# Exploring Monaural Features for Classification-Based Speech Segregation

Yuxuan Wang, Kun Han, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Monaural speech segregation has been a very challenging problem for decades. By casting speech segregation as a binary classification problem, recent advances have been made in computational auditory scene analysis on segregation of both voiced and unvoiced speech. So far, pitch and amplitude modulation spectrogram have been used as two main kinds of time-frequency (T-F) unit level features in classification. In this paper, we expand T-F unit features to include gammatone frequency cepstral coefficients (GFCC), mel-frequency cepstral coefficients, relative spectral transform (RASTA) and perceptual linear prediction (PLP). Comprehensive comparisons are performed in order to identify effective features for classification-based speech segregation. Our experiments in matched and unmatched test conditions show that these newly included features significantly improve speech segregation performance. Specifically, GFCC and RASTA-PLP are the best single features in matched-noise and unmatched-noise test conditions, respectively. We also find that pitch-based features are crucial for good generalization to unseen environments. To further explore complementarity in terms of discriminative power, we propose to use a group Lasso approach to select complementary features in a principled way. The final combined feature set yields promising results in both matched and unmatched test conditions.**

***Index Terms*—Binary classification, computational auditory scene analysis (CASA), feature combination, group Lasso, monaural speech segregation.**

## I. Introduction

SPEECH segregation, also known as the cocktail party problem, refers to the problem of segregating target speech from its background interference. Monaural speech segregation, which is the task of speech segregation from monaural recordings, is important for many real-world applications including robust speech and speaker recognition, audio information retrieval and hearing aids design (see e.g., [1], [7]). However, despite decades of effort, monaural speech segregation still remains one of the hardest problems in signal and speech processing. In this paper, we are concerned with monaural speech segregation from nonspeech interference; in other words, we do not address multitalker separation.

Numerous algorithms have been developed to attack the monaural speech segregation problem. For example, spectral subtraction [4] and Weiner filtering [6] are two representative techniques. However, assumptions regarding background interference are needed to make them work reasonably well. Another line of research relies on source models, e.g., training models for different speakers. Algorithms such as [19], [27], [28] can work well if the statistical properties of the observations correspond well to training conditions. Generalization to different sources usually needs model adaptation, which is a non-trivial issue.

Computational auditory scene analysis (CASA), which is inspired by Bregman's account of auditory scene analysis (ASA) [2], has shown considerable promise in the last decade. The estimation of the ideal binary mask (IBM) is suggested as a primary goal of CASA [35]. The 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). In this work, we use a 0 dB threshold in all the experiments. A series of recent experiments [5], [24], [37] shows that IBM processing of sound mixtures yields large speech intelligibility gains.

The estimation of the IBM may be viewed as binary classification of T-F units. Recent studies have applied this formulation and achieved good speech segregation results in both anechoic and reverberant environments [11], [14], [20], [22], [23], [29], [39]. In [14], [20], the pitch-based features are used in training a classifier to separate target and interference dominant units. However, the pitch-based features cannot deal with unvoiced speech that lacks harmonic structure. Seltzer *et al.* [29] and Weiss *et al.* [39] use comb filter and spectrogram statistics as features. In [11], [22], [23], amplitude modulation spectrogram (AMS) is used, which makes unvoiced speech segregation possible as AMS is a characteristic of both voiced and unvoiced speech. Unfortunately, the generalization ability of AMS is not good [11].

For classification, the use of an appropriate classifier is obviously important. Our previous study [11] focuses on classifier comparisons, and suggests that support vector machines (SVMs) work better than Gaussian mixture models (GMMs). However, this study only uses two existing features. Equally important for classification is the choice of appropriate features, which are less studied. It should be noted that we are concerned with T-F unit level features, i.e., spectral/cepstral features extracted from each T-F unit. Feature extraction is possible be-

Manuscript received February 16, 2012; revised June 05, 2012; accepted September 20, 2012. Date of publication October 02, 2012; date of current version November 21, 2012. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-08-1-0155 and in part by an STTR grant from the AFOSR. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bryan Pardo.

Y. Wang and K. Han are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: wangyuxu@cse.ohio-state.edu; hank@cse.ohio-state.edu).

D. Wang is with the Department of Computer Science and Engineering and the Center for Cognitive Science, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASL.2012.2221459

cause a T-F unit is a signal of a certain length. To our knowledge, aside from the features used in [29], only pitch and AMS have been used as T-F unit level features. On the other hand, in the speech and speaker recognition community, many acoustic features have been explored, such as gammatone frequency cepstral coefficients (GFCC), mel-frequency cepstral coefficients (MFCC), relative spectral transform (RASTA) and perceptual linear prediction (PLP), each having its own advantages. However, they have not been studied as T-F unit level features for classification-based speech segregation.

The objective of this paper is to conduct a comprehensive feature study for classification-based speech segregation. That said, we fix SVM as the classifier and explore the use of existing speech and speaker features under the same classification framework. Our contributions are as follows:

- We propose to extract conventional speech/speaker features within each T-F unit to significantly enlarge the feature repository for unit classification.
- We propose a principled method to identify a complementary feature set. It is shown in speech recognition that complementarity exists between basic acoustic features [9], [42]. To investigate complementary features in terms of discriminative power, we address the corresponding group variable selection problem using a group least absolute shrinkage and selection operator (Lasso) [41].
- We systematically compare the segregation performance of the newly included features and combinations in various acoustic environments.

This paper is organized as follows. We present an overview of the system along with the methodology of extracting features at the T-F unit level in Section II. Section III describes a group Lasso approach to combining different features. Unit labeling results are reported in Section IV. We conclude this paper in Section V.

## II. System Overview and Feature Extraction

We describe the architecture of our segregation system as follows. A sound mixture with the 16 kHz sampling frequency is first fed into a 64-channel gammatone filterbank, with center frequencies equally spaced from 50 Hz to 8000 Hz on the equivalent rectangular bandwidth rate scale. Gammatone filters model human auditory filters (critical bands) [26], and 64 channels provide an adequate frequency representation (see e.g., [37]). The output in each channel is then divided into 20-ms frames with 10-ms overlapping between consecutive frames. This procedure produces a time-frequency representation of the sound mixture, called a cochleagram [36]. Our computational goal is to estimate the ideal binary mask for the mixture. Since the energy distribution of speech signals in different channels can be very different, we train a Gaussian-kernel SVM [11] for each subband channel separately, and ground truth labels are provided by the IBM. We use 5-fold cross validation to determine the hyperparameters. Feature extraction is performed at the T-F unit level in the way described below. After obtaining a binary mask, i.e., estimated IBM, from trained SVM classifiers, the target speech is segregated from the sound mixture in a resynthesis step [36]. Note that we do not perform auditory segmentation, which is usually done for better segregation [11], [20], as we want to directly compare the unit labeling performance of each feature type. Auditory segmentation refers to a stage of processing that breaks the auditory scene into contiguous T-F regions each of which contains acoustic energy mainly from a single sound source.

