IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 25, 2019, accepted November 13, 2019, date of publication November 26, 2019,
date of current version December 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955982

Multimodal Voice Conversion Under Adverse Environment Using a Deep Convolutional Neural Network

JIAN ZHOU[1], YUTING HU[1], HAILUN LIAN[1], HUABIN WANG[1], LIANG TAO[1], AND HON KEUNG KWAN[2], (Life Senior Member, IEEE)

[1]Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei 230601, China
[2]Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada

Corresponding author: Jian Zhou (jzhou@ahu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61301295, in part by the Anhui Provincial Natural Science Foundation under Grant 1708085MF151, and in part by the Natural Science Foundation for the Higher Education Institutions of Anhui Province under Grant KJ2018A0018.

ABSTRACT This paper presents a voice conversion (VC) technique under noisy environments. Typically, VC methods use only audio information for conversion in a noiseless environment. However, existing conversion methods do not always achieve satisfactory results in an adverse acoustic environment. To solve this problem, we propose a multimodal voice conversion model based on a deep convolutional neural network (MDCNN) built by combining two convolutional neural networks (CNN) and a deep neural network (DNN) for VC under noisy environments. In the MDCNN, both the acoustic and visual information are incorporated into the voice conversion to improve its robustness in adverse acoustic conditions. The two CNNs are designed to extract acoustic and visual features, and the DNN is designed to capture the nonlinear mapping relation of source speech and target speech. Experimental results indicate that the proposed MDCNN outperforms two existing approaches in noisy environments.

INDEX TERMS Audio and video feature fusion, convolutional neural network, deep learning, mel-frequency cepstral coefficients, multilayer feedforward neural networks, multimodal voice conversion, noise robustness.

I. INTRODUCTION

Voice conversion (VC) is an emergent problem in speech processing that deals with modifying a speaker's identity. The goal of a voice conversion system is to change a source speaker's speech so that it sounds as if it were spoken by a different speaker (target speaker) without changing the linguistic content [1]. In the past two decades, much attention has been given to VC owing to its wide range of applications, including customization of talking devices, designing damaged voice restoring tools to assist people with voice disorders, disguising speaker identity in communication, dubbing films, translation into different languages, and synthesis of text-to-speech (TTS) where a voice conversion system is used to create natural and intelligible voice.

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

The basic idea behind VC is to identify relevant acoustic features of a source speaker and replace them with those of a target speaker without modifying the message. Typically, a VC system consists of three main modules [2]: extracting representative acoustic features, constructing mapping rules between a source speaker and a target speaker, and synthesizing a target speech.

In the past few decades, a number of statistical methods have been investigated for VC by regarding it as a task of mapping from a source space to a target space [3], [4]. In recent years, Gaussian mixture models (GMMs) [5] and neural networks [6] have been commonly used for spectral mapping. In the GMM-based approach [7], the joint distribution of features extracted from the speech signal of a source speaker and a target speaker is modeled by the sum of weighted Gaussian components. The acoustic feature space of speakers is partitioned into overlapping classes, and the weighted contribution of all classes is considered.

This enables spectral envelopes to be converted successfully without discontinuities. The performance of a GMM-based voice conversion improves as the number of mixture components increases [8].

Although a number of improvements of GMM have been proposed, the usefulness of GMMs still faces some problems due to its limited capability to capture the source-target correspondence for a given parametric representation of speech. For example, if a GMM is based on a single conversion frame, the sequential information between frames is ignored [9]. Additionally, a GMM may fail when the size and/or the dimension of the feature space of a speech is too large [10]. To address this issue, Chen *et al.* [11] proposed using deep neural networks (DNN) to construct a global nonlinear mapping relationship between the spectral envelopes of two speakers. The proposed DNN was trained by cascading two restricted Boltzmann machines (RBMs) to model the distributions of spectral envelopes of a source speaker and a target speaker. Furthermore, a different DNN framework was proposed by Ye and Yu [12] to map the spectral envelopes between a target speaker and a source speaker, in which the neighbor frame's influence is carefully considered. A VC method using deep bidirectional long short-term memory-based recurrent neural networks (DBLSTM) was proposed by Li et al. in [13]. Additionally, Nguyen et al. proposed a comprehensive VC framework using deep neural networks to model high-dimensional features, including both high-resolution timbre spectral features and prosodic features, such as fundamental frequency (F0), intensity, and duration in [14].

Although the effectiveness of these approaches [11]–[14] for clean speech data was confirmed, their utilization in noisy environments has not been considered. The noise in a source speech may degrade conversion performance due to unexpected feature mapping between a source speech and a target speech [15]. Hence, VC technique in noisy environments is a subject of interest.

Human speech is an articulatory-to-auditory mapping process in which mouth, vocal tract, and lip movements produce an audible acoustic signal. In addition to speech signals, visual information is important in human-human or human-machine interactions. A study by McGurk [16] indicated that the shape of lips or mouth could play an important role in speech processing. Accordingly, audio-visual multimodality has been adopted in a number of areas in speech processing [17]–[21].

In [22], Masaka et al. proposed a multimodal VC using nonnegative matrix factorization (NMF) based on the idea of sparse representation. Input noisy audio-visual features were decomposed into a linear combination of clean audio-visual features and noise features. By replacing a source speaker's joint audio-visual feature with a target speaker's audio feature, the voice individuality of a source speaker was converted to that of a target speaker. However, this method requires that the activity of each atom in the dictionary be estimated and consequently requiring a high computation cost.

In this paper, motivated by the feature-extraction ability of convolutional neural networks (CNNs) [23], [24] and the nonlinear mapping ability of DNNs [25], [26], we propose a multimodal voice conversion method based on a deep convolutional neural network (MDCNN) for VC in a noisy environment. We utilize a CNN to extract visual features from lip movement sequences. In addition, two convolutional kernels of different sizes are used to effectively extract audio features. Subsequently, the extracted audio and visual features are merged and fed into a fully connected neural network (FCNN) trained by the backpropagation algorithm [28] to obtain the corresponding converted speech waveform. The effectiveness of the proposed method was evaluated by comparing it to the NMF-based multimodal VC method [22] and the conventional DNN-based method [27]. In this paper, the MDCNN was implemented by software on a computer. For portable applications, the MDCNN can be realized and implemented by simple hardware as described in [28]–[35].

The rest of this paper is organized as follows: In Section II, the details of the proposed MDCNN are presented. Section III describes the data preparation and evaluation methods. In Section IV, the experimental conditions and results are presented. The final section is devoted to conclusions.

II. MDCNN ARCHITECTURE

Voice conversion changes a source speech by modifying the acoustic characteristics of the speech signal while preserving its linguistic details. VC methods generally only focus on acoustic information [36], [37] without visual information. Recently, methods for processing speech signals using multimodal information have attracted the attention of researchers [19]–[21]. The experimental results obtained show that the addition of visual information can improve the performance of the MDCNN under noisy environments. In this paper, visual information is added to the voice conversion, and the proposed multimodal voice conversion model is shown in Fig. 1.

