

MULTIMODAL VOICE CONVERSION USING NON-NEGATIVE MATRIX FACTORIZATION IN NOISY ENVIRONMENTS

Kenta MASAKA, Ryo AIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI

Graduate School of System Informatics, Kobe University
1-1, Rokkodai, Nada, Kobe, 6578501, Japan

ABSTRACT

This paper presents a multimodal voice conversion (VC) method for noisy environments. In our previous NMF-based VC method, source exemplars and target exemplars are extracted from parallel training data, in which the same texts are uttered by the source and target speakers. The input source signal is then decomposed into source exemplars, noise exemplars obtained from the input signal, and their weights. Then, the converted speech is constructed from the target exemplars and the weights related to the source exemplars. In this paper, we propose a multimodal VC that improves the noise robustness in our NMF-based VC method. By using the joint audio-visual features as source features, the performance of VC is improved compared to a previous audio-input NMF-based VC method. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional Gaussian Mixture Model (GMM)-based method.

Index Terms— voice conversion, multimodal, image features, non-negative matrix factorization, noise robustness

1. INTRODUCTION

Background noise is an unavoidable factor in speech processing. In the task of automatic speech recognition (ASR), one problem is that the recognition performance remarkably decreases under noisy environments, and it becomes a significant problem seeking to develop a practical use of ASR. The same problem occurs in voice conversion, which can modify nonlinguistic information, such as voice characteristics, while keeping linguistic information unchanged. The noise in the input signal is not only output with the converted signal, but may also degrade the conversion performance itself due to unexpected mapping of source features. To address the problem, in this paper, we propose a noise-robust VC method that is based on sparse representations.

Approaches based on sparse representations have gained interest in a broad range of signal processing in recent years. Non-negative matrix factorization (NMF) [1], which is based on the idea of sparse representations is a well-known approach for source separation and speech enhancement [2, 3]. In these approaches, the observed signal is represented by a linear combination of a small number of atoms, such as the exemplar and basis of NMF. In some approaches for source separation, the atoms are grouped for each source, and the mixed signals are expressed with a sparse representation of these atoms. By using only the weights of the atoms related to the target signal, the target signal can be reconstructed. Gemmeke et al. [4] proposed an exemplar-based method for noise-robust speech recognition using NMF. In that method, the observed speech is decomposed into the speech atoms, noise atoms, and their weights. Then

the weights of the speech atoms are used as phonetic scores (instead of the likelihoods of hidden Markov models) for speech recognition.

In [5], we discussed a noise-robust voice conversion (VC) technique using NMF. In that method, source exemplars and target exemplars are extracted from the parallel training data, in which the same texts are uttered by the source and target speakers. Also, the noise exemplars are extracted from the before- and after-utterance sections in an observed signal. For this reason, no training processes related to noise signals are required. The input source signal is expressed with a sparse representation of the source exemplars and noise exemplars. Only the weights related to the source exemplars are picked up, and the target signal is constructed from the target exemplars and the picked-up weights. This method showed better performances than the conventional Gaussian Mixture Model (GMM)-based method [6] in VC experiments using noise-added speech data. However, the performance of our method was not good enough for practical use.

As one of the techniques used for robust speech recognition under noisy environments, audio-visual speech recognition, which uses lip dynamic visual information and audio information has been studied. In audio-visual speech recognition, there are mainly three integration methods: early integration [7], which connects the audio feature vector with the visual feature vector; late integration [8], which weights the likelihood of the result obtained by a separate process for audio and visual signals; and synthetic integration [9], which calculates the product of output probability in each state and so on.

In this paper, we propose a multimodal VC technique using NMF that uses visual information together with audio information as an input feature. The visual information is extracted from videos, which captured lip movement of the utterances. The extracted visual features are connected with the audio features and used as source exemplars. The input noisy audio-visual feature is represented by a linear combination of source exemplars and noise exemplars. Then, the source exemplars are replaced with related parallel target exemplars, which are extracted from clean audio features. The effectiveness of this method was confirmed by comparing it with that of the conventional audio input NMF-based method and the conventional GMM-based method.

The rest of this paper is organized as follows: In Section 2, related works are introduced. In Section 3, our proposed method is described. In Section 4, the experimental data are evaluated, and the final section is devoted to our conclusions.