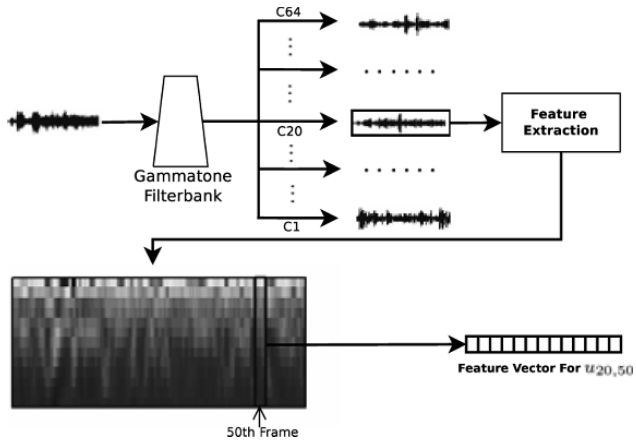



Fig. 1. Illustration of deriving RASTA-PLP features for the T-F unit in channel 20 and at frame 50 ($u_{20,50}$).

Acoustic features are usually derived at the frame level. But since a binary decision needs to be made for each T-F unit, we need to find an appropriate representation for each T-F unit (recall that each T-F unit contains a slice of a subband signal). This can be done in a straightforward way as follows. To get acoustic features for the T-F unit $u_{c,m}$ in channel $c$ and at frame $m$, we take the filtered output $x_c(t)$ in channel $c$. Treating $x_c(t)$ as the input, conventional frame-level acoustic feature extraction is carried out and the feature vector at frame $m$ is taken as the feature representation for $u_{c,m}$. The unit level features derived this way obviously contain redudancy, as the subband signals are limited to the bandwidth of the corresponding gammatone filters. Nevertheless, such redundancy does no harm to classification in our experiments. We also proposed a method to reduce the dimensionality for unit level features, which derives different acoustic features based on bandlimited spectral features. Interested readers are referred to our technical report [38]. Fig. 1 illustrates how to derive a 12th order RASTA-PLP feature vector (including zeroth cepstral coefficient) for the T-F unit in channel 20 and at frame 50.

In the following, we describe the features used in our experiments. These features have been successfully used in many speech processing tasks. We use the RASTAMAT toolbox [8] for extracting MFCC, PLP, and RASTA-PLP features.

### A. Amplitude Modulation Spectrogram

AMS features have been applied to speech segregation problems recently [23]. To extract AMS features, we extract the envelope of the mixture signal by full-wave rectification and decimate it by a factor of 4. The decimated envelope is Hanning windowed and zero-padded for a 256-point FFT. The resulted FFT magnitudes are integrated by 15 triangular windows uniformly spaced from 15.6 to 400 Hz, producing a 15-D AMS feature vector.

### B. Perceptual Linear Prediction

PLP [12] is a popular representation in speech recognition, and it is designed to find smooth spectra consisting of resonant

peaks. To derive PLPs, we first warp the power spectrum to a 20-channel Bark scale using trapezoidal filters. Then, equal loudness preemphasis is applied, followed by applying an intensity loudness law. Finally, cepstral coefficients from linear predictions form the PLP features. Following common practice in speech recognition, we use a 12th order linear prediction model, yielding 13-D (including zeroth cepstral coefficient) PLP features.

### C. Relative Spectral Transform-PLP

RASTA filtering [13] is often coupled with PLP for robust speech recognition. In our experiments, we use a log-RASTA filtering approach. After the power spectrum is warped to the Bark scale, we $\log$-compress the resulted auditory spectrum, filter it by the RASTA filter (single pole at 0.94), and expand it again by an exponential function. Subsequently, PLP analysis is taken on this filtered spectrum. In essence, RASTA filtering serves as a modulation-frequency bandpass filter, which emphasizes the modulation frequency range most relevant to speech while discarding lower or higher modulation frequencies. Same as PLP, we use 13-D RASTA-PLP in this paper.

### D. Gammatone Frequency Cepstral Coefficient

To get GFCC features [31], a signal is decomposed by a 64-channel gammatone filterbank first. Then, we decimate a filter response to an effective sampling rate of 100 Hz, resulting in a 10-ms frame shift. The magnitudes of the decimated filter outputs are then loudness-compressed by a cubic root operation. Finally, discrete cosine transform (DCT) is applied to the compressed signal to yield GFCC. As suggested in [30], we use 31-D GFCC in this paper.

### E. Mel-Frequency Cepstral Coefficient

We follow the standard procedure to get MFCC. The signal is first preemphasized, followed by a 512-point short-time Fourier transform with a 20-ms Hamming window to get its power spectrogram. The power spectra are then warped to the mel scale followed by a $\log$ operation and DCT. Note that we warp the magnitudes to a 64-channel mel scale, for fair comparisons with GFCCs in which a 64-channel gammatone filterbank is used for subband analysis. We use 31-D MFCC in this paper.

### F. Pitch-Based Features

Pitch is a primary cue for ASA. In our experiments, we use a set of pitch-based features originally proposed in [14], and its effectiveness has been confirmed in both anechoic and reverberant environments with additive noise [17], [20]. Although we are only concerned with nonspeech interference in this paper, it should be noted that pitch can also be effective for segregating target speech from competing speech. To get pitch-based features for the T-F unit $u_{c,m}$, we first calculate the normalized autocorrelation function at each time lag $\tau$, denoted by $A(c,m,\tau)$:

$$A(c,m,\tau) = \frac{\sum_n x_c(mT_m - nT_n)x_c(mT_m - nT_n - \tau T_n)}{\sqrt{\sum_n x_c^2(mT_m - nT_n)}\sqrt{\sum_n x_c^2(mT_m - nT_n - \tau T_n)}} \quad (1)$$

where $T_m = 10$ ms is the frame shift and $T_n$ is the sampling period. The summation is over a 20-ms frame. If the signal in $u_{c,m}$ is voiced and dominated by the target speech, it should have a period close to the pitch period at frame $m$. That is, given the pitch period of the target speech $\tau_m$ at frame $m$, $A(c,m,\tau_m)$ measures how well the signal in $u_{c,m}$ is consistent with the target speech.

The second and third features involve the average instantaneous frequency $\bar{f}(c,m)$ derived from the zero-crossing rate of $A(c,m,\tau)$. If the signal in $u_{c,m}$ belongs to target speech, the product of $\bar{f}(c,m)$ and $\tau_m$ gives a harmonic number. Hence, we set the second feature to be the nearest integer of $\bar{f}(c,m)\tau_m$ and the third feature to be the difference between the actual value of the product and its nearest integer. These two features have complementary information to the first feature $A(c,m,\tau_m)$ [17].

The next three features are the same as the first three except that they are extracted from the envelopes of filter responses. The envelopes are calculated by using a low-pass FIR filter with passband $[0, 1\ \text{kHz}]$ and a Kaiser window of 18.25 ms. The resulting 6-D feature vector is:

$$\mathbf{x}_{c,m} = \begin{pmatrix} A(c,m,\tau_m) \\ [\bar{f}(c,m)\tau_m] \\ |\bar{f}(c,m)\tau_m - [\bar{f}(c,m)\tau_m]| \\ A_E(c,m,\tau_m) \\ [\bar{f}_E(c,m)\tau_m] \\ |\bar{f}_E(c,m)\tau_m - [\bar{f}_E(c,m)\tau_m]| \end{pmatrix} \quad (2)$$

where $[\cdot]$ denotes the round operation, and subscript $E$ indicates envelope. It should be noted that pitch exists only in voiced speech. In this study, classifiers are trained on ground truth pitch extracted from clean speech by PRAAT [3], but tested on pitch estimated by a recently proposed multipitch tracker [21].

