Technical Report OSU-CISRC-11/11-TR37
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2011**
File: **TR37.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

Exploring Monaural Features for Classification-Based Speech Segregation

Yuxuan Wang
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
wangyuxu@cse.ohio-state.edu

Kun Han
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
hank@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive Science
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

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, only pitch and amplitude modulation spectrogram have been used as 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). 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 and unmatched test conditions, respectively. We also find that pitch-based features are crucial for good generalization. To further explore complementarity

in terms of discriminative power, we propose to use a group Lasso approach to combine different features in a principled way. The final combined feature yields very promising results in both matched and unmatched test conditions.

Index Terms – Computational auditory scene analysis (CASA), monaural speech segregation, binary classification, feature combination, group Lasso.

1 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. 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 non-speech interference.

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 [16, 24, 25] 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 [31]. 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 [1, 5, 20, 33] 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. To our knowledge, the first attempt to formulate speech segregation as a binary classification problem was made in the binaural domain [23]. Recent studies have applied this formulation in the monaural domain and achieved good speech segregation results in both anechoic and reverberant environments [9, 12, 17, 19, 26]. In [12, 17], the pitch-based feature is used in training a classifier to separate target and interference dominant units. However, the pitch-based feature cannot deal with unvoiced speech that lacks harmonic structure. In [9, 19], 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 [9].

For classification, the use of an appropriate classifier is obviously important. The study in [9] suggests that support vector machines (SVMs) are more powerful than Gaussian mixture models (GMMs). Equally important is the choice of appropriate features. So far, only pitch and AMS have been studied. On the other hand, in the speech and speaker recognition

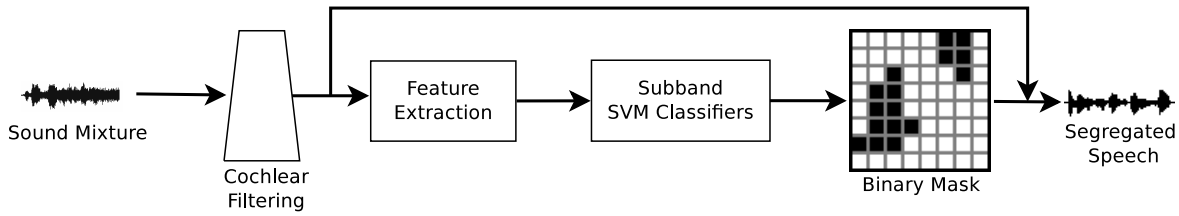


Figure 1: Schematic diagram of a classification-based speech segregation system. First, a sound mixture is fed into a 64-channel gammatone filterbank to produce a cochleagram. Acoustic features for each T-F unit are then extracted. SVM classifiers are trained for each channel, and their classification yields an estimate of the IBM. After gating the cochleagram by the estimated mask, target speech is segregated.

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.

In this paper we explore the use of existing speech and speaker features, in order to enlarge possible features for speech segregation. It is shown in speech recognition that complementarity exists between basic acoustic features [8, 35]. Even if individual features do not perform well, their combination may yield large performance boosts. In order to investigate complementary features in terms of discriminative power in a principled way, we propose to address the corresponding group variable selection problem using a group least absolute shrinkage and selection operator (Lasso) [34]. Group Lasso extends the widely used Lasso [30] to perform feature selection at the group level.

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 2. Section 3 describes a group Lasso approach to combining different features. Unit labeling results in matched and unmatched test conditions are reported in Section 4. To simplify the system, a dimension reduction method based on bandwidth analysis is presented in Section 5. We conclude this paper in Section 6.

2 System overview and Feature Extraction

The architecture of our segregation system is shown in Fig. 1. A sound mixture with 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. 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 [32]. 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 [9] with all parameters cross-validated for each subband channel separately, and ground truth labels are provided by the IBM. Feature