In Fig. 1, mel-frequency cepstral coefficients (MFCC) are adopted as speech acoustic features to reduce the model complexity. MFCC_s and MFCC_t were extracted from a source speech and a target speech, respectively. Since speech features are typically speaker-dependent, the duration of speech between two people is always different. Therefore, we adopt dynamic time warping (DTW) to align the corresponding source and target speech features to obtain aligned acoustic features. MFCC_s' and MFCC_t' were aligned acoustic features of the corresponding source speech and target speech, respectively. Since both the frame rates of the audio and the video from a source speaker were set to 60 frames per second (fps), the acoustic feature and the visual feature of a source speech were aligned automatically. The aligned acoustic features of a source speech were also used to align the corresponding source lip images. Two-dimensional discrete cosine transform (DCT) was used to transform the lip image of a current frame, which was then straightened by zigzagging

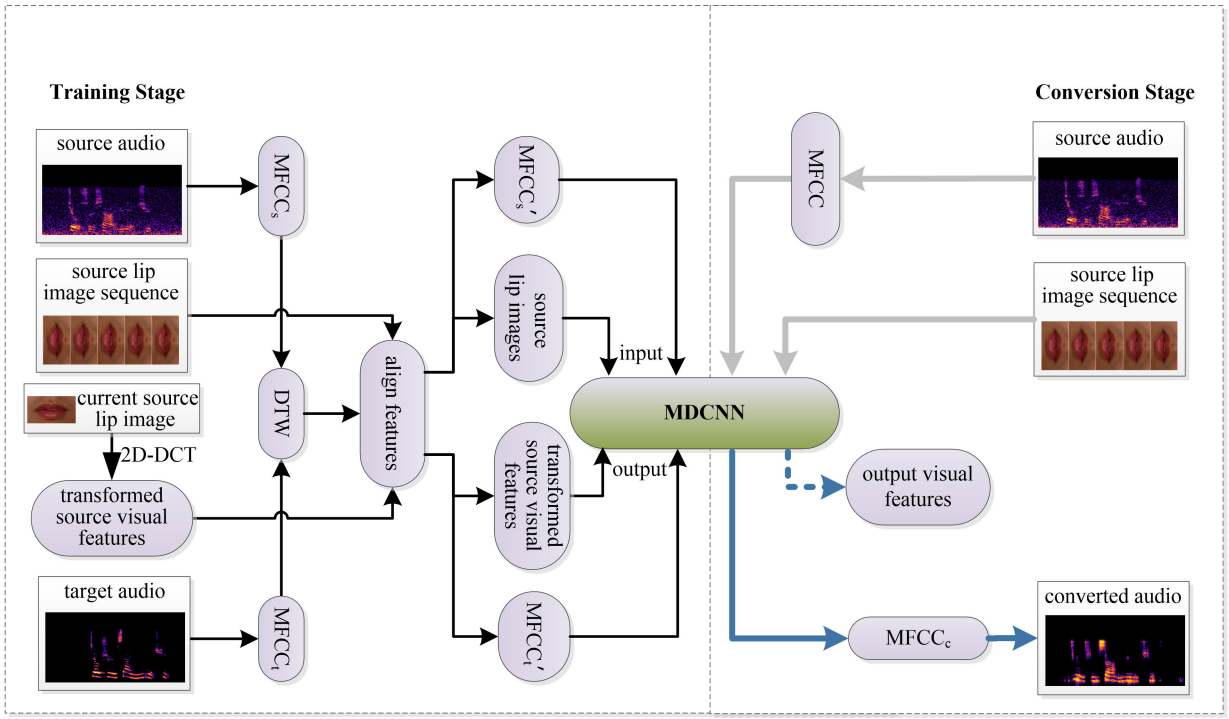


FIGURE 1. Block diagram of the proposed multimodal voice conversion model.

to obtain a 1-dimensional DCT vector shown as transformed source visual features in Fig. 1.

For training, the aligned lip images and the MFCC$_s$' of a source speaker were input to the MDCNN. Additionally, the aligned MFCC$_t$' was used as the output of the MDCNN in the training phase. It has been shown that using the transformed visual features of a source speech as the output of a deep learning network can enhance speech by suppressing noise [18]. The transformed visual features of the output of the MDCNN were extracted from a source speaker. Note that the mapping provided by the transformed source visual features at the output serves to suppress noise to improve the robustness of the MDCNN.

The MDCNN can characterize and learn the nonlinear mapping relation between the audio-visual features of a source speech and the audio features of a target speech. At the conversion stage, the audio and visual features were extracted from a source speaker and then sent to the MDCNN to obtain the converted acoustic features MFCC$_c$. As shown in Fig. 2, the architecture of the proposed MDCNN comprises two CNNs and one DNN. We used one CNN to extract the acoustic features of a source speech to reflect interframe correlation and another CNN to extract the interframe features of the visual information of the source speech.

Note that CNN is often formatted to accept two-dimensional input, so the MFCC$_s$ of every 5 successive frames of each speech were used to form an input to the audio CNN. The target output of the MDCNN is the MFCC$_t$ of the clean speech from a current frame. Since audio signals are the main information for audio-visual voice conversion, we adopted two convolution kernels of different sizes for a source audio to capture two feature maps to improve the performance of the MDCNN. As shown in Fig. 3, the audio CNN in the proposed MDCNN adopted a bichannel convolution architecture to effectively extract audio features [38]. In Fig. 2 and Fig. 3, the Merge1 module concatenated the outputs of previous pooling layers along z-index. Since lip images were introduced only for supplementary information, only one convolution kernel was used for processing the lip images. The joint visual-audio features were derived by concatenating each column vector of the 2-dimensional visual feature matrix obtained from Pool5 and the 2-dimensional audio feature matrix obtained from Conv9. The joint visual-audio features were then sent to a four-layer DNN to map the nonlinear relation between the audio-visual features of the source speech and the acoustic features of its target speech.

In the conversion stage, for a given source speech, two CNNs were used to extract their audio features and visual features, which are then input to the MDCNN to obtain the features of the converted speech. The mapped MFCC$_c$ features of the converted speech undergo an inverse transform from which the waveform of the converted speech was synthesized [39].

