16 target sentences, the performance of EVC is comparable to that of the conventional VC. Although the conversion-accuracy of EVC is slightly inferior to that of the conventional VC when using 32 target sentences, EVC still has an advantage of allowing unsupervised adaptation.

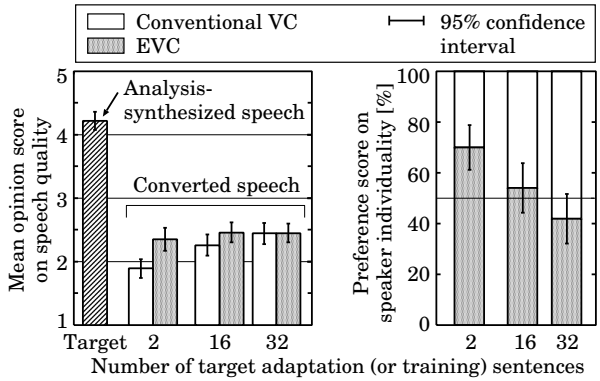


Fig. 5. Results of subjective evaluations in one-to-many VC.

5. CONCLUSIONS

This paper described flexible frameworks of voice conversion (VC) such as one-to-many VC and many-to-one VC. Eigenvoice conversion (EVC) was applied to both frameworks. Results of various experimental evaluations demonstrated the effectiveness of the proposed VC frameworks. We will apply EVC to many-to-many VC.

6. REFERENCES

[1] H. Kuwabara and Y. Sagisaka. Acoustic characteristics of speaker individuality: control and conversion. *Speech Communication*, Vol. 16, No. 2, pp. 165–173, 1995.

[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] A. Kain and M.W. Macon. Spectral voice conversion or text-to-speech synthesis. *Proc. ICASSP*, pp. 285–288, Seattle, USA, May 1998.

[4] A. Mouchtaris, J.V. der Spiegel, and P. Mueller. Non-parallel training for voice conversion by maximum likelihood constrained adaptation. *Proc. ICASSP*, Vol. 1, pp. 1–4, Montreal, Canada, May 2004.

[5] N. Iwahashi and Y. Sagisaka. Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks. *Speech Communication*, Vol. 16, No. 2, pp. 139–151, 1995.

[6] T. Toda, Y. Ohtani, and K. Shikano. Eigenvoice conversion based on Gaussian mixture model. *Proc. ICSLP*, pp. 2446–2449, Pittsburgh, USA, Sep. 2006.

[7] 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.

[8] T. Toda, A.W. Black, and K. Tokuda. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. *Proc. ICASSP*, Vol. 1, pp. 9–12, Philadelphia, USA, Mar. 2005.

[9] JNAS: Japanese Newspaper Article Sentences. *http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html*

[10] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an 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.

[11] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. KItamura. Eigenvoices for HMM-based speech synthesis. *Proc. ICSLP*, Vol. 1, pp. 1269–1272, Denver, USA, Sep. 2002.