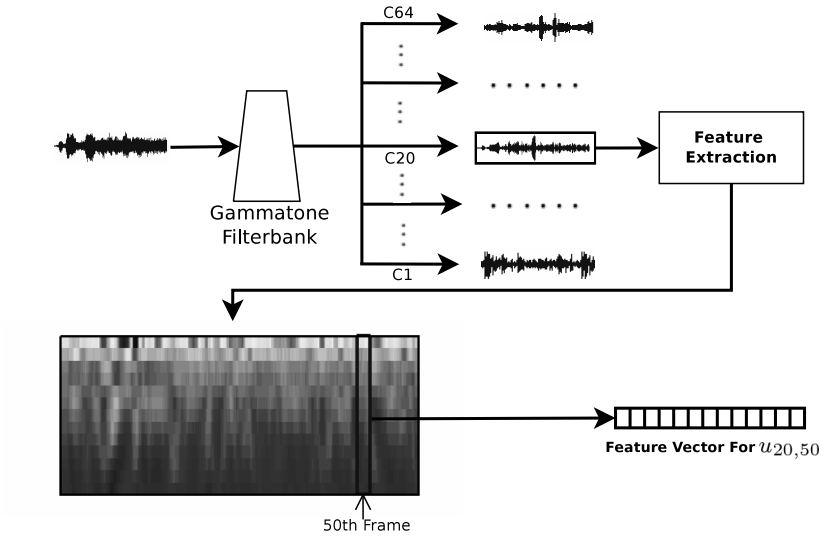


Figure 2: Illustration of deriving RASTA-PLP features for the T-F unit in channel 20 and at frame 50 ($u_{20,50}$). The mixture signal is bandpassed by a 64-channel gammatone filterbank. Then, the output in channel 20 is fed into the conventional RASTA-PLP feature extractor, which in essence extracts RASTA-PLPs for the subband signal. Finally, we take the values at frame 50 as the feature vector for $u_{20,50}$.

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 [32]. Note that we do not perform auditory segmentation, which is usually done for better segregation [9, 17], as we want to directly compare the unit labeling performance of each feature.

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 $u_{c,m}$. For features derived in a frame-by-frame manner (such as MFCC and PLP), this procedure is equivalent to a spectral/cepstral analysis solely based on the slice of the subband signal contained in each T-F unit; i.e., it is equivalent to windowing (with overlapping) the subband signal, and performing spectral/cepstral analysis afterwards. But our procedure also enables us to derive T-F unit features involving neighboring frames, as done in RASTA filtering, in a convenient way. Note that the described procedure is certainly not the only way to derive unit level features, and obviously the features derived in this way contain redundant information. Nevertheless, this simple procedure is effective in our experiments and we also present a method to reduce the dimensionality for unit features based on bandwidth analysis in Section 5. Fig. 2 illustrates how to derive a 12th order RASTA-PLP feature (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 characterize different aspects of the speech signal. We make use of the RASTAMAT toolbox [7] for extracting MFCC, PLP, and RASTA-PLP features.

2.1 Pitch-based Feature

Pitch is a primary cue for ASA. In our experiments, we use a set of pitch-based features originally proposed in [12], and its effectiveness has been confirmed in both anechoic and reverberant environments with additive noise [14, 17]. To get pitch-based features for $u_{c,m}$, we first calculate the normalized autocorrelation function at each time lag τ, denoted by $A(c, m, \tau)$:

$$\begin{aligned} &A(c, m, \tau) \\ &= \frac{\sum_n x_c(mT_m - nT_n)x_c(mT_m - nT_n - \tau T_n)}{\sqrt{\sum_n x_c^2(mT_m - nT_n)}\sqrt{\sum_n x_c^2(mT_m - nT_n - \tau T_n)}} \end{aligned} \tag{1}$$

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 τ_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 τ_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 $A(c, m, \tau_m)$ [14].

The next three features are the same as the first three except that they are extracted from the envelopes of filter responses. The resulting six-dimensional feature is:

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

where $[\cdot]$ denotes the round operation, and subscript E indicates envelope.

2.2 Amplitude Modulation Spectrogram

AMS features have been applied to speech segregation problems recently [19]. To extract AMS features, we extract the envelope of the mixture signal 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, producing a 15-dimensional AMS feature.

2.3 Gammaotone Frequency Cepstral Coefficient

GFCCs are shown to be effective for robust speaker identification [28]. To get GFCC features, 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.

2.4 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 a fair comparison with GFCCs in which a 64-channel gammatone filterbank is used for subband analysis.

2.5 Perceptual Linear Prediction

PLP [10] is a popular feature 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 feature.