The clean target speech as well as the noisy source speech and the lip images of a source speaker were used to train the MDCNN. Let $\boldsymbol{n}_i$ and $\boldsymbol{C}_i$ denote the *ith* frame audio features extracted from the *ith* frame noisy source and the *ith* frame

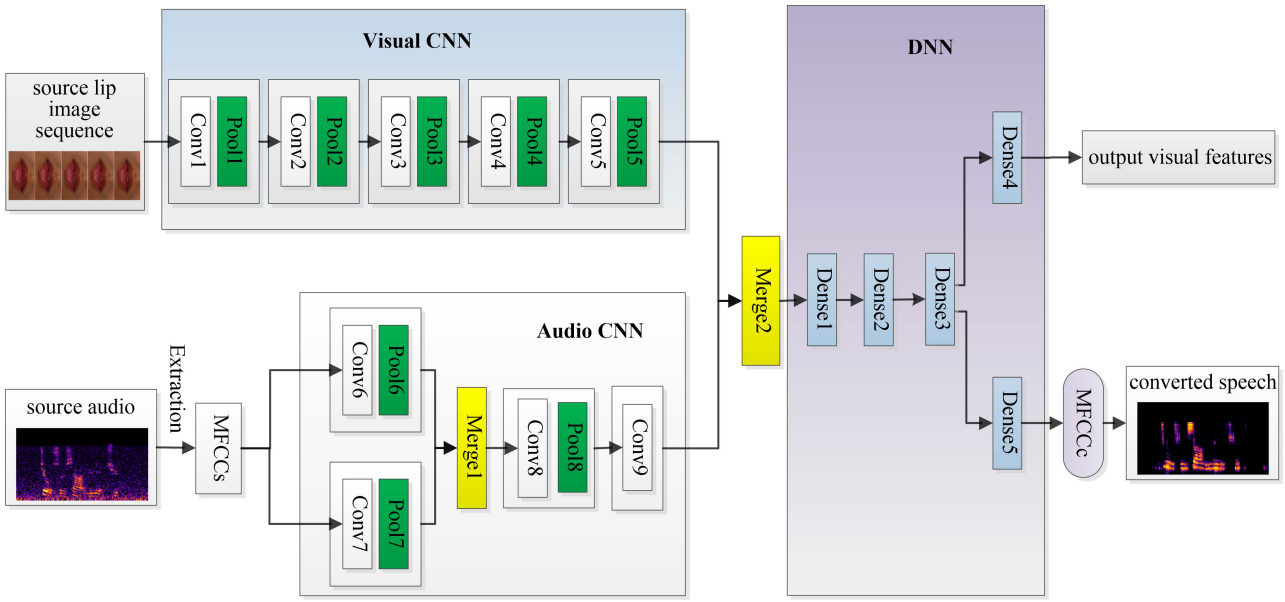


FIGURE 2. Architecture of the MDCNN.

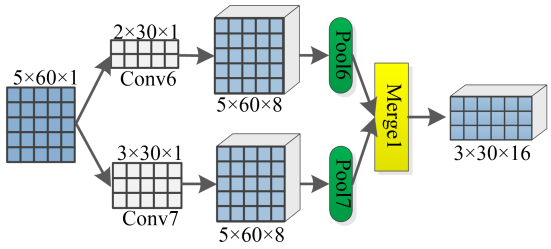


FIGURE 3. Architecture of the Audio CNN.

clean target speech, respectively, and $\boldsymbol{N}_i = [\boldsymbol{n}_{i-2}^T, \boldsymbol{n}_{i-1}^T, \boldsymbol{n}_i^T, \boldsymbol{n}_{i+1}^T, \boldsymbol{n}_{i+2}^T]$. Let $\boldsymbol{p}_i$ represent the *ith* frame lip image corresponding to the *ith* frame noisy source speech, and $\boldsymbol{V}_i$ denote the *ith* frame visual features transformed from $\boldsymbol{p}_i$, $\boldsymbol{P}_i = [\boldsymbol{p}_{i-2}^T, \boldsymbol{p}_{i-1}^T, \boldsymbol{p}_i^T, \boldsymbol{p}_{i+1}^T, \boldsymbol{p}_{i+2}^T]$.

Let K represent the number of samples, *Poolj* denote the max-pooling operation of the *jth* max-pooling layer, *Convj* denote the convolutional operation of the *jth* convolution layer, and *Densej* denote the fully connected operation of the *jth* fully connection layer. In the MDCNN, the extracted visual features $\vec{\boldsymbol{P}}_i$ can be computed as

$$\vec{\boldsymbol{P}}_i = Pool5(Conv5(Pool4(Conv4(Pool3(Conv3(Pool2(Conv2(Pool1(Conv1(\boldsymbol{P}_i)))))))))), \quad i = 1, 2, 3 \ldots\ldots K. \tag{1}$$

The extracted audio features $\vec{\boldsymbol{N}}_i$ can be computed as

$$\vec{\boldsymbol{N}}_i = Conv9(Pool8(Conv8(Pool6(Conv6(\boldsymbol{N}_i)) + Pool7(Conv7(\boldsymbol{N}_i))))), \quad i = 1, 2, 3 \ldots\ldots K. \tag{2}$$

Let $\boldsymbol{F}_i = [\vec{\boldsymbol{P}}_i; \vec{\boldsymbol{N}}_i]$, the converted speech feature $\vec{\boldsymbol{C}}_i$ and the visual feature $\vec{\boldsymbol{V}}_i$ of the *ith* frame are computed as

$$\vec{\boldsymbol{C}}_i = Dense5(Dense3(Dense2(Dense1(\boldsymbol{F}_i)))), \quad i = 1, 2, 3 \ldots\ldots K.$$

$$\vec{\boldsymbol{V}}_i = Dense4(Dense3(Dense2(Dense1(\boldsymbol{F}_i)))), \quad i = 1, 2, 3 \ldots\ldots K. \tag{3}$$

The mean square error (MSE) [40] is adopted as the loss function for the MDCNN as defined by

$$loss = \left(\frac{1}{K}\sum_{i=1}^{K}(\|\boldsymbol{C}_i - \vec{\boldsymbol{C}}_i\|^2) + \omega * \frac{1}{K}\sum_{i=1}^{K}(\|\boldsymbol{V}_i - \vec{\boldsymbol{V}}_i\|^2)\right). \tag{4}$$

where ω is the mixed weight, which is used to regulate the impact of visual information on voice conversion.

III. DATA PREPARATION AND EVALUATION METHODS

A. DATA PREPARATION

To evaluate the performance of the proposed MDCNN, 300 (100 × 3) clean utterances spoken by three speakers were selected from the Audio-Visual Whisper Database (AVWD) recorded by us in a quiet room due to an absence of a suitable open access multimodal corpus database. The database consists of 100 utterances from a male speaker and 200 utterances from two female speakers. The AVWD contains video recordings of 100 utterances of Chinese sentences, each of which is spoken by 5 native female speakers and 5 native male speakers, generating 1,000 (100 × 10) utterances. The length of each utterance is approximately 2-3 seconds. Videos were recorded at a sampling rate of 30 fps with a resolution of 1920 × 1080 in pixels. The speech signals were recorded with a sampling rate of 44.1 kHz. The AVWD corpus database is open access and available at ftp://210.45.212.96 with username: download, and password: download.