## III. Feature Combination: A Group Lasso Approach

Different acoustic features characterize different properties of the speech signal. As observed in speech recognition, feature combination may lead to significant performance improvement [9], [42]. Here, feature combination is usually done in three ways. The simplest method is to directly try different combinations. The exponential number of possibilities renders this method unrealistic when the number of features is large. The second way is to perform unsupervised feature transformation such as kernel-PCA [32] on the concatenated feature vector. The third way is to apply supervised feature transformation such as linear discriminant analysis (LDA) [9] to the concatenated feature vector. However, an issue with feature transformation relates to complementarity; i.e., it is unclear which feature types are complementary after transformation. Here, by complementarity, we mean that each feature type provides complementary information to boost classification and thus their combination (concatenation in paper) should outperform an individual type.

Therefore, our goal is to find a principled way to select a set of complementary features, and such complementarity should be related to the discrimination of target-dominance and interference-dominance. This problem can be cast as a group variable selection problem, which is to find important groups of explanatory factors for prediction in the regression framework.

Group Lasso [41], a generalization of the widely used Lasso operator [34], is designed to tackle this problem by incorporating a mixed-norm regularization over regression coefficients. Since our labels are binary, we use the logistic regression extension of group Lasso [25], which can be efficiently solved by block coordinate gradient descent. The estimator is

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta},a} \sum_i \log\left(1+\exp(-y_i(\boldsymbol{\beta}^T\mathbf{x}_i+a))\right)+\lambda\sum_{g=1}^{G}\|\boldsymbol{\beta}_{\mathcal{I}_g}\|_2 \tag{3}$$

where $\mathbf{x}_i$ is the $i$ th training sample, $y_i$ is the ground truth label scaled to $\{-1,1\}$, and $a$ is the intercept. $\|\cdot\|_2$ refers to the $\ell_2$ norm. $\boldsymbol{\beta}$ consists of $G$ predefined non-overlapping groups and $\mathcal{I}_g$ is the index set of the $g$ th group. The first term in the minimization is a standard log loss that concerns discrimination. The second term is an $\ell_1/\ell_2$ mixed-norm regularization, which imposes an $\ell_1$ regularization between groups and an $\ell_2$ regularization within each group. It is well known that the $\ell_1$ norm induces sparsity, therefore the $\ell_1/\ell_2$ regularization results in group sparsity hence group level feature selection. Regularization parameter $\lambda$ controls the level of sparsity of the resulting model. In practice, we usually calculate $\lambda_{\max}$ first, above which $\hat{\boldsymbol{\beta}}_\lambda$ is very close to zero. We then use $\gamma\cdot\lambda_{\max}$ with $\gamma\in[0,1]$ as $\lambda$ in (3) for the ease of choosing appropriate parameter values.

To do feature combination, all the features are concatenated together to form a long feature vector, and each feature type is defined as a group; e.g., AMS (all 15 feature elements) is defined as the first group, PLP as the second, and so on. Then, for a fixed $\gamma$ (hence $\lambda$), we solve (3) to get $\hat{\boldsymbol{\beta}}_\lambda$. Since group sparsity is induced, $\hat{\boldsymbol{\beta}}_{\mathcal{I}_g}$ shall be zeros (or small numbers) for some groups $g$, meaning that these groups (feature types) contribute little to discrimination in the presence of the other groups. Groups shall be selected if the magnitudes of their regression coefficients are greater than zero. Since (3) is solved at each channel separately, different types of features may get selected for different channels. A subband SVM classifier is then trained on the selected features and a cross-validation accuracy is obtained. To select a "global" set of complementary features, we average the cross-validation accuracies and corresponding regression coefficients across frequency channels. Features having significant average responses or peaks are considered to be complementary for the particular choice of $\gamma$. This is done for $\gamma$ varying from 0 to 1 with the step size of 0.05. To achieve a good trade-off between discrimination power and model complexity which is the number of groups selected, we empirically determine the final combination by leveraging the averaged cross-validation accuracies with the corresponding model complexity.

## IV. Evaluation Results

### A. Experimental Setup

We use the IEEE corpus [18] for most of our evaluations. All utterances are downsampled to 16 kHz. For training, we mix 50 utterances recorded by a female talker with three types of noise at 0 dB. The three noises are: N1—bird chirps, N2—crow noise, and N3—cocktail party noise [14]. We choose 20 new utterances from the IEEE corpus for testing. The test utterances are different from those in training. Unless stated otherwise, test utterances from the same female talker are used, i.e., a speaker-dependent setting. This enables us to directly compare with [23] where the same speaker is used in training and testing. Relaxing speaker dependency is examined in Section IV-I. Two test conditions are employed. In the matched-noise condition, we mix the test utterances with different cuts from the trained noises (i.e., N1-N3) in order to test the performance on unseen utterances. In the unmatched-noise condition, the test utterances are mixed with three unseen noises: N4—crowd noise at a playground, N5—electric fan noise, and N6—traffic noise. The test mixtures are all mixed at 0 dB except in Section IV-H. There are approximately 800 seconds of mixtures for training in most of the experiments. The experiments in Section IV-G use longer training data as the number of training utterances is increased. For testing, there are approximately 650 seconds of mixtures for the IEEE test set and 700 seconds for the TIMIT test set (see Section IV-I). The number of T-F units to be classified is about $30{,}000\times 64 = 1{,}920{,}000$ for the IEEE test set and $35{,}000\times 64 = 2{,}240{,}000$ for the TIMIT test set.

The dimensionality of each feature is described in Section II. As mentioned before, for the pitch-based features, ground truth pitch and estimated pitch are used in training and testing, respectively. We use PITCH to denote the 6-D pitch-based features.

To put the performance of our classification-based segregation in perspective, we include results from a recent CASA system, the tandem algorithm [17], which jointly performs voiced speech segregation and pitch estimation in an iterative fashion. The tandem algorithm is initialized by the same estimated pitch from [21]. We use ideal sequential grouping for the tandem algorithm, because the algorithm does not deal with the issue of sequential grouping, i.e., it does not have a way to group pitch contours (and their associated masks) of the same speaker across time to form a segregated sentence. So these results represent the ceiling performance of the tandem algorithm.

Aside from the tandem algorithm which tries to estimate the IBM explicitly, we focus on comparisons between different features under the same framework. Comparisons with fundamentally different techniques are not included in this study which is about feature exploration for classification-based speech separation.

### B. Evaluation Criteria

Since the task is classification, it is straightforward to measure the performance using classification accuracy. However, simply using accuracy as the evaluation criterion may not be appropriate, as miss and false-alarm errors are treated equally. Speech intelligibility studies [23], [24] have shown that false-alarm (FA) errors are far more detrimental to human speech intelligibility than miss errors. Kim *et al.* have thus proposed the HIT-FA rate as an evaluation criterion, and shown that this rate is well correlated to intelligibility [24]. The HIT rate is the percent of correctly classified target dominant T-F units in the IBM. The FA rate is the percent of wrongly classified interference-dominant T-F units in the IBM. Therefore, we use HIT-FA as our main evaluation criterion. Another criterion is the IBM-modulated SNR of the segregated speech. When computing SNRs, the target speech resynthesized from the IBM is

