Auditory Feature for Monaural Speech Segregation

Yi Jiang, Runsheng Liu
Department of Electronic Engineering
Tsinghua University
Beijing, P.R. China
e-mail: jiangyi09@mails.tsinghua.edu.cn

Yuanyuan Zu
Quartermaster Equipment Research Institute
General Logistics Department
Beijing, P.R. China
e-mail: yvette_zu@126.com

***Abstract*—Monaural speech segregation has been a very challenging problem for speech signal processing. The implication of the ideal binary masks to an auditory mixture has been shown to yield substantial improvements in signal-to-noise-ratio (SNR) and intelligibility. In this paper, we use the time-frequency (T-F) unit level gammatone frequency cepstral coefficients (GFCC) auditory feature to estimate the ideal binary mask for monaural speech segregation. The paper reports the successful attempt to use GFCC as the segregation cue with deep neural networks (DNNs) classifier. Results show that robust performance can be achieved across noisy and reverberant conditions.**

Keywords-gammatone frequency cepstral coefficients (GFCC); monaural speech segregation; binary classification; time-frequency(T-F) unit

I. INTRODUCTION

Monaural speech segregation has been a very challenging problem for decades, while human listening excel in "hearing out" a target source from sound mixtures in noisy and reverberant conditions. Inspired by human auditory processing, computational auditory scene analysis (CASA) [1] aims to separate a mixture of sources into different auditory streams based on perceptual principles, and has shown considerable promise in speech segregation. The ideal binary mask (IBM) is a time-frequency (T-F) binary mask, constructed from premixed target and interference. A mask value 1 for a T-F unit indicates that the signal-to-noise ratio (SNR) within the unit exceeds a threshold (target-dominant), and 0 otherwise (interference-dominant). The IBM has been suggested as a primary computational goal of CASA [2]. The estimation of the IBM is viewed as a binary classification of the time frequency (T-F) units [3].

So far, pitch, mel-frequency cepstral coefficients (MFCC) has been used as monaural acoustic features in T-F units' classification. Some papers attempt to utilize gammatone frequency cepstral coefficients (GFCC) in automatic speech recognition too [4], though the investigation is far from being extensive and the improvement in results being insignificant.

In this paper, we provide a new implementation of GFCC for monaural speech segregation, which has shown consistent and significant ASR performance gains in various noise types and SNR levels conditions [5]. We propose to extract GFCC within each T-F unit to significantly enlarge the GFCC implication. We also employ deep neural networks (DNNs) [6] as a classifier to estimate the ideal binary mask.

In the following section, we present an overview of the classification-based monaural speech segregation system and DNN classifiers. Section Ⅲ describes how to extract GFCC and others monaural features. We present the results of several experiments in Section Ⅳ. We discuss related issues and conclude the paper in Section Ⅴ.

II. SYSTEM OVERVIEW

The proposed system consists of four stages. As shown in Fig. 1.

The auditory filterbanks is used to decompose the input mixture signals into T-F representation units. A T-F unit corresponds to a certain channel in filter banks at a certain time frame. Then the monaural feature sets are calculated for each T-F unit. We introduce a novel GFCC auditory feature as the inputs to a binary DNN classifier for each frequency channel. The training label is provided by the IBM. We also get the binary mask form the DNN classifier as the estimated IBM in testing. The estimated binary masks are used together to compose target speech streams.