The audio signals were resampled to 16 kHz. The time-frequency audio spectrum was extracted by taking the squared magnitude of the short-time discrete Fourier transform (DFT). A fixed frame size of 32 ms was used with 48% overlap between frames. A 512-point DFT was used to

calculate the short-term power spectrum of the speech signal of each frame. In the MFCC computing process, the speech spectrum of each frame was passed through a series of triangular filters that were spaced linearly in a perceptual mel scale. The mel filter bank log energy (MFLE) of each of the filters was computed. Finally, MFCC was computed by taking the DCT of the MFLE. For visual information, the video was increased to 60 fps using FFMPEG [41] software to synchronize the number of audio and visual frames. The Dlib tool [42] was used to detect the lip area. The size of the lip images was adjusted to 32×64 in pixels.

Noises were selected from the NOISEX-92 standard noise library [43], which contains 15 kinds of real scene noise. Source clean utterances spoken by a male and a female were artificially contaminated by 6 kinds of noise (Volvo, Gaussian white noise, factory noise, pink noise, F16 double cockpit noise and HF channel noise) at 7 different signal-to-noise ratios (SNRs) of -5, -3, 0, 3, 5, 10, and 20 dB. Ninety percent utterances were randomly selected as the training set, and the remaining 10% was used as the testing set.

Male to female and female to female voice conversions were performed by different conversion methods for performance evaluation. In these two conversions, we used the same target female speaker.

B. EVALUATION METHODS

Mel cepstral distance (CD), perceptual evaluation of speech quality (PESQ), and short-time objective intelligibility (STOI) were used to evaluate the performance of different voice conversion methods objectively. The CD value [44] is a common objective evaluation method for speech quality. The formula for the calculation of CD value can be written as

$$CD = (10/ln(10)) \sqrt{2 \sum_{d=1}^{D} (C_d - C'_d)^2}. \quad (5)$$

where C_d and C'_d represent the dth MFCC of the reference source speech and the converted speech, respectively. D represents the dimension of MFCC, which was set to 26 in the following experiments.

The average CD value of all the frames of a speech (utterance) was used as the CD value for the speech. A higher CD value indicates a greater difference between a converted and a reference source speech. Therefore, the lower the CD value, the better the performance of a conversion. The PESQ [45] score, which ranges from -0.5 to 4.5, denotes the overall quality of a converted speech. The STOI [46] score is an indicator of speech intelligibility, and it ranges from 0 to 1. Larger PESQ and STOI indicate better quality and higher intelligibility of a converted speech.

For subjective evaluation, 5 males and 5 females participated in the listening tests. The ABX preference test [47] and the mean opinion score (MOS) were adopted to evaluate the naturalness, speech similarity, and quality of converted speech. In the ABX preference test, the participants were asked to choose which of the converted speech is (a) more similar to the target speech and (b) more natural. If the difference between two converted speeches was small, "fair" was chosen. The MOS score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad).

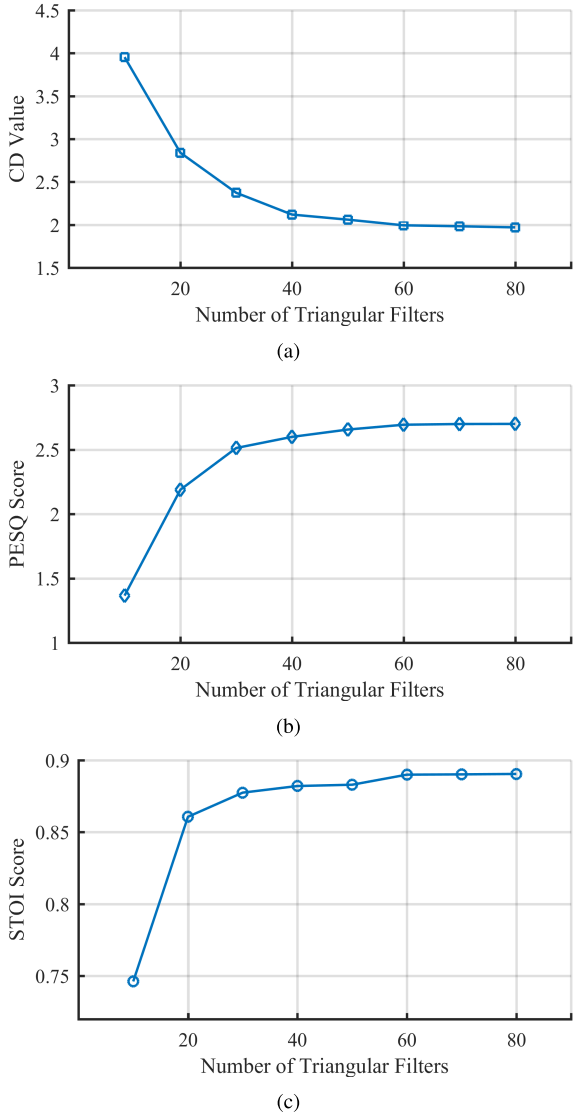


FIGURE 4. CD, PESQ, and STOI performance using different numbers of triangular filters on reconstructed speech. (a) CD with different numbers of triangular filters; (b) PESQ with different numbers of triangular filters; (c) STOI with different numbers of triangular filters.

IV. EXPERIMENTS

A. EXPERIMENTAL CONDITIONS

Fig. 4 shows the effect of different numbers of triangular filters on speech reconstruction. With an increase in the number of triangular filters, the quality of reconstructed speech further improved. When the number of triangular filters reached 60, the conversion performance was basically stable. Therefore, we set the number of triangular filters to 60 for MFCC extraction in the following experiments.

For visual features, the lip image of the current frame was transformed by 2D-DCT, which was then straightened by

zigzagging to obtain a 1-dimensional DCT vector of 50 coefficients.

Since the DNN module of the MDCNN was formatted to accept one-dimensional data as input and output, the data output through the convolutional layer is multi-dimensional data. Therefore, after audio CNN and Visual CNN, it was necessary to straighten and reduce the dimensions of the previous output to obtain one-dimensional data. The visual CNN output was straightened to 1,280 ($1 \times 10 \times 128$), and the audio CNN output was straightened to 960 ($2 \times 15 \times 32$). Then, these two inputs were combined to obtain 2240-dimensional data as DNN input.

TABLE 1. Configuration of the Proposed MDCNN (Keys: Cf = Convolutional filter; Nf = Number of filters).

