Multimodal Exemplar-based Voice Conversion using Lip Features 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
{ makka, aihara } @me.cs.scitec.kobe-u.ac.jp, { takigu, ariki } @kobe-u.ac.jp

Abstract

This paper presents a multimodal voice conversion (VC) method for noisy environments. In our previous exemplar-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 method that improves the noise robustness of our previous exemplar-based VC method. As visual features, we use not only conventional DCT but also the features extracted from Active Appearance Model (AAM) applied to the lip area of a face image. Furthermore, we introduce the combination weight between audio and visual features and formulate a new cost function in order to estimate the audio-visual exemplars. By using the joint audio-visual features as source features, the VC performance is improved compared to a previous audio-input exemplar-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. DCT is widely used as a visual feature in audio-speech recognition. In [10], we proposed audio-visual speech recognition using a visual feature extracted from AAM [11]. The feature contains shape information, which expresses the lip movement and texture information which express intensity changes such as teeth.

In this paper, we propose a multimodal VC technique using NMF with a combination weight between audio and visual features. The visual information is extracted from videos, which captured the lip movement of the utterances. As visual features, we use not only DCT but also a visual feature extracted from AAM. 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 conversion, 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 [12, 13]. VC is also being adopted as assistive technology that reconstructs a speaker's individuality in electrolaryngeal speech [14], disordered speech [15] or speech recorded by NAM microphones [16]. In recent years, VC has been used for automatic speech recognition (ASR) or speaker adaptation in text-to-speech (TTS) systems [17].

The statistical approaches to VC are most widely studied [6, 18, 19]. 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. [20] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander et al. [21] 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 [22] or eigen-voice GMM (EV-GMM) [23, 24].

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.

$$\mathbf{x}_l \approx \sum_{j=1}^{J} \mathbf{w}_j h_{j,l} = \mathbf{W}\mathbf{h}_l \tag{1}$$

$\mathbf{x}_l$ represents the l-th frame of the observation. $\mathbf{w}_j$ and $h_{j,l}$ represent the j-th basis and the weight, respectively. $\mathbf{W} = [\mathbf{w}_1 \ldots \mathbf{w}_J]$ and $\mathbf{h}_l = [h_{1,l} \ldots h_{J,l}]^T$ 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 $\mathbf{W}$ and the weight vector $\mathbf{h}_l$ are called the 'dictionary' and 'activity', respectively.

Figure 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.

An input source feature matrix $\mathbf{X}^s$ is decomposed into a linear combination of bases from the source dictionary $\mathbf{W}^s$ using NMF. The weights of the bases are estimated as an activity $\mathbf{H}^s$. 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 the converted spectral feature matrix $\hat{\mathbf{X}}^t$, which is represented by a linear combination of bases from the target dictionary. Because the source and target dictionaries are parallel phonemically, the bases used in the converted features are phonemically the same as those of the source features.

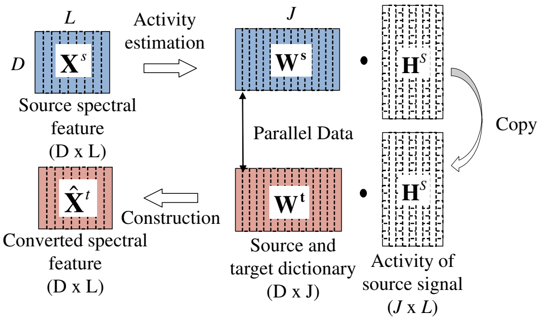
L
Activity estimation
J
D
X^s
W^s
•
H^S
Copy
Source spectral feature (D x L)
Parallel Data
X̂^t
Construction
W^t
•
H^S
Converted spectral feature (D x L)
Source and target dictionary (D x J)
Activity of source signal (J x L)

Figure 1: Basic approach of NMF-based voice conversion

3.2. Multimodal Dictionary Construction

Figure 2 shows the process for constructing a parallel dictionary. In order to construct a parallel dictionary, some pairs of parallel utterances are needed, where each pair consists of the same text. The source dictionary W^s consists of jointed audio-visual features, while the target dictionary W^t consists of audio features only.

For the audio features, a simple magnitude spectrum calculated by short-time Fourier transform (STFT) is extracted from clean parallel utterances. Mel-cepstral coefficients are calculated from the STRAIGHT spectrum in order to obtain alignment information in DTW.