2. RELATED WORKS

VC is a technique for converting specific information in speech while maintaining the other information in the utterance. One of the most popular VC applications is speaker conversion [6]. In speaker con-

version, a source speaker’s voice individuality is changed to a specified target speaker’s so that the input utterance sounds as though a specified target speaker had spoken it.

There have also been studies on several tasks that make use of VC. Emotion conversion is a technique for changing emotional information in input speech while maintaining linguistic information and speaker individuality [10, 11]. VC is also being adopted as assistive technology that reconstructs a speaker’s individuality in electrolaryngeal speech [12], disordered speech [13] or speech recorded by NAM microphones [14]. In recent years, VC has been used for automatic speech recognition (ASR) or speaker adaptation in text-to-speech (TTS) systems [15].

The statistical approaches to VC are most widely studied [6, 16, 17]. Among these approaches, the Gaussian mixture model (GMM)-based mapping approach [6] is widely used. In this approach, the conversion function is interpreted as the expectation value of the target spectral envelope. The conversion parameters are evaluated using Minimum Mean-Square Error (MMSE) on a parallel training set. A number of improvements in this approach have been proposed. Toda et al. [18] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander et al. [19] proposed transforms based on partial least squares (PLS) in order to prevent the over-fitting problem associated with standard multivariate regression. There have also been approaches that do not require parallel data that make use of GMM adaptation techniques [20] or eigen-voice GMM (EV-GMM) [21, 22].

However, the effectiveness of these approaches was confirmed with clean speech data, and their utilization in noisy environments was not considered. The noise in the input signal may degrade the conversion performance itself due to unexpected mapping of source features. To address the problem, in this paper, we propose exemplar-based multimodal VC. Joint audio-visual features are used as the source feature of NMF-based VC [5]. Because the audio features are not affected by background noise, our method improved the noise robustness of NMF-based VC.

3. MULTIMODAL VOICE CONVERSION

3.1. Basic Approach

In the approaches based on sparse representations, the observed signal is represented by a linear combination of a small number of bases.

$$\mathbf{x}_l \approx \sum_{j=1}^J \mathbf{a}_j h_{j,l} = \mathbf{A} \mathbf{h}_l \quad (1)$$

\mathbf{x}_l represents the l -th frame of the observation. \mathbf{a}_j and $h_{j,l}$ represent the j -th basis and the weight, respectively. $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_J]$ and $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^T$ represent the collection of the bases and the stack of weights. When the weight vector \mathbf{h}_l is sparse, the observed signal can be represented by a linear combination of a small number of bases that have non-zero weights. In this paper, each basis denotes the exemplar of the speech or noise signal, and the collection of exemplar \mathbf{A} and the weight vector \mathbf{h}_l are called the ‘dictionary’ and ‘activity’, respectively.

Fig. 1 shows the basic approach of our exemplar-based VC using NMF. D , d , L , and J represent the number of dimensions of source features, dimensions of target features, frames of the dictionary, and basis of the dictionary, respectively. Our VC method needs two dictionaries that are phonemically parallel, where one dictionary (source dictionary) is constructed from source features and the

other dictionary (target dictionary) is constructed from target features. These two dictionaries consist of the same words and are aligned with dynamic time warping (DTW). Hence, these dictionaries have the same number of bases.

Input source features X^s are decomposed into a linear combination of bases from the source dictionary A^s by NMF. The weights of the bases are estimated as an activity H^s . Therefore, the activity includes the weight information of input features for each basis. Then, the activity is multiplied by a target dictionary in order to obtain converted spectral features \hat{X}^t , which are represented by a linear combination of bases from the target dictionary. Because the source and target dictionary are parallel phonemically, the bases used in the converted features are phonemically the same as those of the source features.

Fig. 2 shows an example of the activity matrices estimated from a word Japanese word “ikioi” (“vigor” in English). To show an intelligible example, each dictionary was structured from just the one word “ikioi” and aligned with DTW. The source/target features and each atom in the dictionary are a spectral envelope extracted by STRAIGHT analysis [23]. When the source/target signals and its dictionary are the same word, the estimated activity will have high energies through the diagonal line. The reason some areas far from the diagonal line, such as the red-circled areas, also have high energies are that those areas correspond to the same utterance ‘i’.