2.6 Relative Spectral Transform-PLP

RASTA filtering [11] 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 filer (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.

3 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 [8, 35]. 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 [29] on the concatenated feature vector. The third way is to apply supervised feature transformation such as linear discriminant analysis [8] to the concatenated feature vector. However, an issue with feature transformation relates to complementarity; i.e., it is unclear which features are complementary after transformation.

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 [34], a generalization of the widely used Lasso operator [30], 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 [22], which can be efficiently solved by block coordinate gradient descent. The estimator is

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

where $\mathbf{x}_i$ is the ith training sample, y_i is the ground truth label scaled to $\{-1, 1\}$, and a is a parameter (intercept). $\|\cdot\|_2$ refers to the ℓ_2 norm. $\boldsymbol{\beta}$ consists of G predefined non-overlapping groups and $\mathcal{I}_g$ is the index set of the gth group. The first term in the minimization is a standard log loss that concerns discrimination. The second term is an ℓ_1/ℓ_2 mixed-norm regularization, which imposes an ℓ_1 regularization between groups and an ℓ_2 regularization within each group. It is well known that the ℓ_1 norm induces sparsity, therefore the ℓ_1/ℓ_2 regularization results in group sparsity hence group level feature selection. Regularization parameter λ controls the level of sparsity of the resulting model. In practice, we usually calculate $\lambda_{\max}$ first, above which $\hat{\boldsymbol{\beta}}_\lambda$ is all zero. We then use $\gamma \cdot \lambda_{\max}$ with $\gamma \in [0, 1]$ as λ in Equation (3) for the ease of choosing appropriate parameter values. In our experiments, we make use of the SLEP sparse learning package [21] for efficient optimization.

To do feature combination using group Lasso, all the features are concatenated together to form a long feature vector, and each feature is defined as a group. Then, for a fixed γ

(hence λ), we solve Equation (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 (features) contribute little to discrimination in the presence of the other groups. Groups having large regression coefficients shall be included in the complementary feature set. To achieve a good trade-off between discrimination power and model complexity which is the number of groups selected, we cross-validate γ from 0.1 to 0.9 with the step size of 0.1. It should be noted that the above procedure is carried out at each channel separately. That is, we solve Equation (3) for each γ using training samples from each channel. A subband SVM classifier is then trained on the resulting combined feature for cross-validation, yielding $64 \times 9 = 576$ cross-validation accuracies, which are then averaged across channels. We empirically determine the final combination by leveraging the averaged cross-validation accuracies with the corresponding model complexity. For an illustration, see Fig. 5 in Section 4.5.

4 Evaluation Results

4.1 Experimental Setup

We use the IEEE corpus [15] for all 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 with water flowing, N2 – crow noise, and N3 – cocktail party noise [12]. We choose 10 new utterances from the IEEE corpus for testing. Two test conditions are employed. In condition one, we mix the test utterances with the trained noises (i.e., N1-N3) in order to test the performance on unseen utterances. In condition two, the test utterances are mixed with three unseen noises: N4 – crowd noise at a playground, N5 – traffic noise, and N6 – electric fan noise. Unless stated otherwise, the test mixtures are mixed at 0 dB.

As mentioned in Section 2, the dimensionality of the pitch-based feature and AMS feature is 6 and 15, respectively. We use a 31-D GFCC feature as suggested in [27]. Following common practice in speech recognition, we use a 12th order linear prediction model, yielding 13-D (including zeroth cepstral coefficient) PLP and RASTA-PLP features. Initially, we take the first 13 DCT coefficients to form a 13-D MFCC feature, as usually done in speech recognition. To include more harmonic structure in the representation, we also increase the MFCC dimension to 31. For comparison, we also use a 13-D GFCC feature, and we denote MFCC/GFCC with two dimensionalities as MFCC13/GFCC13 and MFCC31/GFCC31. For the pitch-based feature, 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 [18]. We use PITCH to denote the pitch-based feature.