For visual features, we use DCT and the feature extracted from AAM [10] of lip motion images of the source speaker's utterance is used. AAM is a technique to express the face model by the low-dimensional parameter. The subspace is constructed by applying PCA to shape and texture of face feature points, where $\mathbf{s}$ and $\mathbf{g}$ which represent the shape vector and the texture vector are expressed by each mean vector $\bar{\mathbf{s}}$ and $\bar{\mathbf{g}}$ and eigenvector matrices $\mathbf{P_s}$ and $\mathbf{P_g}$.

$$
\begin{aligned}
\mathbf{s} &= \bar{\mathbf{s}} + \mathbf{P_s b_s} && (2)\\
\mathbf{g} &= \bar{\mathbf{g}} + \mathbf{P_g b_g} && (3)
\end{aligned}
$$

$\mathbf{b_s}$ and $\mathbf{b_g}$ represent the shape parameter and the texture parameter, respectively. They are combined and reduced as shown in Eq. (4) by applying PCA because there is a correlation in shape and texture.

$$
\mathbf{b} = \begin{pmatrix} \mathbf{R_s b_s} \\ \mathbf{b_g} \end{pmatrix} = \begin{pmatrix} \mathbf{R_s P_s}^T(\mathbf{s} - \bar{\mathbf{s}}) \\ \mathbf{P_g}^T(\mathbf{g} - \bar{\mathbf{g}}) \end{pmatrix} = \mathbf{Qc} \tag{4}
$$

$\mathbf{R_s}$ is the matrix that normalizes the difference in the unit of the shape vector and the texture vector. $\mathbf{Q}$ is an eigenvector matrix. $\mathbf{c}$ is a vector of combined shape and texture parameters. By controlling parameter $\mathbf{c}$ it becomes possible to operate shape and texture together.

Then they are adjusted to satisfy the non-negativity constraint of NMF. The visual features are interpolated by spline interpolation in order to fill the sampling rate gap between audio 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 visual feature of the noise dictionary is extracted in the same way of audio feature.

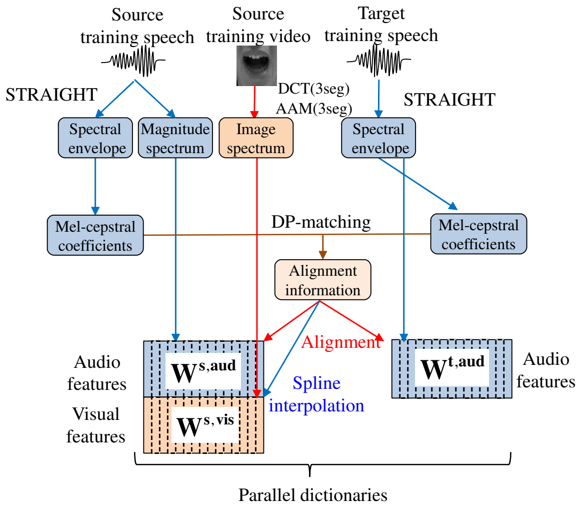

Figure 2: Multimodal dictionary construction

3.3. Estimation of Activity from Noisy Source Signals using NMF with a Combination Weight

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

$$
\begin{aligned}
\mathbf{x} &= \mathbf{x}^s + \mathbf{x}^n \\
&\approx \sum_{j=1}^{J} \mathbf{w}_j^s h_j^{av} + \sum_{k=1}^{K} \mathbf{w}_k^n h_k^n \\
&= [\mathbf{W}^s\ \mathbf{N}] \begin{bmatrix} \mathbf{h}^{av} \\ \mathbf{h}^n \end{bmatrix} \quad s.t. \quad \mathbf{h}^{av}, \mathbf{h}^n \geq 0 \\
&= \mathbf{Wh} \quad s.t. \quad \mathbf{h} \geq 0
\end{aligned} \tag{5}
$$

$\mathbf{x}^s$ and $\mathbf{x}^n$ represent the spectrum of the source signal and the noise, respectively. $\mathbf{W}^s$, $\mathbf{N}$, $\mathbf{h}^{av}$, $\mathbf{h}^n$ represent the source dictionary, noise dictionary, and their activities at a frame, respectively. All spectra are normalized for each frame.