Layer	Kernel	Input	Cf Size	Nf	Output
Conv1	3×3	32×320×3	3×3×3	32	32×320×32
Pool1	2×2	32×320×32	-	-	16×160×32
Conv2	3×3	16×160×32	3×3×32	32	16×160×32
Pool2	2×2	16×160×32	-	-	8×80×32
Conv3	3×3	8×80×32	3×3×32	64	8 ×80×64
Pool3	2×2	8×80×64	-	-	4×40×64
Conv4	3×3	4×40×64	3×3×64	128	4×40×128
Pool4	2×2	4×40×128	-	-	2×20×128
Conv5	1×1	2×20×128	1×1×128	128	2×20×128
Pool5	2×2	2×20×128	-	-	1×10×128
Conv6	2×30	5×60×1	2×30×1	8	5×60×8
Pool6	2×2	5×60×8	-	-	3×30×8
Conv7	3×30	5×60×1	3×30×1	8	5×60×8
Pool7	2×2	5×60×8	-	-	3×30×8
Merge1	-	3×30×8	-	-	
	-	3×30×8	-	-	3×30×16
Conv8	2×15	3×30×16	2×15×16	32	3×30×32
Pool8	2×2	3×30×32	-	-	2×15×32
Conv9	1×1	2×15×32	1×1×32	32	2×15×32
Merge2	-	1280	-	-	
	-	960	-	-	2240
Dense1	-	2240	-	-	512
Dense2	-	512	-	-	512
Dense3	-	512	-	-	512
Dense4	-	512	-	-	50
Dense5	-	512	-	-	60

The overall configuration of the proposed MDCNN is listed in Table 1. For the proposed MDCNN, the following settings were applied: batch size was set to 100, the stride of each convolutional layer was set to 1×1, the stride of each pooling layer was set to 2×2, L2 regularization was used to avoid overfitting, bias was used in each of the convolutional layers and each of the dense layers. The activation function used in each of the convolutional layers, Dense1, Dense2, and Dense3, was ReLU, and the activation function used in each of Dense4 and Dense5 was tanh [34]. Dropout was set to 0.3 and only used at Dense1, Dense2, and Dense3.

To determine a suitable value for the mixing weight ω in (4), we selected 600 utterances with a SNR of 0 dB from 4,200 noisy speech samples. As shown in Table 2, when ω was set to 0.3, the lowest CD value was obtained from a converted speech, with the highest PESQ and STOI scores. Since the visual lip image was only used as auxiliary information, if ω is too large, the network will overfit, resulting in performance degradation of the voice conversion. Hence, ω was set to 0.3 in the following experiments.

TABLE 2. Conversion performance using different mixed weights ω in loss function defined in (4).

ω	CD	PESQ	STOI
0	6.91	0.70	0.41
0.3	6.73	0.83	0.42
0.5	6.88	0.74	0.42
1.0	6.91	0.71	0.41
10	7.04	0.64	0.40

B. BASELINE METHODS

To verify the effectiveness of the proposed MDCNN, DNN for audio feature mapping only (referred to as DNN-COM) [27] and multimodal nonnegative matrix factorization (referred to as MUL-NMF)-based voice conversion [22] were selected as the baseline methods.

For the MUL-NMF-based voice conversion, the audio spectral envelope was extracted from the STRAIGHT [48] model. The audio spectrum envelope and the visual feature, which are the same as in the MDCNN, were concatenated for each frame to obtain a joint audio-visual feature. The voice conversion method was the same as in [22]. The joint audio-visual feature was used to form the source feature $\boldsymbol{A} = [\boldsymbol{A}_s, \boldsymbol{A}_n]$. Input noisy audio-visual features $\boldsymbol{X}$ were decomposed into a linear combination of the clean audio-visual feature $\boldsymbol{A}_s$ and the noise feature $\boldsymbol{A}_n$. $\boldsymbol{H}_s$ and $\boldsymbol{H}_n$ represent the coefficient matrix of source and noise, respectively. $\boldsymbol{H}$ is a coefficient matrix in NMF. By replacing the source speaker's audio-visual feature $\boldsymbol{A}_s$ with the target speaker's audio feature $\boldsymbol{A}_t$, the voice individuality $\boldsymbol{X}_t$ of the source speaker is converted to the target speaker.

$$\begin{aligned} \boldsymbol{X} &\approx [\boldsymbol{A}_s \boldsymbol{A}_n] \begin{bmatrix} \boldsymbol{H}_s \\ \boldsymbol{H}_n \end{bmatrix} s.t. \quad \boldsymbol{H}_s, \boldsymbol{H}_n \geq 0 \\ &= \boldsymbol{A}\boldsymbol{H} \quad s.t. \quad \boldsymbol{H} \geq 0. \end{aligned} \tag{6}$$

$$\boldsymbol{X}_t = \boldsymbol{A}_t \boldsymbol{H}_s. \tag{7}$$

For the DNN-COM-based voice conversion method, the following settings were applied: MFCC of 5 successive frames (the current frame and $\pm$ 2 frames) were concatenated as the input. The target output was the MFCC of the clean speech from the current frame. There were 300 (60×5) nodes in the source audio input layer and 60 nodes in the target audio output layer. The number of hidden layers was set to 3, with 512 nodes in each layer. A dropout rate of 0.3 was adopted, with a batch size of 100, and L2 regularization was used to avoid overfitting. The output layer used tanh as the activation function, and the hidden layers used ReLU as the activation function. A bias was used in each layer. Both the MDCNN and the DNN-COM were trained by the backpropagation algorithm [28].

C. EXPERIMENTAL RESULTS AND DISCUSSION

From Fig. 5, we can see that when SNR is high, the CD value of the converted speech based on the MDCNN and

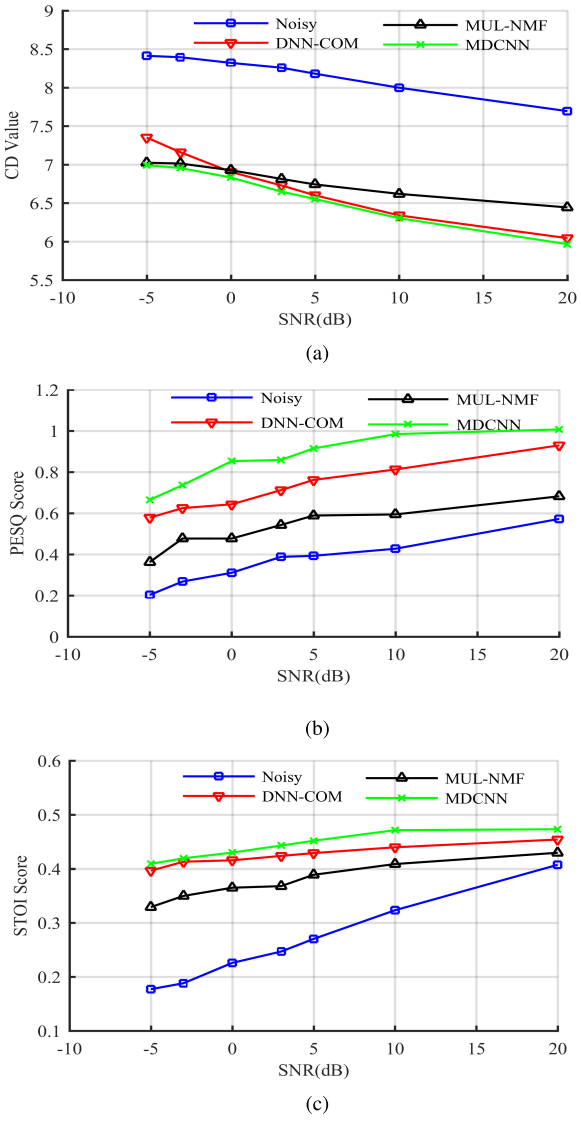


FIGURE 5. Comparison of objective performance of voice conversion methods. (a) CD value under different SNRs; (b) PESQ score under different SNRs; (c) STOI score under different SNRs.