To put the performance of our classification-based segregation in perspective, we include results from a recent CASA system, the tandem algorithm [14], which jointly performs voiced

Table 1: HIT−FA rates for single features in condition one. Boldface indicates best rate

Feature	Overall			Voiced			Unvoiced		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
AMS	70%	7%	63%	76%	8%	68%	47%	5%	42%
PLP	79%	8%	71%	82%	9%	73%	65%	6%	59%
RASTA-PLP	73%	7%	66%	77%	9%	68%	56%	5%	51%
GFCC13	84%	6%	78%	87%	8%	79%	74%	4%	70%
MFCC13	79%	9%	70%	82%	11%	71%	69%	7%	62%
GFCC31	86%	6%	**80%**	89%	7%	**82%**	76%	4%	**72%**
MFCC31	83%	7%	76%	86%	8%	78%	70%	5%	65%
PITCH	N/A	N/A	N/A	76%	16%	60%	N/A	N/A	N/A
TANDEM	N/A	N/A	N/A	74%	4%	70%	N/A	N/A	N/A

Table 2: HIT−FA rates for single features in condition two

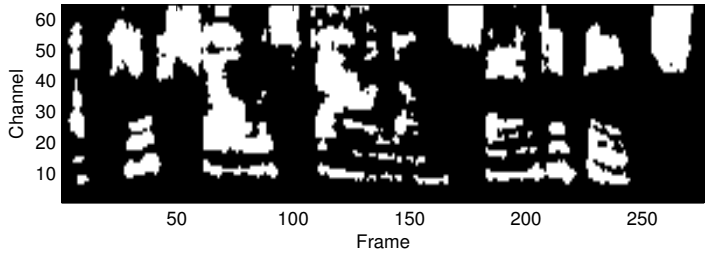
Feature	Overall			Voiced			Unvoiced		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
AMS	59%	23%	36%	64%	22%	42%	41%	25%	16%
PLP	70%	28%	42%	72%	27%	45%	62%	29%	32%
RASTA-PLP	68%	12%	**56%**	70%	13%	57%	58%	10%	**48%**
GFCC13	68%	31%	37%	68%	30%	38%	70%	31%	39%
MFCC13	70%	33%	37%	71%	33%	38%	67%	34%	33%
GFCC31	74%	30%	44%	74%	29%	45%	74%	32%	42%
MFCC31	74%	28%	46%	75%	28%	47%	68%	28%	40%
PITCH	N/A	N/A	N/A	76%	20%	56%	N/A	N/A	N/A
TANDEM	N/A	N/A	N/A	67%	4%	**63%**	N/A	N/A	N/A

speech segregation and pitch estimation in an iterative fashion. The tandem algorithm is initialized by the same estimated pitch. We use ideal sequential grouping for the tandem algorithm as it does not address the problem of grouping pitch contours and their associated binary masks across time. So these results represent the ceiling performance of the tandem algorithm.

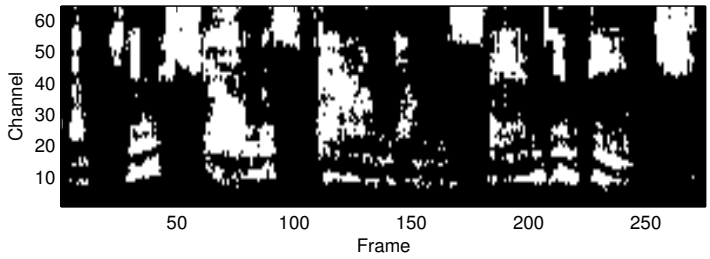
The evaluation criterion used throughout the experiments is the hit minus false alarm rate (HIT−FA). The HIT rate is the percent of correctly classified target-dominant T-F units in the IBM. The FA rate is the percent of wrongly classified interference-dominant T-F units in the IBM. HIT−FA has been shown to be highly correlated with human speech intelligibility [19, 20].

4.2 Single Features

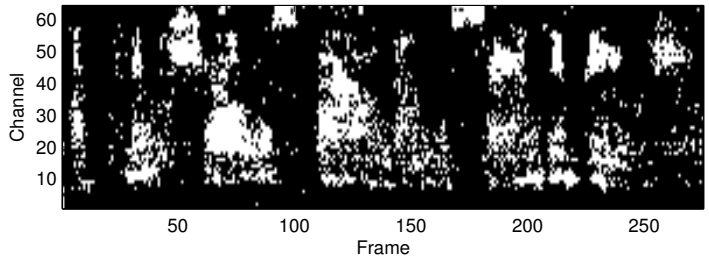
In terms of HIT−FA, we document unit labeling performance at three levels: voiced speech intervals (pitched frames), unvoiced speech intervals (unpitched frames), and overall. Voiced and unvoiced speech intervals are determined by ground truth pitch. Table 1 gives the results in test condition one. In 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,