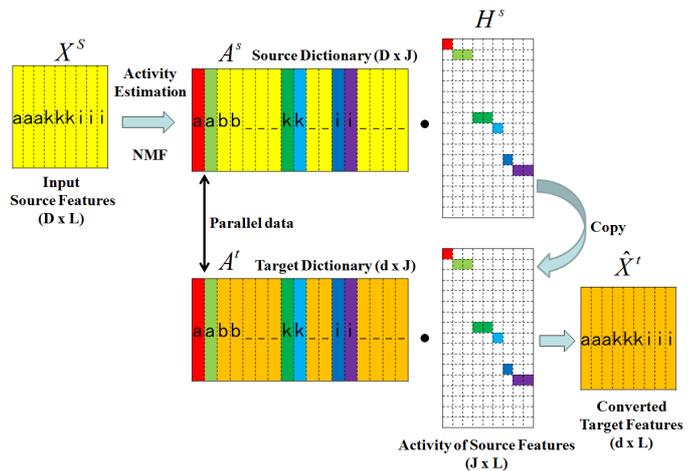


Fig. 1. Basic approach of NMF-based voice conversion

3.2. Multimodal Dictionary Construction

Fig. 3 shows the process for constructing a parallel dictionary. In order to make a parallel dictionary, some pairs of parallel utterances are needed, where each pair consists of the same text. The source dictionary A^s consists of jointed audio-visual features, while the target dictionary A^t consists of audio features only.

For the audio features, STRAIGHT spectrum is extracted from clean parallel utterances. Mel-cepstral coefficients are calculated from the STRAIGHT spectrum in order to get alignment information in DTW. The extracted spectrum envelopes are phonemically aligned with DTW. For visual features, the image spectrum of lip motion images of the source speaker’s utterance is used. In order to confirm the non-negativity constraint of NMF, the image spectrum is calculated by short-time Fourier transform (STFT). The image spectrum is interpolated by spline interpolation in order to fill the sam-

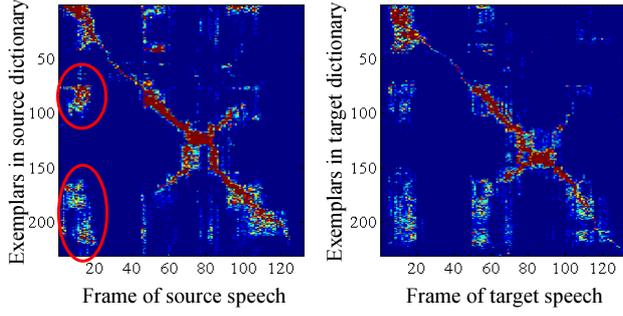


Fig. 2. Activity matrices of the source signal (left) and target signal (right)

pling rate gap between audio and visual features. Aligned audio and visual features of the source speaker are joined and used as a source feature. Source and target dictionaries are constructed by lining up each of the features extracted from parallel utterances.

The audio feature of the noise dictionary is extracted from the before- and after-utterance sections in the input noisy audio signal. The mean vector, which is calculated from all the input visual images is used as the visual feature of the noise dictionary.

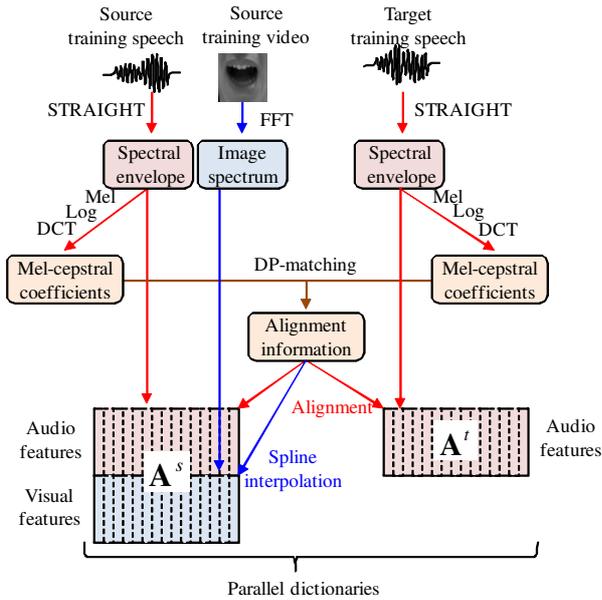


Fig. 3. Multimodal dictionary construction

3.3. Estimation of Activity from Noisy Source Signals

In the exemplar-based approach, the spectrum of the noisy source signal at frame l is approximately expressed by a non-negative linear combination of the source dictionary, noise dictionary, and their

activities.

