A FEATURE STUDY FOR CLASSIFICATION-BASED SPEECH SEPARATION AT VERY LOW SIGNAL-TO-NOISE RATIO

Jitong Chen[1], Yuxuan Wang[1] and DeLiang Wang[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
{chenjit,wangyuxu,dwang}@cse.ohio-state.edu

ABSTRACT

Speech separation is a challenging problem at low signal-to-noise ratios (SNRs). Separation can be formulated as a classification problem. In this study, we focus on the SNR level of -5 dB in which speech is generally dominated by background noise. In such a low SNR condition, extracting robust features from a noisy mixture is crucial for successful classification. Using a common neural network classifier, we systematically compare separation performance of many monaural features. In addition, we propose a new feature called Multi-Resolution Cochleagram (MRCG), which is extracted from four cochleagrams of different resolutions to capture both local information and spectrotemporal context. Comparisons using two non-stationary noises show a range of feature robustness for speech separation with the proposed MRCG performing the best. We also find that ARMA filtering, a post-processing technique previously used for robust speech recognition, improves speech separation performance by smoothing the temporal trajectories of feature dimensions.

Index Terms— Speech separation, classification, multi-resolution cochleagram, ARMA filtering

1. INTRODUCTION

Monaural speech separation in low SNR conditions is a very challenging task in speech processing. A recent approach to speech separation applies binary masking to the spectrogram or cochleagram of the mixture, where the ideal binary mask (IBM) is considered the computational objective [1]. The IBM assigns the value 1 to a time-frequency (T-F) unit if the SNR within the unit exceeds a threshold, and 0 otherwise. Therefore, the speech separation problem can be treated as estimating the IBM. In other words, the separation problem becomes a binary classification problem [2]. Recent studies have shown that this approach is effective for improving speech intelligibility of human listeners in background noise [3] [4].

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIDCD grant (R01 DC012048), an STTR subcontract from Kuzer and the Ohio Supercomputer Center.

The performance of classification-based separation mainly depends on the choice of classifier and discriminative features extracted from the mixture. In this study, we systematically examine the robustness of many popular features for classification while fixing the classifier to a multilayer perceptron (MLP) [5]. A recent study evaluates the performance of a number of features [6], but the evaluation is done at 0 dB where human listeners perform almost perfectly (i.e. no room to improve in terms of speech intelligibility). Our study is conducted using two challenging non-stationary noises: factory noise and babble noise, mixed at -5 dB SNR where the recognition rate of even normal-hearing listeners is less than 50% [3]. In addition, we include an extensive list of features that have been shown to be effective for robust automatic speech recognition (ASR). Robust features for ASR, such as mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP), can be useful for classification-based speech separation [4]. In addition to evaluating commonly used features in robust ASR in terms of their discriminant power in IBM estimation, we propose a new feature, which we call multi-resolution cochleagram (MRCG), for the purpose of speech separation. As shown later, the MRCG feature produces the best separation performance.

The paper is organized as follows. Section 2 introduces our feature evaluation framework. Candidate features as well as the proposed MRCG feature are described in Section 3, followed by a discussion in feature post-processing in Section 4. We present experimental results in Section 5. Section 6 concludes the paper.

2. FEATURE EVALUATION FRAMEWORK

The computational goal of classification-based speech separation is to estimate the IBM given the features extracted from the mixture. We use a 32-channel gammatone filterbank to calculate the IBM with 20 ms frame length and 10 ms frame shift. The local SNR threshold that is used to label whether a T-F unit is target dominant is set to -10 dB throughout the study. The feature evaluation framework consists of feature extraction and neural network classification, as shown in Fig.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