TABLE I
SEGREGATION PERFORMANCE FOR SINGLE FEATURES IN THE MATCHED-NOISE CONDITION. BOLDFACE INDICATES BEST RESULT. "†" INDICATES THE RESULT IS SIGNIFICANTLY BETTER THAN AMS AT A 5% SIGNIFICANCE LEVEL

| Feature | Overall | | | Voiced | | | Unvoiced | | | Accuracy | SNR (dB) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | HIT | FA | HIT−FA | HIT | FA | HIT−FA | HIT | FA | HIT−FA | | |
| AMS | 70% | 6% | 64% | 76% | 8% | 68% | 49% | 4% | 45% | 84.6% | 13.7 |
| PLP | 79% | 9% | 70%† | 83% | 10% | 73%† | 65% | 6% | 59%† | 86.5%† | 14.9† |
| RASTA-PLP | 74% | 7% | 67%† | 79% | 9% | 70%† | 56% | 4% | 52%† | 85.9%† | 14.2† |
| GFCC | 87% | 8% | **79%**† | 89% | 9% | **80%**† | 77% | 6% | **71%**† | **90.1%**† | **17.3**† |
| MFCC | 82% | 7% | 75%† | 86% | 8% | 78%† | 69% | 5% | 64%† | 88.8%† | **17.3**† |
| PITCH | N/A | N/A | N/A | 77% | 16% | 61% | N/A | N/A | N/A | N/A | N/A |
| TANDEM [17] | N/A | N/A | N/A | 75% | 4% | 71%† | N/A | N/A | N/A | N/A | N/A |

TABLE II
SEGREGATION PERFORMANCE FOR SINGLE FEATURES IN THE UNMATCHED-NOISE CONDITION

| Feature | Overall | | | Voiced | | | Unvoiced | | | Accuracy | SNR (dB) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | HIT | FA | HIT−FA | HIT | FA | HIT−FA | HIT | FA | HIT−FA | | |
| AMS | 60% | 23% | 37% | 64% | 22% | 42% | 44% | 25% | 19% | 72.7% | 7.6 |
| PLP | 71% | 30% | 41%† | 73% | 30% | 43% | 65% | 32% | 33%† | 70.0% | 7.4 |
| RASTA-PLP | 69% | 12% | **57%**† | 71% | 13% | 58%† | 60% | 9% | **51%**† | **83.8%**† | **9.1**† |
| GFCC | 77% | 33% | 44%† | 76% | 32% | 44%† | 77% | 34% | 43%† | 69.4% | 6.6 |
| MFCC | 74% | 29% | 45%† | 75% | 29% | 46%† | 70% | 29% | 41%† | 71.7% | 8.1† |
| PITCH | N/A | N/A | N/A | 76% | 20% | 56%† | N/A | N/A | N/A | N/A | N/A |
| TANDEM [17] | N/A | N/A | N/A | 68% | 4% | **64%**† | N/A | N/A | N/A | N/A | N/A |

used as the ground truth signal [15], [17], as the IBM represents the ground truth of classification. This IBM-modulated SNR complements the above classification-based criteria by taking into account the underlying signal energy of each T-F unit.

We should note that other evaluation criteria have been developed in the speech separation community, including SNR and source to distortion ratio (SDR). Unlike the IBM which is directly motivated by the auditory masking phenomenon, SNR and SDR do not take into consideration perceptual effects. Also, it is well known that SNR may not correlate to speech intelligibility and the relationship between SDR and speech intelligibility is still unknown. Because of its correlation with speech intelligibility, we prefer the HIT-FA rate over SNR and SDR.

### C. Single Features

In terms of HIT-FA, we document unit labeling performance at three levels: voiced speech intervals (pitched frames), unvoiced speech intervals (unpitched frames), and overall. Voiced/unvoiced speech intervals are determined by ground truth pitch. Both classification accuracy and SNR are evaluated at the overall level. Table I gives the results in the matched-noise test condition. In this condition, all features are able to maintain a low FA rate. The performance differences mainly stem from the HIT rate. Clearly, AMS does not perform well compared with the other features as it fails to label a lot of target-dominant units. In contrast, GFCC manages to achieve high HIT rates, with 79% overall HIT-FA, which is significantly better than other single features. The classification accuracy and SNR using GFCC are also significantly higher than those obtained by the other features (except MFCC in terms of SNR). Unvoiced speech is important to speech intelligibility, and its segregation is a difficult task due to the lack of harmonicity and weak energy [16]. Again, AMS performs the worst whereas GFCC does a very good job at segregating unvoiced speech. The good performance of GFCC is probably due to its effectiveness as a speaker identification feature [31]. An encouraging observation in the matched-noise condition is that some general acoustic features such as GFCC and MFCC significantly outperform PITCH even in voiced intervals. This remains true even when ground truth pitch is used in (2), which achieves 72% HIT-FA in voiced intervals. Similarly, the tandem algorithm, which includes auditory segmentation, is not competitive. For systematic comparison, we have produced the receiver operating characteristic (ROC) curves for overall classification obtained by using single features, and interested readers are referred to our technical report [38].

Unlike the matched-noise condition, the unseen broadband noises are more demanding for generalization. The segregation results in the unmatched-noise condition are listed in Table II. We can see that the classification accuracy and both HIT rate and FA rate are affected, and the main degradation comes from substantially increased FA rates. Contrary to the other features, PITCH is the least affected feature type with only 5% reduction in HIT-FA. Using ground truth pitch it is able to achieve 68% HIT-FA in voiced intervals. As the pitch-based features reflect intrinsic properties of speech, we do not expect that the change of interference will dramatically change pitch characteristics in target-dominant T-F units. Similarly, the tandem algorithm obtains a fairly low FA rate and achieves the best HIT-FA result in voiced intervals in this condition. Among others, it is interesting to see that RASTA-PLP becomes the best performing feature type in terms of all three criteria. As shown in [13], RASTA-PLP effectively acts as a modulation-frequency filter, which retains slow modulations corresponding to speech.

We have used Student's $t$-tests at a 5% significance level to examine if an improvement is statistically significant. We use the symbol "†" to denote that a result is significantly better than the previously studied AMS feature. As can be seen in Tables I and II, almost all the improvements are statistically significant.

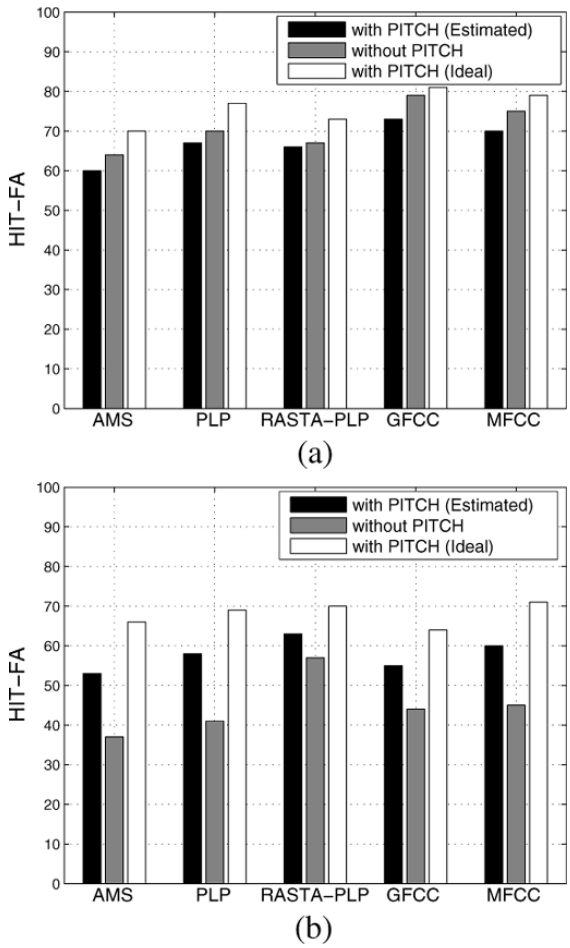



Fig. 2. Overall HIT-FA performance for pairwise combination of single features and pitch-based features in (a) the matched-noise condition, and (b) the unmatched-noise condition. (a) Matched-noise condition. (b) Unmatched-noise condition.