$$\begin{aligned}
 \mathbf{x}_l &= \mathbf{x}_l^s + \mathbf{x}_l^n \\
 &\approx \sum_{j=1}^J \mathbf{a}_j^s h_{j,l}^s + \sum_{k=1}^K \mathbf{a}_k^n h_{k,l}^n \\
 &= [\mathbf{A}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{h}_l^s \\ \mathbf{h}_l^n \end{bmatrix} \quad s.t. \quad \mathbf{h}_l^s, \mathbf{h}_l^n \geq 0 \\
 &= \mathbf{A} \mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0
 \end{aligned} \tag{2}$$

\mathbf{x}_l^s and \mathbf{x}_l^n represent the spectrum of the source signal and the noise, respectively. \mathbf{A}^s , \mathbf{A}^n , h_l^s , h_l^n represent the source dictionary, noise dictionary, and their activities at frame l , respectively. Given the spectrogram, (2) can be written as follows:

$$\begin{aligned}
 \mathbf{X} &\approx [\mathbf{A}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{H}^s \\ \mathbf{H}^n \end{bmatrix} \quad s.t. \quad \mathbf{H}^s, \mathbf{H}^n \geq 0 \\
 &= \mathbf{A} \mathbf{H} \quad s.t. \quad \mathbf{H} \geq 0.
 \end{aligned} \tag{3}$$

In order to consider only the shape of the spectrum, \mathbf{X} , \mathbf{A}^s and \mathbf{A}^n are first normalized for each frame or exemplar so that the sum of the magnitudes over frequency bins equals unity.

$$\begin{aligned}
 \mathbf{M} &= \mathbf{1}^{(D \times D)} \mathbf{X} \\
 \mathbf{X} &\leftarrow \mathbf{X} ./ \mathbf{M} \\
 \mathbf{A} &\leftarrow \mathbf{A} ./ (\mathbf{1}^{(D \times D)} \mathbf{A})
 \end{aligned} \tag{4}$$

$\mathbf{1}$ is an all-one matrix and $./$ denotes element-wise division, respectively. The joint matrix \mathbf{H} is estimated based on NMF with the sparse constraint that minimizes the following cost function [4]:

$$d(\mathbf{X}, \mathbf{A} \mathbf{H}) + \|(\lambda \mathbf{1}^{(1 \times L)}) .* \mathbf{H}\|_1 \quad s.t. \quad \mathbf{H} \geq 0. \tag{5}$$

The first term is the Kullback-Leibler (KL) divergence between \mathbf{X} and $\mathbf{A} \mathbf{H}$. The second term is the sparse constraint with the L1-norm regularization term that causes \mathbf{H} to be sparse. $.*$ denotes element-wise multiplication. The weights of the sparsity constraints can be defined for each exemplar by defining $\lambda^T = [\lambda_1 \dots \lambda_J \dots \lambda_{J+K}]$. In this paper, the weights for source exemplars $[\lambda_1 \dots \lambda_J]$ were set to 0.1, and those for noise exemplars $[\lambda_{J+1} \dots \lambda_{J+K}]$ were set to 0. \mathbf{H} minimizing (5) is estimated iteratively applying the following update rule:

$$\begin{aligned}
 \mathbf{H}_{n+1} &= \mathbf{H}_n .* (\mathbf{A}^T (\mathbf{X} ./ (\mathbf{A} \mathbf{H}))) \\
 &\quad ./ (\mathbf{1}^{((J+K) \times L)} + \lambda \mathbf{1}^{(1 \times L)}).
 \end{aligned} \tag{6}$$

3.4. Target Speech Construction

From the estimated joint matrix \mathbf{H} , the activity of the source signal \mathbf{H}^s is extracted, and by using the activity and the target dictionary, the converted spectral features are constructed. Then, the target dictionary is also normalized for each frame in the same way the source dictionary was.

$$\mathbf{A}^t \leftarrow \mathbf{A}^t ./ (\mathbf{1}^{(d \times d)} \mathbf{A}^t) \tag{7}$$

Next, the normalized target spectral feature is constructed, and the magnitudes of the source signal calculated in (4) are applied to the normalized target spectral feature.

$$\hat{\mathbf{X}}^t = (\mathbf{A}^t \mathbf{H}^s) .* \mathbf{M} \tag{8}$$

The input source and converted spectral feature is expressed as a STRAIGHT spectrum. Hence, the target speech is synthesized using a STRAIGHT synthesizer. The other features extracted by STRAIGHT analysis, such as F0 and the aperiodic components, are used to synthesize the converted signal without any conversion.