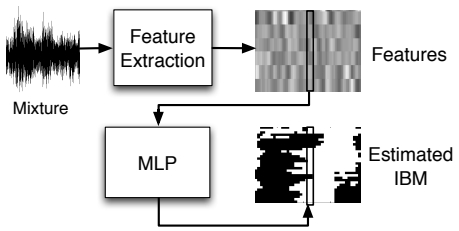

Fig. 1. Diagram of the feature evaluation framework

1. We extract frame-level acoustic features from a mixture and feed them into an MLP to estimate each frame of the IBM.

The estimated binary mask is compared against the ground truth IBM. There are several criteria to measure the similarity between the estimated mask and the IBM. One criterion is classification accuracy where percentage of correctly labeled T-F units is calculated. Classification error alone is not sufficient to measure the performance since it does not account for different error types. Therefore, we add the HIT-FA criterion where HIT is the percentage of correctly predicted target-dominant T-F units and FA is the percentage of wrongly predicted interference-dominant T-F units. HIT-FA rate is shown to be well correlated with human speech intelligibility [3].

Besides evaluating individual features, we also explore complementary features. Feature combination may yield better performance [6]. We use group lasso to select the complementary features [7]. Group lasso imposes ℓ_1/ℓ_2 mixed norm regularization on logistic regression. ℓ_1/ℓ_2 regularization is known to yield sparsity between feature groups (i.e. feature types). The input to group lasso is a high dimensional vector by concatenating all features and the target is the IBM. Regression is applied channel by channel, and the average magnitudes of regression coefficients across channels indicate the discriminative power of each feature type. The complementary features selected by the group lasso are further evaluated by the MLP.

3. FEATURE DESCRIPTION

3.1. Existing Features

First, we evaluate three popular speech recognition features including mel-frequency cepstral coefficient (MFCC), perceptual linear prediction (PLP) and relative spectral transform-PLP (RASTAPLP) [8]. In this paper, we follow standard procedures to extract 31-D MFCC, 13-D PLP and 13-D RASTAPLP. In addition, we examine two autocorrelation-related speech recognition features, namely relative autocorrelation sequence-MFCC (RAS-MFCC) [9] and phase autocorrelation-MFCC (PAC-MFCC) [10]. The idea of RAS is to reduce slow-varying components in the signal by filtering temporal autocorrelation trajectories of the signal. RAS-MFCC is computed by applying the MFCC feature extraction procedure on the filtered autocorrelation sequences. In PAC-MFCC, the angle between a signal and its shifted version is computed, then MFCC is applied to the angle sequences to derive cepstral coefficients. We use 31-D RAS-MFCC and 31-D PAC-MFCC in this paper. We also evaluate power normalized cepstral coefficients (PNCC) and Gabor filterbank (GBFB) features. PNCC employs power-law nonlinearity, asymmetric filtering and temporal masking [11]. GBFB utilizes spectrotemporal modulation frequencies to improve feature robustness [12].

Besides the speech recognition features, we also evaluate the amplitude modulation spectrogram (AMS) [3] and pitch-based features. We compute 15-D AMS following the standard procedure [3]. As for pitch-based features, we first extract pitch tracks using PEFAC [13]. Based on the pitch tracks, we compute 6-D features described in [6] for every T-F unit after passing the signal through a 64-channel gammatone filterbank. The pitch-based features therefore have 64×6 dimensions. Pitch-based features can not be used alone, as no pitch exists in unvoiced intervals. For training, we use the pitch extracted from the clean speech. During the test phase, the pitch is estimated from mixtures by PEFAC.

Gammatone domain features such as gammatone frequency cepstral coefficient (GFCC) might also provide useful information for IBM estimation. To compute GFCC, the input signal is decomposed into sub-band signals by a 64-channel gammatone filterbank. Then, each sub-band signal is down-sampled to 100 Hz, followed by point-wise loudness compression using cubic root operation. The time shift between two contiguous point of the down-sampled signal is 10 ms, which matches the frame shift used in MFCC. We further apply DCT to derive the first 31 coefficients as features.