the DNN-COM were lower than that of the MUL-NMF. This result was attributed to the nonlinear fitting ability of neural networks. However, for SNR less than 0 dB, the CD values of the MUL-NMF and the MDCNN were better than those of the DNN-COM. This may be attributed to the effectiveness of the auxiliary visual information in the MDCNN. With a decrease in SNR, visual information in the MDCNN improved the robustness of voice conversion. In Fig. 5(a), the CD values of the converted speech based on the MDCNN were lower than that of the MUL-NMF. This means that the performance of the proposed MDCNN was better than that of the MUL-NMF with audio-visual information. Fig. 5(b) and Fig. 5(c) show that the MDCNN performs best in terms of PESQ and STOI.

Fig. 6(a)-(e) presents the spectrograms of clean target speech, noisy source speech, and converted speech based on the MUL-NMF, the DNN-COM, and the MDCNN.

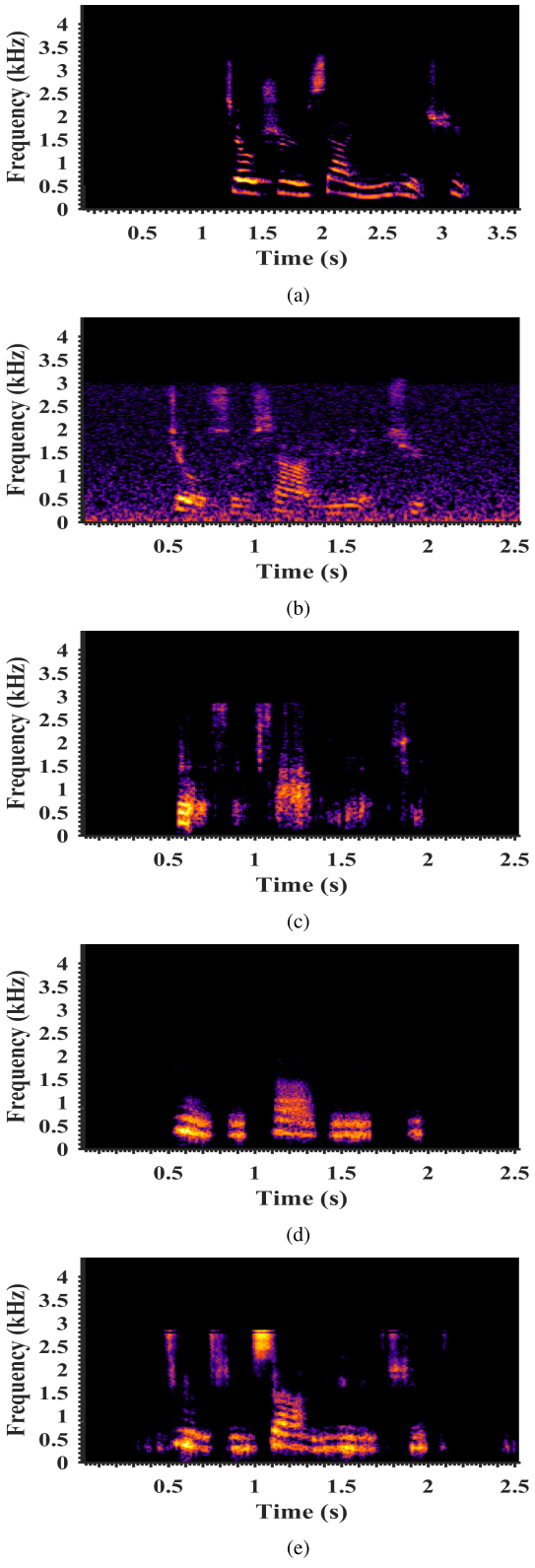


FIGURE 6. Comparison of spectrograms. (a) Clean target speech; (b) Noisy speech with pink noise at 5 dB SNR; (c) MUL-NMF; (d) DNN-COM; (e) MDCNN.

Although the MUL-NMF approach retained abundant high-frequency information, it could not efficiently capture low frequency information, resulting in a fuzzy voiceprint as

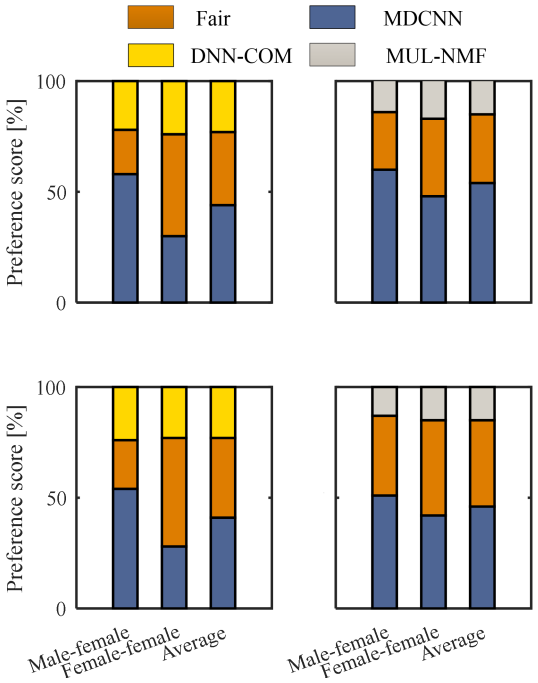


FIGURE 7. ABX preference test of speaker similarity (top) and speech naturalness (bottom).

shown in Fig. 6(c). The DNN-COM method obtained a clearer voiceprint as shown in Fig. 6(d) than the MUL-NMF as shown in Fig. 6(c). However, the DNN-COM could not effectively restore the high-frequency information.

Compared to the MUL-NMF and the DNN-COM, the MDCNN as shown in Fig. 6(e) obtained the clearest voiceprint that was the most similar to the target speech and the high-frequency details were also well captured.

TABLE 3. MOS Results of Converted Voices using Different Methods.