4. EXPERIMENTAL RESULTS

4.1. Experimental Conditions

The proposed multimodal VC technique was evaluated by comparing it with an exemplar-based audio-input method [5] and a conventional GMM-based method [6] in a speaker-conversion task using clean speech data and noise-added speech data. The source speaker and target speaker were one male and one female speaker, whose speech is stored in the ATR Japanese speech database [24], respectively. Source speaker speech and visual data are taken from the M2TINIT database [25]. The sampling rate of the audio data was 16 kHz. The frame rate of the visual data was 1/29.97 sec and the pixel number is 720 x 840.

A total of 40 sentences of clean speech were used to construct parallel dictionaries in the methods based on sparse representation and used to train the GMM in the GMM-based method. In the exemplar-based method, the number of exemplars of the source and target dictionaries was 68,580. Ten sentences of clean speech or noisy speech were used in the evaluation. The noisy speech was created by adding a noise signal recorded in a restaurant (taken from the CENSREC-1-C database [26]) to the clean speech sentences. The SNR was 24 dB. The noise dictionary is extracted from the before- and after-utterance sections in the evaluation sentence.

In the methods based on sparse representation, a 1,025-dimensional STRAIGHT spectrum was used for the source and target dictionary. The number of iterations used to estimate the activity was 300. In the GMM-based method, the 1st through 24th linear-cepstral coefficients obtained from the STRAIGHT spectrum were used as the feature vectors. The number of mixtures was 64.

4.2. Results and Discussion

Fig. 4 shows the spectral distortion improvement ratio (SDIR) [dB] for the noisy input source signal. The SDIR is defined as follows:

$$\text{SDIR}[\text{dB}] = 10 \log_{10} \frac{\sum_d |\mathbf{X}^t(d) - \hat{\mathbf{X}}^t(d)|^2}{\sum_d |\mathbf{X}^t(d) - \mathbf{X}^s(d)|^2} \quad (9)$$

Here, \mathbf{X}^s , \mathbf{X}^t and $\hat{\mathbf{X}}^t$ are normalized so that the sum of the magnitudes over frequency bins equals unity. As shown in Fig. 4, the distortion improvements of the proposed method were higher than other two methods.

We performed a mean opinion score (MOS) test [27] on the naturalness and noise suppression of the converted speech. The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). The tests were carried out with 6 subjects. For the evaluation of naturalness, each subject listened to the converted speech and evaluated how natural the sample sounded. For the evaluation of noise suppression, each subject listened to the converted speech and evaluated how the noise of the sample is suppressed.

Fig. 5 shows the results of the MOS test. The error bars show 95% confidence intervals. As shown in this figure, the performances of the GMM-based method degraded considerably, particularly in naturalness. This might be because the noise caused unexpected mapping in the GMM-based method, and the speech was converted with a lack of naturalness. On the other hand, the degradations of the performances of the VC methods based on our proposed audio-visual NMF and audio NMF were less than those of GMM-based method. Moreover, in the noise suppression test, our proposed method obtained a higher score than the other two methods. This result showed the noise robustness of our multimodal VC method.

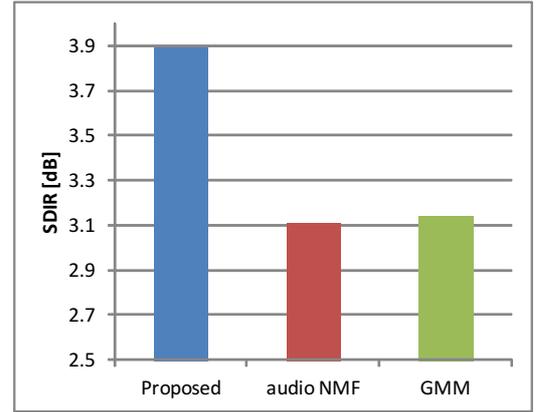


Fig. 4. Spectral distortion improvement ratio calculated from converted speech using each method