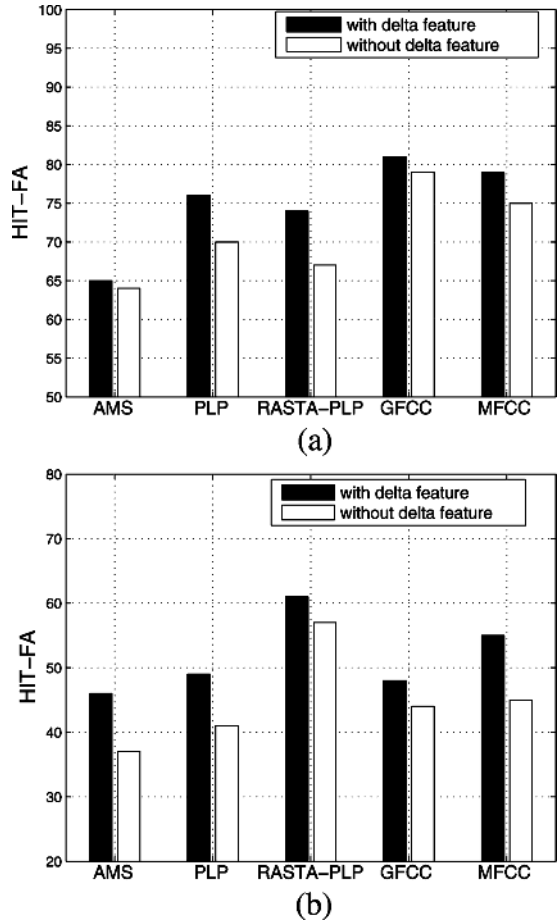



Fig. 3. Effects of delta features on overall HIT-FA performance in (a) the matched-noise condition, and (b) the unmatched-noise condition. (a) Matched-noise condition. (b) Unmatched-noise condition.

### *D. Combining With Pitch-Based Features*

Considering the excellent performance of some features in the matched-noise condition and the robustness of the pitch-based features in the unmatched-noise condition, it seems sensible to combine the single features with the pitch-based features. If the pitch tracker dose not detect pitch in a frame, we simply set pitch-based features to all zeros in the combination. Fig. 2(a) shows the overall HIT-FA results for pairwise combinations in the matched-noise condition. Due to pitch estimation errors, the combination does not improve the performance in this test condition. However, it can be seen that the combination using the ideal (ground-truth) pitch significantly improves the performance for all the features. Results for the unmatched-noise condition are listed in Fig. 2(b). Even with estimated pitch, the performance of all the features is significantly boosted by the combination, demonstrating the role of the pitch-based features in generalization to unseen noises. As before, RASTA-PLP leads the overall performance in this combination. We note here that all the improvements are statistically significant.

### *E. Adding Delta Features*

Difference features, also known as delta features, are found to be useful in speech processing as they capture variations. We now investigate the effects of including delta features. A positive effect of adding delta features with AMS has been shown in [23]. Fig. 3 shows the overall HIT-FA results by adding first-order delta features (denoted by $\Delta$) along time in matched and unmatched-noise conditions. We can clearly see improvements in both test conditions. Two observations are in order. First, adding deltas is helpful for unvoiced speech segregation (not shown). Second, all features benefit from adding deltas in the unmatched-noise condition, indicating their effect in improving generalization. We note here that all the improvements are statistically significant.

We have also experimented with adding additional deltas along frequency channel as suggested in [23]. This also yields some improvements yet at the expense of added dimensionality. As a trade-off, in the next few experiments, we add deltas along frequency only for PITCH which has a low dimensionality, producing a 18-D feature representation denoted by $\mathrm{PITCH}\Delta\Delta$.

### *F. Feature Combination*

In this subsection, we evaluate feature combination as described in Section III. Since we want the selected features to be general, the mixtures from both IEEE female and male talkers are used to form the training data for the group Lasso. As outlined in Section III, we concatenate AMS, PLP, RASTA-PLP, MFCC, GFCC, PITCH and their deltas together and define each feature type as a group. Group Lasso feature selection is then performed

TABLE III
SEGREGATION PERFORMANCE FOR FEATURE COMBINATION IN THE MATCHED-NOISE CONDITION. "‡" INDICATES THAT THE RESULT IS SIGNIFICANTLY BETTER THAN ALL THE OTHER FEATURES AT A 5% SIGNIFICANCE LEVEL

| Feature Combination | Overall | | | Voiced | | | Unvoiced | | | Accuracy | SNR (dB) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | HIT | FA | HIT−FA | HIT | FA | HIT−FA | HIT | FA | HIT−FA | | |
| AMS+RASTA-PLP +MFCC | 86% | 5% | **81%**‡ | 89% | 6% | **83%** | 75% | 3% | **72%**‡ | **91.8%**‡ | **19.6**‡ |
| AMS+RASTA-PLP | 80% | 6% | 74% | 84% | 7% | 77% | 64% | 5% | 59% | 88.6% | 16.9 |
| AMS +MFCC | 83% | 6% | 77% | 87% | 7% | 80% | 70% | 4% | 66% | 90.1% | 18.5 |
| RASTA-PLP +MFCC | 84% | 6% | 78% | 87% | 7% | 80% | 74% | 4% | 70% | 90.5% | 17.8 |
| AMS+RASTA-PLPΔ +MFCC+PITCHΔΔ | 84% | 7% | 77% | 89% | 6% | **83%** | 69% | 8% | 61% | 89.7% | 15.6 |
| LDA | 80% | 10% | 70% | 85% | 9% | 76% | 63% | 13% | 50% | 85.9% | 12.6 |

TABLE IV
SEGREGATION PERFORMANCE FOR FEATURE COMBINATION IN THE UNMATCHED-NOISE CONDITION

| Feature Combination | Overall | | | Voiced | | | Unvoiced | | | Accuracy | SNR (dB) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | HIT | FA | HIT−FA | HIT | FA | HIT−FA | HIT | FA | HIT−FA | | |
| AMS+RASTA-PLP +MFCC | 80% | 20% | 60% | 80% | 21% | 59% | 80% | 20% | 60% | 80.0% | 9.0 |
| AMS+RASTA-PLP | 72% | 15% | 57% | 74% | 15% | 59% | 65% | 14% | 51% | 81.3% | 8.7 |
| AMS +MFCC | 75% | 25% | 50% | 76% | 25% | 51% | 67% | 24% | 43% | 74.9% | 8.1 |
| RASTA-PLP +MFCC | 77% | 17% | 60% | 76% | 17% | 59% | 76% | 15% | 61% | 81.9% | 9.0 |
| AMS+RASTA-PLPΔ +MFCC+PITCHΔΔ | 81% | 11% | **70%**‡ | 83% | 12% | **71%**‡ | 72% | 10% | **62%**‡ | **87.0%**‡ | **10.5**‡ |
| LDA | 70% | 11% | 59% | 73% | 13% | 60% | 55% | 8% | 47% | 84.6% | 8.2 |

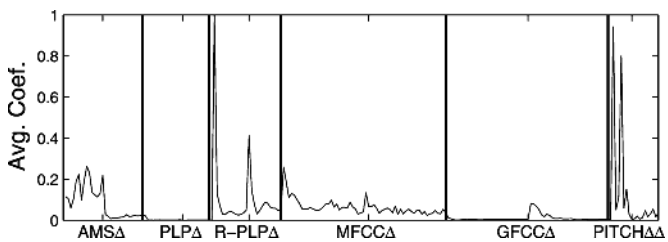



Fig. 4. Averages of the magnitudes of regression coefficients across channels, where R-PLP stands for RASTA-PLP.