Conversion	MUL-NMF	DNN-COM	MDCNN (Proposed)
Male-female	2.51	2.95	3.11
Female-female	2.49	2.87	2.93

Experimental results for subjective hearing tests from 5 males and 5 females are displayed in Fig. 7 and listed in Table 3. The ABX preference test of speaker similarity (top) and speech naturalness (bottom) are shown in Fig. 7. For the top-left subplot in Fig. 7, the blue bar denotes the percentage of participants who preferred the voice converted by the MDCNN. The yellow bar denotes the percentage of participants who preferred the voice converted by the DNN-COM. For the top-right subplot in Fig. 7, the gray bar denotes the percentage of participants who preferred the voice converted by the MUL-NMF. Throughout Fig. 7, the orange bar denotes the percentage of participants who had no preference and did not distinguish which one of the two converted voices was more similar to the ground truth target.

The results shown in Fig. 7 indicate that the MDCNN outperformed the MUL-NMF and the DNN-MUL in terms

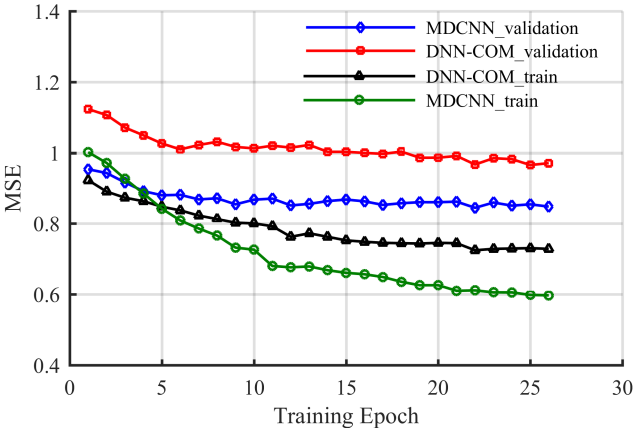


FIGURE 8. MSE versus training epoch for DNN-COM and MDCNN on training set and validation set.

of both speaker similarity and speech naturalness, which may due to the fusion features extracted by the MDCNN can represent the original information better; and also the use of lip images as supplementary information can improve the robustness of voice conversion in noisy environments. From Table 3, the subjective evaluation of the MOS results indicates that the quality of converted voice obtained by the MDCNN was better than those obtained by the two state-of-the-art methods.

Fig. 8 compares the training and validation MSEs of the MDCNN and the DNN-COM versus epoch on both the training set and the validation set. The two methods were evaluated with 5% of the constructed data as the validation set and there is no intersection between the validation set and training set. We observed that the MDCNN converged faster and achieved lower MSEs than those of the DNN-COM.

TABLE 4. Comparison of Objective Evaluations between CNN (audio-only) and MDCNN.

SNR(dB)	-5	-3	0	3	5	10	20
CD (CNN)	7.01	6.99	6.89	6.69	6.56	6.32	6.00
CD (MDCNN)	6.99	6.95	6.83	6.64	6.54	6.30	5.96
PESQ (CNN)	0.58	0.67	0.68	0.74	0.79	0.82	0.95
PESQ (MDCNN)	0.66	0.74	0.85	0.86	0.92	0.99	1.01
STOI (CNN)	0.39	0.41	0.42	0.43	0.43	0.44	0.44
STOI (MDCNN)	0.40	0.42	0.43	0.45	0.45	0.47	0.47

The DNN-COM voice conversion method uses only speech audio feature which cannot characterize the relationship between successive frames but can be tackled by the MDCNN. In addition, we compared the CNN-based voice conversion method using only audio information with the MDCNN to verify the contribution of the visual information. The CNN was constructed by removing the visual CNN and the visual output in the MDCNN, and the experimental conditions of this CNN were the same as the MDCNN. The experimental results are shown in Table 4. The performance of the MDCNN was better than the CNN in terms of CD values, PESQ scores, and STOI scores under different SNRs.

The results indicate that the addition of visual information can improve the performance of voice conversion.

V. CONCLUSION

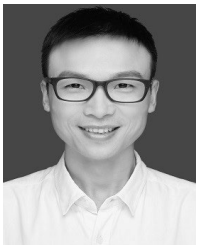
We have presented a multimodal deep convolutional neural network for voice conversion tasks under noisy environments. The MDCNN aims to improve the intelligibility, quality, and naturalness of a converted speech using the visual information of the corresponding source lip movement as auxiliary information on top of the audio information of a source audio. Two different CNNs were adopted to extract visual and acoustic features, which were then merged and fed into a deep neural network. Experimental results show that the performance of the proposed MDCNN outperformed the MUL-NMF and the DNN-COM in both subjective and objective evaluations under a variety of noise environments.

The converted voice audios are posted on the demo website at http://101.37.150.44:8088/hyt.aspx.