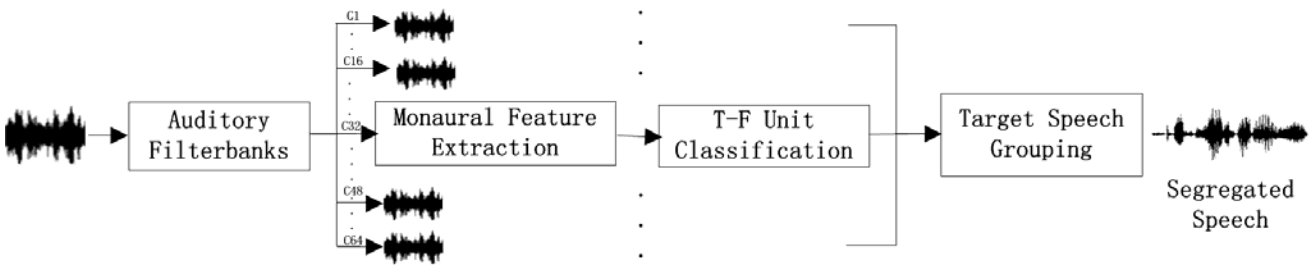

Figure 1. Schematic diagram of the monaural speech segregation

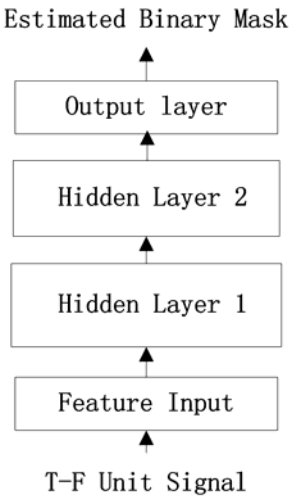

Figure 2. Structure of the DNN classfier

We follow a DNN architecture to classify the T-F units to target speech and noise as shown in Fig. 2. In this architecture, a DNN classifier consists of an input layer, two hidden layers and an output layer. The monaural features are the inputs. Considering performance and computational complexity, each hidden layer has 200 neurons. A restricted Boltzmann machine (RBM) is used to conduct initialization. The output layer labels the T-F unit to '1' for the target speech or '0' to interference source. This is a standard DNN classifier in implementation.

III. FEATURE EXTRACTION AND LEARNING

In this paper, a gammatone cochlear model proposed by Roy Patterson is used as the auditory Filterbanks [7]. The number of the filter channels is 64, with center frequencies spaced from 50 Hz to 8000 Hz. The order of the gammatone filterbanks is 4 for usual implementation.

With gammatone filterbanks, the mixtures are decomposed into different frequency channels. Then a time frame window is used to segment the signals to T-F units. The T-F units are 20-ms time frames length with 10-ms overlapping between consecutive frames.

A. Gammatone Frequency Cepstral Coefficient

The GFCC can be derived from GF-generated cochleagrams. We use 64 channels gammatone filter banks to extract the GFCC. A pre-emphasis algorithm is used in MFCC feature extraction to reduce the dynamic range of spectrum and intensifying the low frequency components. This method can retain more information of speech signals. Following the same idea, we implement a 2-order low-pass filter to pre-emphasis the input signals. We also use an average approach to down sampling the signals, which uses a window covering K points and shifting every L points to frame. In this paper, we choose K = 320, L = 160. For 16 kHz signals, these settings result in 100 frames per second, exactly the same as the down-sampling approach and the regular frame rate of MFCC. With the GF cochleagrams, the discrete cosine transform (DCT) is requested to obtain component-uncorrelated cepstral features as (1).

$$G(m,d)=(\frac{2}{C})^{0.5}\sum_{i=1}^{C}\{\frac{1}{3}\log(\bar{x}(i,m)\cos[\frac{\pi d}{2C}(2i-1)])\}\ . \quad (1)$$

There m is the index of time frames. C is the channel number of gammatone filterbanks, which is 64 in this paper. d is the frequency point index of the DCT. The energy of speech signals always distributes on lower frequency, and we just use the first 12 components to reduce the feature dimensions.

We extract the first and second order dynamic features as follows forms (2) and (3), which are generally helpful in capturing temporal information. k is 2.

$$\Delta G(m,d)=\left[\sum_{k=-2}^{2}kG(m+k,d)\right]\Big/\left[\sum_{k=-2}^{2}k^2\right]. \quad (2)$$

$$\Delta^2 G(m,d)=\left[\sum_{k=-2}^{2}k\Delta G(m+k,d)\right]\Big/\left[\sum_{k=-2}^{2}k^2\right]. \quad (3)$$

Above all, we get 36 dimension GFCC vectors to perform the speech segregations on multiple interferences and reverberant conditions.

B. Mel-Frequency Cepstral Coefficient

We follow the standard procedure to get MFCC feature. The signal is first pre-emphasized, followed by a 512-point short-time Fourier transform with a 20-ms Hamming window to get its power spectrogram. Note that we warp the magnitudes to a 64-channel mel scale, for fair comparisons with GFCC in which a 64-channel gammatone filterbanks is used for sub-band analysis. The power spectra are then warped to the mel scale followed by a operation and DCT. We use 36-D MFCC in this paper.