on the normalized concatenated feature vector. We empirically found that $\gamma = 0.1$ offers a good trade-off between model complexity and cross-validation accuracy. We plot the averages of the magnitudes of regression coefficients across channels in Fig. 4. It is clear that AMS, RASTA-PLP, MFCC and PITCH are associated with larger regression coefficients, while the coefficients of PLP are zero in almost all channels. GFCC's contribution to model fitting is relatively weak (i.e., its regression coefficients are relatively small), making it almost redundant given AMS, RASTA-PLP, MFCC and PITCH. We set the *final combined feature set* to AMS+RASTA-PLPΔ+MFCC+PITCHΔΔ, resulting in a 90-D feature vector. We do not include deltas for AMS and MFCC because we found that they improve performance only slightly at the expense of nearly doubling the dimensionality. Since we have already validated the effectiveness of PITCH, we will also present comparisons with AMS+RASTA-PLP+MFCC, which comes from the feature selection and is referred as the *complementary feature set* in the rest of the paper.

The segregation results of feature combination in the matched and unmatched-noise conditions are shown in Tables III and IV. To show that the feature combination is not redundant, we also include results from AMS+RASTA-PLP, AMS+MFCC, and RASTA-PLP+MFCC. As a comparison, we also present results using LDA for feature combination. LDA is applied to the same concatenated feature vector on which group Lasso is applied. We use the symbol "‡" to denote that a result is significantly better than all the other features. We can see that the complementary feature set AMS+RASTA-PLP+MFCC performs the best (equaling GFCCΔ, see Fig. 3(a)) in the matched test condition, and is significantly better than all the other single features in the unmatched test condition (see Table II). The final combined feature set generalizes well to unseen noises as shown in Table IV. For reference, the final combined feature set using ground truth pitch achieves 84% and 76% HIT-FA rates in the two test conditions, respectively. LDA does not achieve comparable results in either test condition.

### G. Training Corpus Size

As mentioned in Section IV-A, our training set is created from 50 clean utterances. In the following, we examine the dependence on the number of training utterances. We retrain SVM classifiers using 20, 100, and 200 utterances mixed with the same noises N1-N3 for representative features. The overall HIT-FA results are given in Fig. 5(a) and (b) for matched and unmatched-noise conditions.

In the matched-noise condition, more utterances for training enable each feature type to improve the unit labeling performance. Specifically, we obtain about 5% improvements by increasing the number of training utterances from 20 to 200, except for RASTA-PLP, which seems to saturate when 200 utterances are used. In the unmatched-noise condition, no significant performance gain is achieved beyond 50 for GFCC and

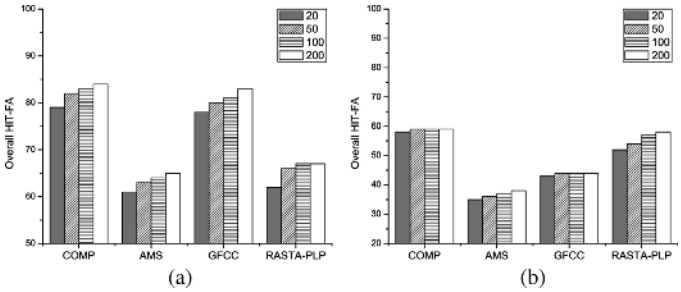



Fig. 5. Overall HIT-FA rates of representative features as a function of the number of training utterances. "COMP" stands for the complementary feature set AMS+RASTA-PLP+MFCC (a) Matched-noise condition. (b) Unmatched-noise condition.

the complementary feature set. However, for RASTA-PLP, a 5% gain is achieved by using 100 utterances compared to 20, and the performance seems to keep increasing with more training utterances. It is worth noting that the performance of the complementary feature set using only 20 training utterances surpasses the other features using more training utterances. In summary, there is a clear benefit of training on more utterances for the matched-noise condition, which is consistent with the results in [22]; yet the performance dependence on the number of training utterances in the unmatched-noise condition is significant only for certain feature types. In future research, it would be interesting to study the performance profile using even more utterances for RASTA-PLP and the complementary feature set (which contains RASTA-PLP), especially in the unmatched-noise condition.

### *H. Evaluation in Different SNR Conditions*

From a practical point of view, it is interesting to know how well a model trained on a single SNR condition generalizes to different SNR conditions. To examine this question, we use the subband SVMs already trained on 0 dB mixtures described in Section IV-A to segregate the same test mixtures at −5 dB, 5 dB, and 10 dB. Tables V and VI give the overall HIT-FA and SNR results for matched and unmatched-noise conditions. All features are impacted by the input SNR mismatch. The reason for the performance degradation seems twofold. First, a change of SNR leads to a change of power spectrum distribution at the T-F unit level, leading to a deviation from training. Second, a change of SNR also leads to a change of the IBM, which becomes denser (sparser) as SNR increases (decreases). Such a change in the prior probability of unit labels presents an issue to discriminative classifiers such as SVM. This is a clear trend in the 10 dB case, in which we observe that the HIT rate decreases significantly. Relatively speaking, MFCC and RASTA-PLP hold up well, especially at the lower SNR level. Again, the inclusion of the pitch-based features clearly helps each feature type to stabilize the labeling performance. The final combined feature set significantly outperforms the other features in each SNR condition. When ground truth pitch is used, it achieves 86%, 81%, and 72% HIT-FA in the matched-noise condition, and 75%, 75%, and 68% in the unmatched-noise condition, at −5, 5 and 10 dB SNR respectively. These results are comparable to the matched-SNR scenarios. In terms of reconstruction SNR, the combined feature set consistently and significantly improves for each input SNR condition.

TABLE V
SEGREGATION PERFORMANCE IN THE MATCHED-NOISE CONDITION WHEN TESTED ON DIFFERENT SNR CONDITIONS

| Feature | -5 dB | | 5 dB | | 10 dB | |
|---|---|---|---|---|---|---|
| | HIT−FA | SNR (dB) | HIT−FA | SNR (dB) | HIT−FA | SNR (dB) |
| AMS | 61% | 6.7 | 56% | 15.2 | 40% | 15.7 |
| RASTA-PLP | 68% | 11.3 | 62% | 15.0 | 55% | 16.0 |
| MFCC | 76% | 14.4 | 71% | 18.7 | 63% | 19.4 |
| AMS+RASTA-PLP +MFCC | **80%**$^{\dagger}$ | **14.5**$^{\ddagger}$ | **77%**$^{\dagger}$ | **21.1**$^{\ddagger}$ | 67% | **21.2**$^{\ddagger}$ |
| AMS +PITCHΔΔ | 61% | 8.9 | 61% | 14.8 | 54% | 16.2 |
| RASTA-PLP +PITCHΔΔ | 63% | 9.7 | 65% | 16.6 | 58% | 18.3 |
| MFCC +PITCHΔΔ | 71% | 10.3 | 70% | 16.0 | 64% | 17.8 |
| AMS+RASTA-PLPΔ +MFCC+PITCHΔΔ | 77% | 10.6 | 76% | 18.9 | **68%**$^{\dagger}$ | 20.4 |