REFERENCES

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, Apr. 2017.
[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.
[3] S. W. Fu, P. C. Li, Y. H. Lai, C. C. Yang, L. C. Hsieh, and Y. Tsao, "Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 11, pp. 2584–2594, Nov. 2017.
[4] R. Aihara, T. Takiguchi, and Y. Ariki, "Multiple non-negative matrix factorization for many-to-many voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1175–1184, Jul. 2016.
[5] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Salt Lake City, UT, USA, May 2001, pp. 841–844.
[6] L. J. Liu, L. H. Chen, Z. H. Ling, and L. R. Dai, "Spectral conversion using deep neural networks trained with multi-source speakers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 4849–4853.
[7] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-constrained trajectory training algorithm for GMM-based Voice Conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 4859–4863.
[8] T. Hashimoto, D. Saito, and N. Minematsu, "Many-to-many and completely parallel-data-free voice conversion based on eigenspace DNN," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 2, pp. 332–341, Feb. 2019.
[9] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory HMMs," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 2, pp. 417–430, Feb. 2011.
[10] J. Zhou, Y. Dou, R. Liu, H. Wang, and L. Tao, "Whisper to normal conversion based on low dimension feature mapping," *Acta Acustica*, vol. 43, no. 5, pp. 855–863, Sep. 2018.
[11] L. Chen, Z. Ling, L. Liu, and L. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1859–1872, Dec. 2014.
[12] W. Ye and Y. Yu, "Voice conversion using deep neural network in superframe feature space," in *Proc. 6th Int. Conf. Intell. Control Inf. Process. (ICICIP)*, Wuhan, China, Nov. 2015, pp. 465–468.
[13] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 4869–4873.
[14] H. Q. Nguyen, S. W. Lee, X. Tian, M. Dong, and E. S. Chng, "High quality voice conversion using prosodic and high-resolution spectral features," *Multimedia Tools Appl.*, vol. 75, no. 9, pp. 5265–5285, May 2015.
[15] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 3, pp. 806–817, Mar. 2012.
[16] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.
[17] H. Meutzner, N. Ma, R. Nickel, C. Schymura, and D. Kolossa, "Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 5320–5324.
[18] J. C. Hou, S. S. Wang, Y. H. Lai, Y. Tsao, H. W. Chang, and H. M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.
[19] Y. Yuan, C. Tian, and X. Lu, "Auxiliary loss multimodal GRU model in audio-visual speech recognition," *IEEE Access*, vol. 6, pp. 5573–5583, 2018.
[20] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.
[21] J. S. Brumberg, K. M. Pitt, and J. D. Burnison, "A noninvasive brain-computer interface for real-time speech synthesis: The importance of multimodal feedback," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 874–881, Apr. 2018.
[22] K. Masaka, R. Aihara, T. Takiguchi, and Y. Ariki, "Multimodal voice conversion using non-negative matrix factorization in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 1542–1546.
[23] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Proc. 28th Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2014, pp. 2042–2050.
[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[25] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 9, pp. 1469–1477, Sep. 2015.
[26] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 352–364, Feb. 2018.
[27] A. K. Bhuyan and J. H. Nirmal, "Comparative study of voice conversion framework with line spectral frequency and Mel-Frequency Cepstral Coefficients as features using artficial neural networks," in *Proc. Int. Conf. Comput., Commun., Syst. (ICCCS)*, Kanyakumari, India, Nov. 2015, pp. 230–235.
[28] C. Z. Tang and H. K. Kwan, "Parameter effects on convergence speed and generalization capability of backpropagation algorithm," *Int. J. Electron.*, vol. 74, no. 1, pp. 35–46, Jan. 1993.
[29] H. K. Kwan, "Multiplierless designs for artificial neural networks," in *Neural Networks and Systolic Array Design* (Machine Perception and Artificial Intelligence), vol. 49, D. Zhang and S. K. Pal, Eds. Singapore: World Scientific, Jun. 2002, ch. 13, pp. 301–325, doi: 10.1142/9789812778086_0013.
[30] H. K. Kwan and C. Z. Tang, "Multiplierless multilayer feedforward neural network design using quantised neurons," *Electron. Lett.*, vol. 38, no. 13, pp. 645–646, Jun. 2002.
[31] C. Z. Tang and H. K. Kwan, "Multilayer feedforward neural networks with single powers-of-two weights," *IEEE Trans. Signal Process.*, vol. 41, no. 8, pp. 2724–2727, Aug. 1993.
[32] H. K. Kwan and C. Z. Tang, "Multiplierless multilayer feedforward neural network design suitable for continuous input-output mapping," *Electron. Lett.*, vol. 29, no. 14, pp. 1259–1260, Jul. 1993.
[33] H. K. Kwan and C. Z. Tang, "Designing multilayer feedforward neural networks using simplified sigmoid activation functions and one-powers-of-two weights," *Electron. Lett.*, vol. 28, no. 25, pp. 2343–2344, Dec. 1992.
[34] H. K. Kwan, "Simple sigmoid-like activation function suitable for digital hardware implementation," *Electron. Lett.*, vol. 28, no. 15, pp. 1379–1380, Jul. 1992.
[35] H. K. Kwan, "One-layer feedforward neural network for fast maximum/minimum determination," *Electron. Lett.*, vol. 28, no. 17, pp. 1583–1585, Aug. 1992.

[36] Z. Wu, T. Virtanen, E. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1506–1521, Oct. 2014.
[37] T. Nakashika and Y. Minami, "Speaker adaptive model based on Boltzmann machine for non-parallel training in voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5530–5534.
[38] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, 2014, pp. 1746–1751.
[39] L. E. Boucheron, P. L. De Leon, and S. Sandoval, "Low bit-rate speech coding through quantization of mel-frequency cepstral coefficients," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 610–619, Feb. 2012.
[40] F.-L. Xie, Y. Qian, F. K. Soong, and H. Li, "Pitch transformation in neural network based voice conversion," in *Proc. 9th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Singapore, Sep. 2014, pp. 197–200.
[41] FFmpeg Team. *Fast Forward Mpeg*. Accessed: Nov. 6, 2018. [Online]. Available: http://ffmpeg.org/download.html
[42] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1867–1874.
[43] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
[44] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 2, pp. 242–248, Feb. 1988.
[45] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, UT, USA, May 2001, pp. 749–752.
[46] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 4214–4217.
[47] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "ATTS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 6805–6809.
[48] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Munich, Germany, Apr. 1997, pp. 1303–1306.



JIAN ZHOU received the Ph.D. degree in information and communication engineering from Southeast University, China, in 2013. Since 2014, he has been an Associate Professor with the School of Computer Science and Technology, Anhui University, China. His main research interests include speech and image processing and pattern recognition.

YUTING HU received the B.S. degree in information and computing science from the Anhui University of Science and Technology, China, in 2017. She is currently pursuing the M.S. degree in computer science and technology with Anhui University. Her main research interests include speech signal processing and image processing.

HAILUN LIAN received the B.S. degree in computer science and technology from Anhui Jianzhu University, China, in 2017. He is currently pursuing the M.S. degree in computer science and technology with Anhui University, China. His research interests include speech signal processing and deep learning.

HUABIN WANG received the B.S. degree in computer science and technology from the Anhui University of Finance and Economics, China, in 2005, the M.S. degree in signal and information processing and the Ph.D. degree in computer application technology from Anhui University, Hefei, China, in 2008 and 2011, respectively. He is currently the Deputy Director of the Department of Computer Science and Technology, Anhui University. His research interests include face recognition, virtual reality, and signal processing.

LIANG TAO received the Ph.D. degree in information and communication engineering from the University of Science and Technology of China, China, in 2003. From August 1998 to August 1999, he was a Visiting Scholar with the University of Windsor, Windsor, ON, Canada, supported by the China Scholarship Council. He is currently a Professor with the School of Computer Science and Technology, Anhui University, China. He has published over 100 articles. His main research interests include digital signal and image processing and pattern recognition.

HON KEUNG KWAN (M'81–SM'86–LSM'18) received the D.I.C. and Ph.D. degree in electrical engineering (signal processing) from Imperial College London, U.K., in 1981. His previous experiences include working as a Design Engineer in electronics and computer memory industry, from 1977 to 1978, and serving as a Faculty Member with the Department of Electronic Engineering, The Hong Kong Polytechnic University, in 1981, and then with the Department of Electrical and Electronic Engineering, The University of Hong Kong. He subsequently joined the University of Windsor, where he has been a Professor of electrical and computer engineering, since 1989. His current research interests include digital filter and deep neural network design. He is a licensed Professional Engineer (ON), a Chartered Electrical Engineer (U.K.), and he was elected as a Fellow of the Institution of Engineering and Technology (U.K.), in 1996. He had served as the Chair and various officers for each of the Digital Signal Processing Technical Committee and the Neural Systems and Applications Technical Committee of the IEEE Circuits and Systems Society.