C. Auto Correlation Function

Pitch is a primary cue for auditory scene analysis. We just extract the 80-D normalized autocorrelation functions at each T-F units as the pitch based feature, denoted by $ACF(c,m,\tau)$. It is calculated with time lags τ by the following function as (4)

$$ACF(c,m,\tau)=\frac{\sum_{n}(x_{cm}(n)-\bar{x}_{cm})(x_{cm}(n-\tau)-\bar{x}_{cm})}{\sqrt{\sum_{n}(x_{cm}(n)-\bar{x}_{cm})^2}\sqrt{\sum_{n}(x_{cm}(n-\tau)-\bar{x}_{cm})^2}} \quad (4)$$

Where C is the channel number of gammatone filterbanks. m is the index of time frame. The n indexes a signal sample in each time frame. The time lags τ are 80 in steps of the sampling period. For 16 kHz sample rates, we get 80-D ACF feature.

IV. EVALUATION AND COMPARISON

In this section, we use the DNN based monaural speech segregation system to test the performance of these monaural features.

A. Experimental Setup

In this paper, we use the ROOMSIM package [8] to generate the noisy and reverberant mixtures. The simulated room size is 6m×4m×3m. The position of the microphone is fixed asymmetrically at 2.5m×2.5m×2m. Reflection and absorption coefficients of the wall surfaces are the same, which decided by the construction materials. The reflection paths of a particular sound source are obtained using the image reverberation model for a small rectangular room. The reverberation times (T_{60}) of the simulated room configurations are approximately 0.3s and 0.7s. We also use the anechoic measurements in this paper to generate low SNR noisy environments. The speech signals are chosen from TIMIT corpus. Babble noise materials come from NOISEX corpus. All signals, originally sampled at 16 kHz, are up sampling to 44.1 kHz to match the sample rate of the room impulse functions, then down sampling to 16 kHz to auditory periphery processing. All the DNN classifiers are trained on 0 dB SNR conditions.

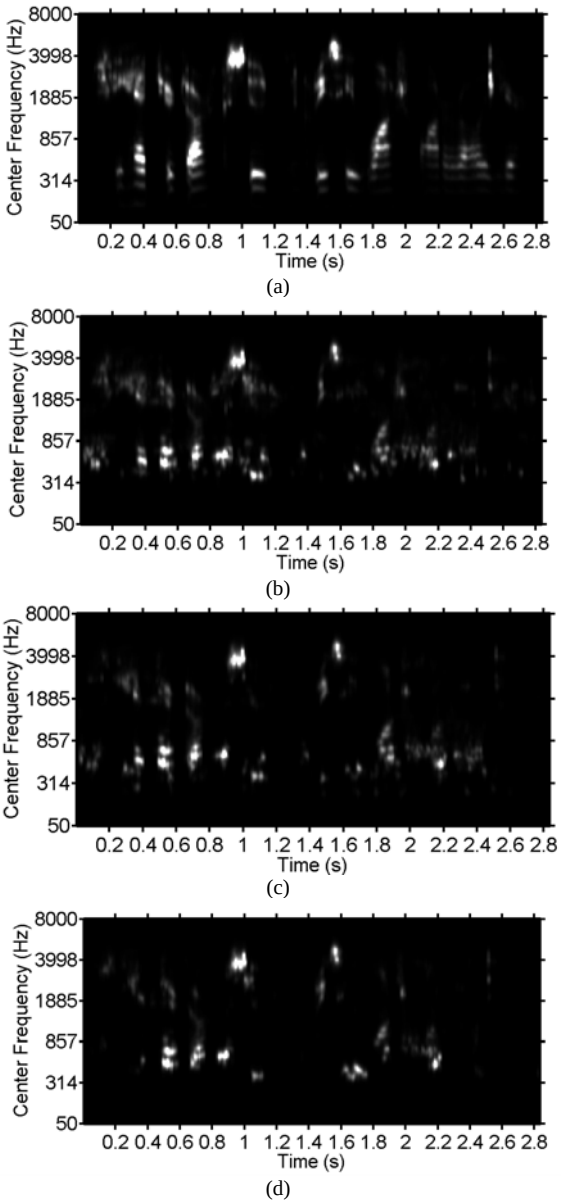

Figure 3. Cochleagrams of the target signal and segregation results

B. Evaluation Criteria

Since the task is time-frequency unit classification, we use HIT-FA as evaluation criterion for assessing classification-based speech segregation systems. The HIT-FA rate has been shown to be well correlated to intelligibility. The HIT rate is the percent of correctly classified target-dominate T-F units in IBM. The FA rate is the percent of wrong classified interference-dominate T-F units. We also use the term standard SNR algorithm to evaluate the performances of the system.