TABLE VI
SEGREGATION PERFORMANCE IN THE UNMATCHED-NOISE CONDITION WHEN TESTED ON DIFFERENT SNR CONDITIONS

| Feature | -5 dB | | 5 dB | | 10 dB | |
|---|---|---|---|---|---|---|
| | HIT−FA | SNR (dB) | HIT−FA | SNR (dB) | HIT−FA | SNR (dB) |
| AMS | 30% | 1.5 | 39% | 12.2 | 34% | 14.4 |
| RASTA-PLP | 55% | 4.5 | 55% | 12.3 | 49% | 14.6 |
| MFCC | 38% | 2.1 | 50% | 13.2 | 49% | 17.2 |
| AMS+RASTA-PLP +MFCC | 52% | 2.1 | 62% | 14.6 | 58% | 18.8 |
| AMS +PITCHΔΔ | 49% | 4.6 | 58% | 13.5 | 51% | 15.3 |
| RASTA-PLP +PITCHΔΔ | 57% | 4.6 | 64% | 14.6 | 56% | 17.3 |
| MFCC +PITCHΔΔ | 55% | 4.4 | 64% | 14.2 | 58% | 16.8 |
| AMS+RASTA-PLPΔ +MFCC+PITCHΔΔ | **62%**$^{\dagger}$ | **4.7**$^{\dagger}$ | **71%**$^{\dagger}$ | **15.3**$^{\ddagger}$ | **64%**$^{\dagger}$ | **18.9**$^{\ddagger}$ |

TABLE VII
SEGREGATION PERFORMANCE ON THE IEEE MALE TALKER

| Feature | Matched-Noise Condition | | | Unmatched-Noise Condition | | |
|---|---|---|---|---|---|---|
| | HIT−FA | Accuracy | SNR (dB) | HIT−FA | Accuracy | SNR (dB) |
| RASTA-PLP | 61% | 84.1% | 13.3 | 52% | 82.9% | 8.5 |
| RASTA-PLP +PITCHΔΔ | 63% | 85.5% | 12.9 | 59% | 86.4% | 9.9 |
| GFCC | 78% | 90.7% | 15.1 | 41% | 72.5% | 5.7 |
| GFCC +PITCHΔΔ | 76% | 90.1% | 13.9 | 54% | 80.8% | 8.1 |
| AMS+RASTA-PLP +MFCC | 79% | 91.3% | **17.8**$^{\ddagger}$ | 56% | 78.7% | 7.9 |
| AMS+RASTA-PLPΔ +MFCC+PITCHΔΔ | **80%** | **91.6%**$^{\ddagger}$ | 16.3 | **66%**$^{\ddagger}$ | **87.5%**$^{\ddagger}$ | **10.3**$^{\ddagger}$ |

### *I. Generalization to Different Speakers*

Previous experiments are mainly based on the IEEE female talker. We now show that the key conclusions hold for the IEEE male talker as well. The training and testing settings are the same as before, except that data from a male talker are used. Table VII shows the segregation results from representative features. As in the female case, GFCC is good as a single feature, PITCH is effective for generalization, and combined features are better than single features.

To further test generalization to different speakers, we create a new test set for each gender by mixing 20 utterances from the TIMIT corpus [10] with N1-N6 at 0 dB. The new test utterances are chosen from 10 different TIMIT speakers of the same gender, each providing 2 utterances. We use the models previously trained on the IEEE corpus for each gender on the new test set without change. The results of representative features for unseen female and male talkers are shown in Tables VIII and IX, respectively. The classification performance is expected to degrade when tested on unseen speakers, as is evident from the

TABLE VIII
SEGREGATION PERFORMANCE WHEN TESTED ON TIMIT FEMALE SPEAKERS

| Feature | Matched-Noise Condition | | | Unmatched-Noise Condition | | |
|---|---|---|---|---|---|---|
| | HIT−FA | Accuracy | SNR (dB) | HIT−FA | Accuracy | SNR (dB) |
| RASTA-PLP | 56% | 81.2% | 11.3 | 47% | 81.0% | 8.1 |
| RASTA-PLP +PITCH$\Delta\Delta$ | 60% | 81.8% | 12.2 | 59% | 84.8% | 9.8 |
| GFCC | 71% | 85.9% | 13.7 | 40% | 67.3% | 6.2 |
| GFCC +PITCH$\Delta\Delta$ | 67% | 84.2% | 13.0 | 51% | 76.4% | 8.1 |
| AMS+RASTA-PLP +MFCC | **72%** | **87.9%**$^{\ddagger}$ | **15.6**$^{\ddagger}$ | 56% | 81.0% | 9.7 |
| AMS+RASTA-PLP$\Delta$ +MFCC+PITCH$\Delta\Delta$ | **72%** | 87.0% | 14.4 | **66%**$^{\ddagger}$ | **86.1%**$^{\ddagger}$ | **10.7**$^{\ddagger}$ |

TABLE IX
SEGREGATION PERFORMANCE WHEN TESTED ON TIMIT MALE SPEAKERS

| Feature | Matched-Noise Condition | | | Unmatched-Noise Condition | | |
|---|---|---|---|---|---|---|
| | HIT−FA | Accuracy | SNR (dB) | HIT−FA | Accuracy | SNR (dB) |
| RASTA-PLP | 50% | 77.9% | 10.1 | 40% | 78.3% | 7.0 |
| RASTA-PLP +PITCH$\Delta\Delta$ | 59% | 82.5% | 11.1 | 57% | 85.3% | 9.5 |
| GFCC | 64% | 84.4% | 7.8 | 32% | 69.3% | 4.0 |
| GFCC +PITCH$\Delta\Delta$ | 69% | 86.4% | 10.0 | 50% | 78.7% | 6.8 |
| AMS+RASTA-PLP +MFCC | 65% | 84.6% | 11.2 | 47% | 78.0% | 6.9 |
| AMS+RASTA-PLP$\Delta$ +MFCC+PITCH$\Delta\Delta$ | **72%**$^{\ddagger}$ | **87.3%**$^{\ddagger}$ | **12.3**$^{\ddagger}$ | **62%**$^{\ddagger}$ | **85.7%**$^{\ddagger}$ | **9.9**$^{\ddagger}$ |

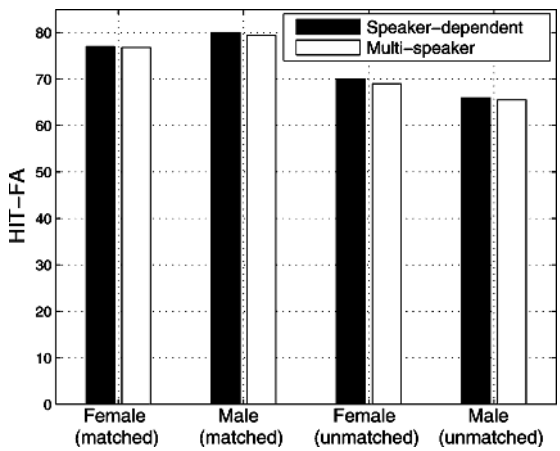



Fig. 6. Overall HIT-FA comparisons between speaker-dependent and multi-speaker classifiers on the IEEE corpus.

performance of single features. Adding PITCH clearly helps. The feature combinations are more robust than single features, and the final combined feature set performs reasonably well compared to the matched-speaker case for both genders.