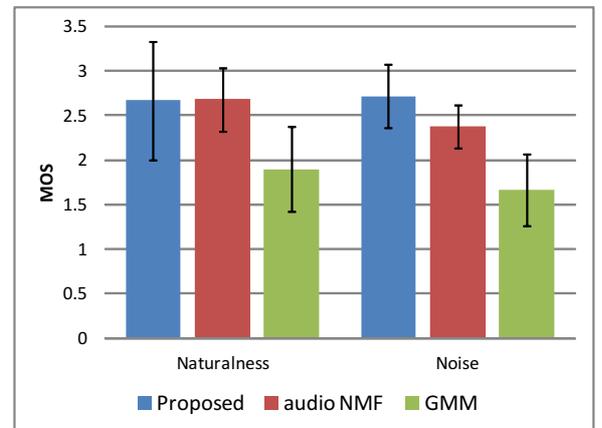


Fig. 5. Mean opinion score for subjective evaluations

5. CONCLUSIONS

We proposed multimodal VC using NMF based on the idea of sparse representation. In our proposed method, the joint audio-visual feature is used as the source feature. Input noisy audio-visual features are decomposed into a linear combination of the clean audio-visual feature and the noise feature. By replacing the source speaker's audio-visual feature with the target speaker's audio feature, the voice individuality of the source speaker is converted to the target speaker. Objective and subjective evaluations show the greater effectiveness of our VC technique compared to conventional audio-input NMF and GMM-based VC.

However, this method requires that the activity of each atom in the dictionary be estimated, and it requires high computation times. In [28], we proposed a framework to train basis matrices of source and target exemplars in order to reduce computational cost. In future work, we will combine that method and our proposed method in this paper. Then we will investigate the optimal number of bases and evaluate the performances. In addition, this method has a limitation in that it can be applied to only one-to-one voice conversations. Hence, we will investigate a method that does not use parallel data. Our future work will also include efforts to investigate other noise conditions, such as a low-SNR, and apply this method to other VC applications.

6. REFERENCES

- [1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Neural Information Processing System*, pp. 556–562, 2001.
- [2] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.
- [3] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Interspeech*, 2006.
- [4] J. F. Gemmeke, T. Viratnen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.
- [5] R. Takashima, T. Takiguchi, and Y. Arika, "Exemplar-based voice conversion in noisy environment," in *SLT*, pp. 313–317, 2012.
- [6] Y. Stylianou, O. Cappe, and E. Moilines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.
- [7] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *ICASSP*, pp. 3733–3736, 1998.
- [8] A. Verma, T. Faruque, C. Neti, S. Basu, and A. Senior, "Late integration in audio-visual continuous speech recognition," in *ASRU*, 1999.
- [9] M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *ICASSP*, pp. 821–824, 1996.
- [10] C. Veaux and X. Robet, "Intonation conversion from neutral to expressive speech," in *Interspeech*, pp. 2765–2768, 2011.
- [11] R. Aihara, R. Takashima, T. Takiguchi, and Y. Arika, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, no. 5, 2012.
- [12] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.
- [13] R. Aihara, R. Takashima, T. Takiguchi, and Y. Arika, "Individuality-preserving voice conversion for articulation disorders based on Non-negative Matrix Factorization," in *ICASSP*, pp. 8037–8040, 2013.
- [14] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech," in *Interspeech*, pp. 148–151, 2006.
- [15] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP*, vol. 1, pp. 285–288, 1998.
- [16] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models," in *ICASSP*, pp. 655–658, 1988.
- [17] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, no. 2-3, pp. 175–187, 1992.
- [18] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.
- [19] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, Issue:5, pp. 912–921, 2010.
- [20] C. H. Lee and C. H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *Interspeech*, pp. 2254–2257, 2006.
- [21] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *Interspeech*, pp. 2446–2449, 2006.
- [22] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Interspeech*, pp. 653–656, 2011.
- [23] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.
- [24] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, pp. 357–363, 1990.
- [25] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "HMM-based text-to-audio-visual speech synthesis - image-based approach," *ICSLP*, vol. III, pp. 25–28, 2000.
- [26] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura, "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments," *Acoustical Science and Technology*, Vol. 30 (2009), No. 5, pp. 363–371, 2009.
- [27] INTERNATIONAL TELECOMMUNICATION UNION, "Methods for objective and subjective assessment of quality," *ITU-T Recommendation P.800*, 2003.
- [28] R. Takashima, R. Aihara, T. Takiguchi, and Y. Arika, "Noise-robust voice conversion based on spectral mapping on sparse space," *SSW8, 8th ISCA Speech Synthesis Workshop*, pp. 71–75, 2013.