C. Evaluation in Noisy Conditions

We generate the noisy mixture speech with SNR on -5dB. The cochleagrams of the target speech and results of classification based on different features are shown on Fig.3. (a) is cochleagrams of the target clean signal. (b) is the segregation result based on ACF. (c) is the result based on MFCC. (d) is the result based on GFCC. The proposed system retains the least noise energy in the result, and we can get the cleanest speech.

TABLE I. SEGREGATION PERFORMACE IN NOISY (-5DB) CONDITIONS

Feature	HIT (%)	FA (%)	HIT-FA (%)	ACC (%)	Output SNR(dB)
ACF	90.03	53.91	36.20	50.54	-7.35
MFCC	70.53	18.48	52.05	80.41	-3.59
proposed GFCC	56.23	9.71	46.52	86.84	-1.83

As shown in Table Ⅰ. The proposed GFCC features get the best SNR performance. It also gets the highest classification accuracy among the three monaural features. The MFCC feature gets the highest HIT-FA rates. It also has the same good intelligibility as the GFCC.

D. Evaluation in Various reverberant Conditions

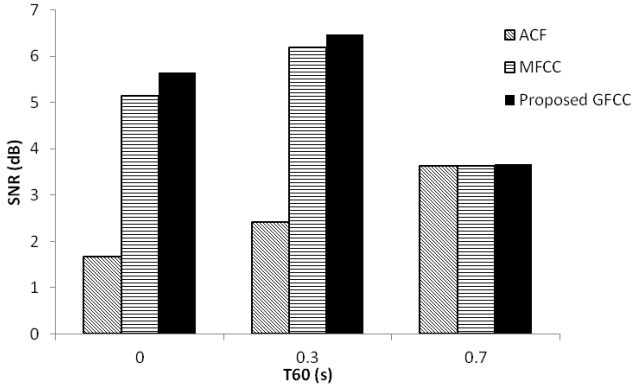

Figure 4. Segregation performance in reverberant environments

As shown in Fig 4. We test the three features' speech segregation ability in two reverberant environments. The SNR of the test mixtures are -5dB. We also use the anechoic condition as a compare system. In these conditions, the proposed GFCC get the best results in all environments. When the reverberant condition increases from 0s to 0.3s, all results are increase. When R_{60} change from 0.3s to 0.7s, only the result of ACF feature increases and results of the GFCC and MFCC are decreases.

V. CONCLUDING REMARKS

In this work, we propose a GFCC feature and test it performance in noisy and reverberant conditions with DNN classifier. The evaluation results show that the proposed GFCC achieves more robust segregation than two existing features, especially in noisy environments. We also find the proposed GFCC has good performance in even low SNR conditions. And the performance decrease gradually in noisy and reverberant conditions.

Future work should consider using monaural features and binaural features together to improve the performance of monaural speech segregation.

REFERENCES

[1] D.L. Wang, G.J. Brown, Eds., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, Wiley/IEEE Press, Hoboken, NJ, 2006.

[2] Y. Jiang, W.Q. Liang, H. Zhou, Z.M. Feng, "Performance of binary time-frequency masks in low signal to noise ratio environments," J. Tsinghua Univ., vol. 52, no. 5, pp. 636-641, 2012.

[3] K. Han, and D.L. Wang, "Towards generalizing classification based speech separation," IEEE Trans. on Audio Speech Lang. Process., vol. 21, no. 1, pp. 166-175, 2013.

[4] X.J. Zhao,Y. Shao, and D.L. WANG, "CASA-based robust speaker identification", IEEE Trans. on Audio Speech Lang. Process., vol. 20, no. 5, pp. 1608-1616, 2012.

[5] Y. Jiang, J. Qi, L. Liu, Q. Chen, and Y. Wang. "Feature Enhancement Based on CASA for Robust Speech Recognition", Proceedings of the 2012 Second International Conference on Electric Information and Control Engineering - Volume 01, IEEE Computer Society, pp. 712-715, 2012.

[6] G.E. Hinton, "Learning multiple a layers of representation," Trends Cogn Sci., vol.11,pp. 428-434, 2007.

[7] Y. Jiang, Y.Y. Zu, X. Chen, and H. Zhou," Performance evaluation of a gammatone filterbank for the embedded system," Applied Mechanics and Materials, vol.336-338, pp. 1459-1462,2013.

[8] D. R. Campbell, "The ROOMSIM User Guide (v3.3)," 2004[Online]. Available: http://media.paisley.ac.uk/~campbell/Roomsim/