3.2. Multi-Resolution Cochleagram

We propose a new speech separation feature called the Multi-Resolution Cochleagram (MRCG). MRCG is constructed by combining multiple cochleagram representations at different resolutions. A high-resolution cochleagram captures local information while a low-resolution cochleagram captures information in a broader spectrotemporal context. Although delta features contain some temporal context, they fail to model temporal and spectral contexts jointly. The multi-resolution contexts of a T-F unit can potentially help classify the T-F unit as target-dominant or interference-dominant. The design of MRCG features is meant to embody the spectrotemporal context systematically, in order to facilitate the estimation of the IBM. The procedure of computing MRCG is as follows:

1. Given an input mixture, compute the first 64-channel cochleagram, CG1, with the frame length of 20 ms and frame shift of 10 ms. This is a commonly used form of cochleagram. A log operation is applied to each T-F

unit.

2. Similarly, compute CG2 with the frame length of 200 ms and frame shift of 10 ms.
3. CG3 is derived by averaging CG1 across a square window of 11 frequency channels and 11 time frames centered at a given T-F unit. If the window goes beyond the given cochleagram, the outside units take the value of zero (i.e. zero padding).
4. CG4 is computed in a similar way to CG3, except that a 23×23 square window is used.
5. Concatenate CG1-4 to obtain a MRCG feature, which has 64×4 dimensions for each time frame.

4. FEATURE POST-PROCESSING

To capture the temporal trajectory of each feature dimension, delta and double-delta features can be used. The delta and double-delta features usually provide extra useful information. Previous study shows that delta and double delta features are also beneficial for speech separation [6]. We add delta and double-delta features to each feature type in this study. Mean variance normalization is usually a necessary post-processing technique, especially for neural network classifiers. We apply mean variance normalization to every feature type in this study.

Recent study shows that applying auto-regression moving-average (ARMA) filtering to mean-variance-normalized features usually leads to better performance in speech recognition [14]. ARMA filtering is defined in Equation (1) where $\hat{C}^{(\tau)}$ is the feature vector at frame τ , $\breve{C}^{(\tau)}$ is the filtered feature vector at frame τ and m is the order of the filter. The idea of ARMA filtering is to smooth the temporal trajectories of each feature dimension, making them less sensitive to noise interference. However, the effect of ARMA filtering on classification-based speech separation is unknown. We treat ARMA filtering as an optional step in feature post-processing. The results are discussed in the next section.

$$\breve{C}^{(\tau)} = \frac{\breve{C}^{(\tau-m)} + \cdots + \breve{C}^{(\tau-1)} + \hat{C}^{(\tau)} + \cdots + \hat{C}^{(\tau+m)}}{2m+1} \tag{1}$$

5. EXPERIMENTAL RESULTS

5.1. Experiment Setting

We create mixtures using the IEEE corpus recorded by a male speaker [15]. A factory noise and a babble noise from the NOISEX noise corpus are used [16]. Each mixture is obtained by mixing a sentence with one type of noise at -5dB SNR. We use 480 sentences for training and another 50 sentences for testing. The 4-minute factory and babble noise recordings are cut into two halves. We randomly select noise segments in the first half to mix with the 480 sentences to form the training set. The test set is created in the same way except that we use the second half of the noise. This guarantees that the noise segments in the test set are not seen in the training set. We train and test on the same type of noise. A one-hidden-layer MLP is used as the classifier for IBM estimation.