The joint matrix $\mathbf{h}$ is estimated based on NMF with the sparse constraint that minimizes a cost function [4]. In [25], we used a simple NMF without considering the weights of the audio and visual parameters when estimating the activity. In this paper, we introduce audio-visual weight α and β because we have to adjust the weight depending on SNR, and we propose a new cost function as follows:

$$
\begin{aligned}
&\alpha d(\mathbf{x}^{s,aud}, [\mathbf{W}^{s,aud}\ \mathbf{N}^{aud}]\mathbf{h}) + \beta d(\mathbf{x}^{s,vis}, [\mathbf{W}^{s,vis}\ \mathbf{N}^{vis}]\mathbf{h}) \\
&\quad + ||\lambda .* \mathbf{h}||_1 \quad s.t. \quad \mathbf{h} \geq 0
\end{aligned} \tag{6}
$$

The first and second terms are the Kullback-Leibler (KL) divergence between $\mathbf{x}^{s,aud}$ and $[\mathbf{W}^{s,aud}\ \mathbf{N}^{aud}]\mathbf{h}$, $\mathbf{x}^{s,vis}$ and $[\mathbf{W}^{s,vis}\ \mathbf{N}^{vis}]\mathbf{h}$, respectively. The third term is the sparse constraint with the L1-norm regularization term that causes $\mathbf{h}$ to be sparse. The symbol of $.*$ denotes element-wise multiplication. The weights of the sparsity constraints can be defined for each exemplar by defining $\lambda^T = [\lambda_1 \ldots \lambda_J \ldots \lambda_{J+K}]$. In this paper, the weights for source exemplars $[\lambda_1 \ldots \lambda_J]$ were set to 0.1, and those for noise exemplars $[\lambda_{J+1} \ldots \lambda_{J+K}]$ were set to 0. $\mathbf{h}$ minimizing (6) is estimated iteratively applying the following update rule:

$$
h_j \leftarrow h_j \frac{\sum_{d}^{D} \mathbf{f}_d + \sum_{e}^{E} \mathbf{g}_e}{\alpha + \beta + \lambda_j} \tag{7}
$$

$$
\mathbf{f}_d = \alpha W_{d,j}^{s,aud} \alpha x_d^a / (\alpha \mathbf{W}^{s,aud} \mathbf{h}^{av})_d \tag{8}
$$

$$
\mathbf{g}_e = \beta W_{e,j}^{s,vis} \beta x_e^v / (\beta \mathbf{W}^{s,vis} \mathbf{h}^{av})_e \tag{9}
$$

D and E represent the dimension of the audio and visual dictionaries, respectively.

3.4. Target Speech Construction

From the estimated joint matrix $\mathbf{h}$, the activity of the source signal $\mathbf{h}^{av}$ is extracted, and by using the activity and the target dictionary, the converted spectral features are constructed.

$$
\hat{\mathbf{x}}^t = \mathbf{W}^{t,aud} \mathbf{h}^{av} \tag{10}
$$

The input source and converted spectral features are 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.

Table 1: SDIR calculated from converted speech using each image feature and weight

0 dB	GMM	(audio NMF) $\beta = 0$	$\beta = 0.1$	$\beta = 1$	$\beta = 10$	$\beta = 20$	$\beta = 30$
DCT	1.3521	2.4857	2.5031	2.631	**2.7847**	2.6809	2.5932
AAM			2.511	2.6719	2.7259	**2.7313**	2.7169
DCT+AAM			2.7672	**2.7673**	2.705	2.6185	2.5427

10 dB	GMM	(audio NMF) $\beta = 0$	$\beta = 0.1$	$\beta = 1$	$\beta = 10$	$\beta = 20$	$\beta = 30$
DCT	1.4197	2.9052	2.9128	2.9713	**3.0777**	2.9871	2.8829
AAM			2.9203	3.0243	**3.0603**	3.0363	2.9853
DCT+AAM			2.9302	**3.0783**	3.0218	2.8735	2.7397

20 dB	GMM	(audio NMF) $\beta = 0$	$\beta = 0.1$	$\beta = 1$	$\beta = 10$	$\beta = 20$	$\beta = 30$
DCT	1.5251	3.2565	3.2605	3.3118	**3.3955**	3.3303	3.2325
AAM			3.2641	3.3562	**3.4501**	3.417	3.3511
DCT+AAM			3.2712	**3.3867**	3.3947	3.2323	3.0742