Our preliminary results on cross-gender generalization show that all the above features perform worse, presumably due to significant deviations of spectro-temporal distributions between the two genders. Two methods can be used to deal with the cross-gender issue. First, one can first identify the gender of the target speech and then use gender-dependent classifiers. Gender identification can be achieved with high accuracy [40]. Second, one can train classifiers by including the multiple speakers of both genders into the training set. We show the results of using the second method by training a classifier on the IEEE female and male talkers and test on mixtures from both. Fig. 6 shows the overall HIT-FA results, and the performance of the multi-speaker classifier is nearly as good as that of using corresponding speaker-dependent classifiers. These results indicate that the selected features perform well across different speakers.

## V. DISCUSSION

Since different subbands in a gammatone filterbank are not independent, it is reasonable to use frame-level features directly in training subband classifiers (see [39]), rather than using T-F unit level features as done in this paper. We have tried such training using conventional frame-level features. We have opted for using T-F unit level features mainly because our experiments show that, although frame-level features produce comparable performance in matched-noise conditions, the performance is significantly worse than unit-level features in unmatched test conditions. Frame-level features, such as GFCC, may be more susceptible to local distortions in a few subbands than unit-level features, as suggested in robust automatic speech recognition (ASR) [33]. Also, features such as pitch-based ones are defined at the T-F unit level, which may create issues for feature combination if other features are derived at the frame level. Nevertheless, it is an interesting question if one can extract unit-level features directly from frame-level ones; if so, feature extraction could be significantly sped up. It may be easy for some features such as energy, but it is unclear how this could be done for cepstral features.

Formulating monaural speech segregation as binary classification has been shown as an effective approach in both speech segregation and robust ASR domains. Nevertheless, only pitch and AMS have been employed as primary T-F unit level features so far. In this paper, we have significantly expanded the unit level feature repository to include features commonly used in speech and speaker processing. For both voiced and unvoiced speech segregation, these newly included features have achieved significant improvements in terms of SNR as well as HIT-FA, a criterion that is well correlated with human speech intelligibility. In terms of single features, GFCC shows excellent performance in the matched-noise test condition, and RASTA-PLP in the unmatched conditions.

The complementarity among these features is systematically exploited by using a group Lasso approach, which selects a compact set of important feature types contributing to target and interference discrimination. The complementary feature set AMS+RASTA-PLP+MFCC has shown stable performance in various test conditions and outperforms each of its components significantly.

Generalization is a critical issue for classification-based speech segregation. We have examined the generalization performance of each feature type in several unmatched conditions. These results point to the robustness of the pitch-based features, which are parameterized by estimated pitch. Pitch-based features have also been shown to generalize well to reverberant conditions in classification-based segregation [20]. Nevertheless, the pitch-based features need to be combined with general acoustic features in order to segregate unvoiced speech and improve voiced speech segregation. The final combined feature set achieves promising segregation results in various test conditions. We plan to address reverberant speech segregation in future work using this combined feature set.

In addition to pitch, our results suggest that RASTA filtering also plays an important role in good generalization. RASTA filtering effectively captures low modulation frequencies corresponding to speech. The inclusion of this speech property sig-

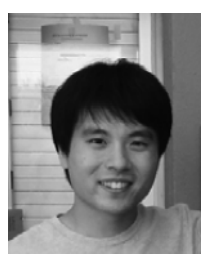
nificantly reduces FA rates, which degrade significantly in unmatched conditions. It would be interesting to explore new features that characterize both pitch and low modulation frequencies in future research.

## ACKNOWLEDGMENT

The authors would like to thank Z. Jin for providing his pitch tracking code.

## REFERENCES

[1] J. Allen, *Articulation and Intelligibility*. San Rafael, CA: Morgan & Claypool, 2005.
[2] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1994.
[3] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer (Version 4.3.14)," 2005 [Online]. Available: http://www.fon.hum.uva.nl/praat
[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.
[5] D. Brungart, P. Chang, B. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4007–4018, 2006.
[6] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Aug. 2006.
[7] H. Dillon, *Hearing Aids*. New York: Thieme, 2001.
[8] D. Ellis, "PLP and RASTA (and MFCC, and Inversion) in Matlab," 2005 [Online]. Available: http://www.ee.columbia.edu/dpwe/resources/matlab/rastamat/
[9] G. Garau and S. Renals, "Combining spectral representations for large-vocabulary continuous speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 508–518, Mar. 2008.
[10] J. Garofolo, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," NIST, 1993.
[11] K. Han and D. Wang, "An SVM based classification approach to speech separation," in *Proc. ICASSP*, 2011, pp. 5212–5215.
[12] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.
[13] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
[14] G. Hu, "Monaural speech organization and segregation," Ph.D. dissertation, The Ohio State Univ., Biophysics Program, Columbus, OH, 2006.
[15] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.
[16] G. Hu and D. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Amer.*, vol. 124, pp. 1306–1319, 2008.
[17] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.
[18] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, Sep. 1969.
[19] G. Jang and T. Lee, "A maximum likelihood approach to single-channel source separation," *J. Mach. Learn Res.*, vol. 4, pp. 1365–1392, 2003.
[20] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 625–638, May 2009.
[21] Z. Jin and D. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1091–1102, Jul. 2011.
[22] G. Kim and P. Loizou, "Improving speech intelligibility in noise using environment-optimized algorithms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2080–2090, Nov. 2010.
[23] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, pp. 1486–1494, 2009.
[24] N. Li and P. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1673–1682, 2008.
[25] L. Meier, S. V. D. Geer, and P. Bühlmann, "The group Lasso for logistic regression," *J. R. Stat. Soc. Series B*, vol. 70, no. 1, pp. 53–71, 2008.
[26] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," APU Report, 1988.
[27] S. Roweis, "One microphone source separation," *NIPS*, pp. 793–799, 2001.
[28] M. Schmidt and R. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. ICSLP*, 2006.
[29] M. Seltzer, B. Raj, and R. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 379–393, 2004.
[30] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Proc. ICASSP*, 2009, pp. 4625–4628.
[31] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. ICASSP*, 2008, pp. 1589–1592.
[32] T. Takiguchi and Y. Ariki, "Robust feature extraction using kernel PCA," in *Proc. ICASSP*, 2006, pp. 509–512.
[33] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," in *Proc. ICASSP*, 1997, pp. 1255–1258.
[34] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Series B*, vol. 58, no. 1, pp. 267–288, 1996.
[35] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.
[36] *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. Wang and G. Brown, Eds. Hoboken, NJ: Wiley-IEEE Press, 2006.
[37] D. Wang, U. Kjems, M. Pedersen, J. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, pp. 2336–2347, 2009.
[38] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," Dept. of CSE, Ohio State Univ., 2011, Tech. Rep. TR37.
[39] R. Weiss and D. Ellis, "Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking," in *Proc. Workshop Statist. Percept. Audition*, 2006.
[40] K. Wu and D. Childers, "Gender recognition from speech. Part I: Coarse analysis," *J. Acoust. Soc. Amer.*, vol. 90, no. 4, pp. 1828–1840, 1991.
[41] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Stat. Soc. Series B*, vol. 68, no. 1, pp. 49–67, 2006.
[42] A. Zolnay, D. Kocharov, R. Schlüter, and H. Ney, "Using multiple acoustic feature sets for speech recognition," *Speech Commun.*, vol. 49, no. 6, pp. 514–525, 2007.



**Yuxuan Wang** received his B.E. degree in network engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009. He is currently pursuing his Ph.D. degree at The Ohio State University. He is interested in machine learning, optimization, speech separation, and computational neuroscience.

**Kun Han,** photograph and biography not available at the time of publication.

**DeLiang Wang,** photograph and biography not available at the time of publication.