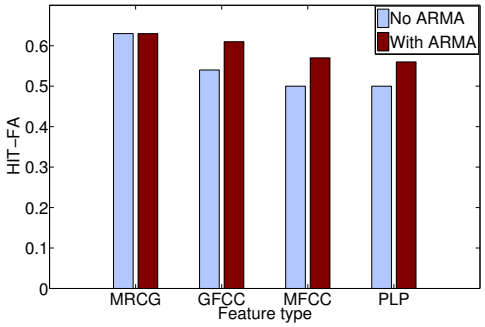

Fig. 2. Effect of ARMA filtering for the factory noise at -5dB

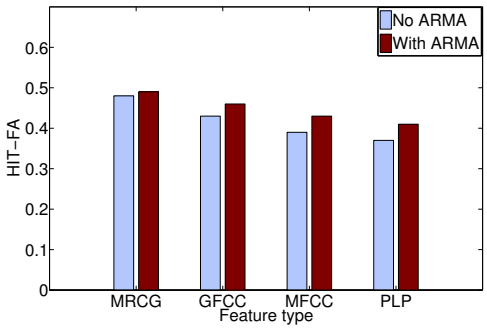

Fig. 3. Effect of ARMA filtering for the babble noise at -5dB

5.2. Result Analysis

We first examine the effect of employing ARMA filtering. Experimental results show that the second order ($m = 2$) ARMA filtering improves all evaluated features except MRCG. The effects of ARMA filtering on MRCG, GFCC, MFCC and PLP are shown in Fig. 2 and Fig. 3 for the factory noise and the babble noise, respectively. HIT-FA rates for GFCC, MFCC and PLP are clearly improved by ARMA filtering while MRCG is almost not affected. MRCG contains low-resolution cochleagram features that are derived by smoothing high-resolution cochleagram feature, which makes ARMA (only temporal smoothing) less needed. The accuracies for all features except MRCG are also improved by about 1% on average.

Although the proposed MRCG is not improved by ARMA filtering, it performs the best among all evaluated features in terms of HIT-FA and accuracy for both factory noise and babble noise. As shown in Table 1 and Table 2, the performance of MRCG is consistently better than the other features in both voiced interval and unvoiced interval. We want to

Table 1. Classification performance for the factory noise with ARMA post-processing at -5dB

Feature	Overall			Voiced			Unvoiced			Accuracy
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA	
MRCG	70%	7%	**63%**	76%	9%	**67%**	41%	5%	**36%**	**88.0%**
GFCC	67%	6%	61%	74%	8%	66%	32%	4%	28%	87.7%
RAS-MFCC	63%	6%	57%	70%	9%	61%	29%	4%	25%	87.0%
MFCC	64%	7%	57%	70%	9%	61%	31%	5%	26%	86.5%
PLP	62%	6%	56%	70%	8%	62%	28%	4%	24%	87.0%
GBFB	64%	7%	57%	72%	9%	63%	22%	5%	17%	86.3%
PNCC	62%	6%	56%	69%	9%	60%	26%	4%	22%	86.6%
RASTAPLP	58%	6%	52%	64%	8%	56%	28%	4%	24%	86.0%
AMS	43%	6%	37%	50%	8%	42%	11%	5%	6%	82.2%
PAC-MFCC	22%	5%	17%	23%	2%	21%	14%	7%	7%	77.9%
PITCH	N/A	N/A	N/A	58%	6%	52%	N/A	N/A	N/A	N/A

Table 2. Classification performance for the babble noise with ARMA post-processing at -5dB

Feature	Overall			Voiced			Unvoiced			Accuracy
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA	
MRCG	62%	13%	**49%**	66%	20%	**46%**	47%	8%	**39%**	**79.5%**
GFCC	60%	14%	46%	65%	20%	45%	41%	10%	31%	78.3%
RAS-MFCC	55%	14%	41%	59%	20%	39%	39%	9%	30%	76.9%
MFCC	57%	14%	43%	62%	21%	41%	39%	9%	30%	77.5%
PLP	53%	12%	41%	58%	18%	40%	37%	8%	29%	77.4%
GBFB	59%	18%	41%	65%	26%	39%	35%	13%	22%	74.5%
PNCC	58%	14%	44%	63%	22%	41%	40%	10%	30%	77.2%
RASTAPLP	52%	14%	38%	56%	19%	37%	38%	10%	28%	75.9%
AMS	35%	9%	26%	42%	13%	29%	7%	7%	0%	73.6%
PAC-MFCC	19%	8%	11%	20%	6%	14%	14%	9%	5%	69.8%
PITCH	N/A	N/A	N/A	58%	25%	33%	N/A	N/A	N/A	N/A

point out that GBFB is also a multi-resolution feature, but MRCG clearly performs better than GBFB on the tested noises. We also compute short-time objective intelligibility (STOI) scores by comparing the clean speech and the resynthesized speech using the estimated IBM [17]. MRCG also performs the best in terms of STOI score.