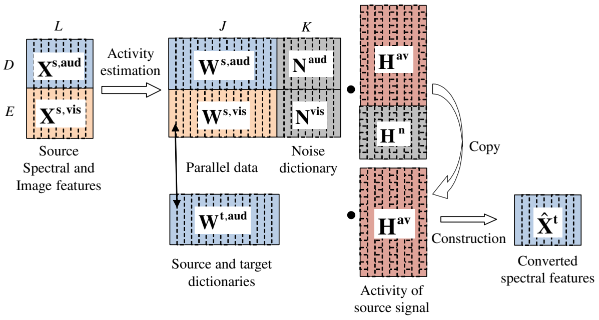

Figure 3: Flow chart of multimodal conversion

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, respectively, whose speech is stored in the ATR Japanese speech database [26]. Source speaker speech and visual data are taken from the M2TINIT database [27]. The sampling rate of the audio data was 16 kHz. The frame rate of the visual data was 1/29.97 sec and the image size is 720 x 840.

A total of 50 sentences of clean speech were used to construct the parallel dictionary for each method 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 80,868. 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 [28]) to the clean speech sentences. The SNR was 0 dB, 10 dB and 20 dB. The noise dictionary is extracted from the before- and after-utterance sections in the evaluation sentence. For the visual dictionary, we used a 40-dimension DCT feature and a 23-dimension c parameter extracted from AAM and its segment feature, which is constructed by concatenating spectra at each current frame ± three frames. For the weights, α is 1, β is changed 0.1, 1, 10, 20, 30.

In the methods based on sparse representation, a 513-dimensional magnitude spectrum was used for the source and noise dictionaries and a 1025-dimensional STRAIGHT spectrum was used for the target dictionary.

The number of iterations used to estimate the activity was 300. In the GMM-based method, the 1^{st} through 24^{th} 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

Table 1 shows the spectral distortion improvement ratio (SDIR) [dB] for the noisy input source signal. The SDIR is defined as follows:

$$\mathrm{SDIR[dB]} = 10 \log_{10} \frac{\sum_d |\mathbf{X}^t(d) - \mathbf{X}^s(d)|^2}{\sum_d |\mathbf{X}^t(d) - \hat{\mathbf{X}}^t(d)|^2} \tag{11}$$

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 Table 1, the distortion improvements of the proposed method were higher than conventional NMF and GMM-based method, where the performance of AAM is slightly better than that of DCT. Also, as the SNR decreases, an optimum β may tend to be large. By changing the weight, we can obtain the best SDIR value in various SNR environments.

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. Furthermore we introduced audio-visual weight and formulate a new cost function. By selecting optimal weight and image features, we achieve the best performance of transformation accuracy. Evaluations show the greater effectiveness of our VC technique compared to conventional audio-input NMF and GMM-based VC.

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. Ariki, “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. Faruquie, 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] Y. Komai, N. Yang, T. Takiguchi, and Y. Ariki, “Robust aam-based audio-visual speech recognition against face direction changes,” *ACM Multimedia*, pp. 1161–1164, 2012.

[11] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 681–685, 2001.

[12] C. Veaux and X. Robet, “Intonation conversion from neutral to expressive speech,” *in Interspeech*, pp. 2765–2768, 2011.

[13] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, “GMM-based emotional voice conversion using spectrum and prosody features,” *American Journal of Signal Processing*, vol. 2, no. 5, 2012.

[14] 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.

[15] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, “Individuality-preserving voice conversion for articulation disorders based on Non-negative Matrix Factorization,” *in ICASSP*, pp. 8037–8040, 2013.

[16] 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.

[17] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” *in ICASSP, vol. 1, pp. 285-288*, 1998.

[18] 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.

[19] H. Valbret, E. Moulines, and J. P. Tubach, “Voice transformation using PSOLA technique,” *Speech Communication, vol. 11, no. 2-3, pp. 175-187*, 1992.

[20] 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.

[21] 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.

[22] 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.

[23] T. Toda, Y. Ohtani, and K. Shikano, “Eigenvoice conversion based on Gaussian mixture model,” *in Interspeech*, pp. 2446–2449, 2006.

[24] 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.

[25] K. Masaka, R. Aihara, T. Takiguchi, and Y. Ariki, “Mutimodal voice conversion using non-negative matrix factorization in noisy environments,” *in ICASSP 2014*, 2014.

[26] 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.

[27] 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.

[28] 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.