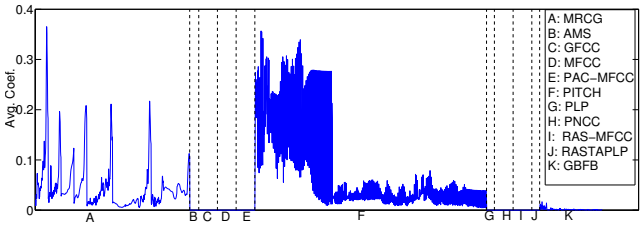

Fig. 4. Average magnitudes of regression coefficients resulted from group lasso

The features selected by group lasso for the factory noise are shown in Fig. 4. The two clusters of spikes correspond to MRCG and pitch-based features while other features have near-zero coefficients, indicating MRCG and pitch-based features are complementary. We combine these two feature types, but only obtain 53% HIT-FA for the factory noise, which is worse than MRCG alone. This is due to the fact that ground truth pitch is used for both training and group lasso while estimated pitch is used in testing. Pitch estimation at -5dB is very difficult. If we use ground truth pitch in testing, the combination of MRCG and pitch features produces 70% HIT-FA, which is substantially better than any individual feature type in this study. Similar trend is observed for the babble noise.

6. CONCLUDING REMARKS

In this study, we have systematically evaluated robust ASR features for classification-based speech separation at -5 dB SNR. We have also proposed a new feature called MRCG. Experimental results show that MRCG outperforms the existing features in terms of classification accuracy and HIT-FA. We have also found that ARMA filtering, previously used for feature post-processing in speech recognition, improves many existing features for separation.

In addition, we have explored feature combination using group lasso and found that MRCG and pitch-based features form the best combination in our feature pool. However, classification results show MRCG combined with pitch does not perform well due to poor pitch estimation in testing; pitch estimation at such a low SNR is very difficult. We expect better separation as pitch estimation improves in very noisy conditions in the future.

7. REFERENCES

[1] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, P. Divenyi, Ed., pp. 181–197. Kluwer Academic Pub., 2005.

[2] K. Han and D. L. Wang, "A classification based approach to speech segregation," *J. Acoust. Soc. Am.*, pp. 3475 –3483, 2012.

[3] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, pp. 1486–1494, 2009.

[4] E. W. Healy, S. E. Yoho, Y. Wang, and D. L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, pp. 3029–3038, 2013.

[5] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 625–638, 2009.

[6] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 270–279, 2013.

[7] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. Roy. Stat. Soc. Ser. B*, vol. 70, no. 1, pp. 53–71, 2008.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Process.*, vol. 2, no. 4, pp. 578–589, 1994.

[9] K. Yuo and H. Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences," *Speech Communication*, vol. 28, no. 1, pp. 13–24, 1999.

[10] S. Ikbal, H. Misra, and H. Bourlard, "Phase autocorrelation (PAC) derived robust speech features," in *ICASSP*, 2003, vol. 2, pp. 133–6.

[11] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *ICASSP*, 2012, pp. 4101–4104.

[12] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 131, pp. 4134–4151, 2012.

[13] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. Euro. Sig. Process. Conf.*, 451-455, p. 2011.

[14] C. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 257–270, 2007.

[15] "IEEE recommnded pratice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, no. 3, pp. 225–246, 1969.

[16] 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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.