(a) IBM



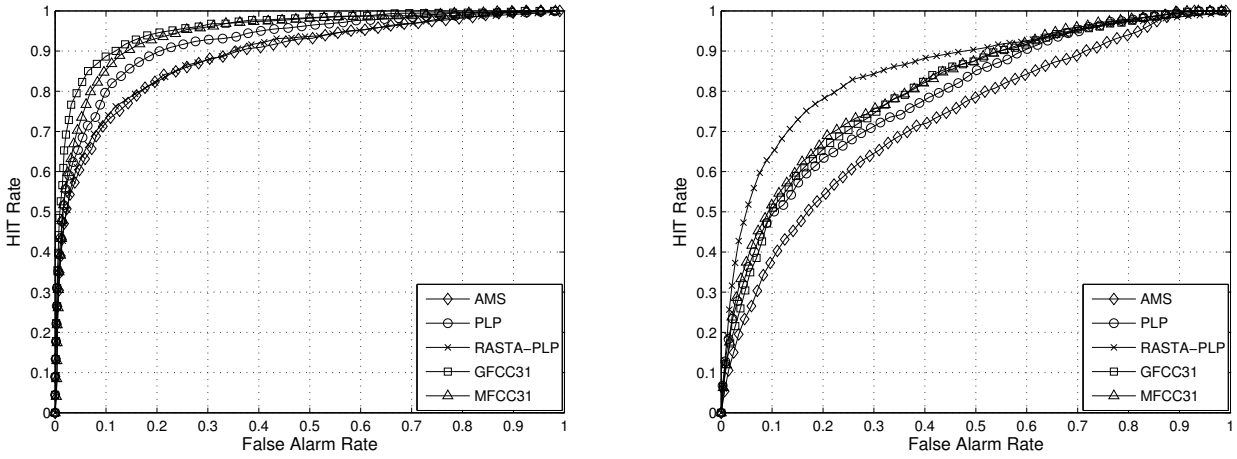
(b) GFCC31 Mask



(c) AMS Mask

Figure 3: Binary masks for a test utterance mixed with the cocktail party noise at 0 dB. Top panel shows the IBM. Middle and bottom panel show the estimated IBMs obtained by using GFCC31 and AMS, respectively.

GFCC31 manages to achieve high HIT rates, with 80% overall HIT−FA, which is significantly better than other single features. Unvoiced speech is important to speech intelligibility, and its segregation is a difficult task due to the lack of harmonicity and weak energy [13]. Again, AMS performs the worst whereas GFCCs do a very good job at segregating unvoiced speech. With the same dimensionality, GFCC outperforms MFCC. The good performance of GFCC is probably due to its effectiveness as a speaker feature [28]. It is interesting to note that increasing the dimensionality of MFCC from 13 to 31 significantly improves its performance in speech segregation, which indicates that keeping more harmonic structure in the representation is helpful. An encouraging observation in the matched-interference test condition is that some general features such as GFCC31 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. Note that the HIT−FA rates with GFCC31 significantly exceed those reported in [19] which uses GMM classifiers on AMS features. Fig. 3 illustrates the binary



(a) Test condition one

(b) Test condition two

Figure 4: ROC curves for overall classification obtained by single features in condition one and two.

masks obtained by using GFCC31 and AMS for a test utterance mixed with the cocktail party noise. For systematic comparison, the receiver operating characteristic (ROC) curves for overall classification obtained by using the single features are shown in Fig. 4(a). The ROC curves are generated from the SVM decision values.

Unlike test condition one, the unseen broadband noises are more demanding for generalization. The HIT−FA results in this condition are listed in Table 2. We can see that 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 with only 4% reduction in HIT−FA. Using ground truth pitch it is able to achieve 68% HIT−FA in voiced intervals. As the pitch-based feature reflects 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. It is interesting to see that RASTA-PLP becomes the best performing feature in condition two. As shown in [11], RASTA-PLP effectively acts as a modulation-frequency filter, which retains slow modulations corresponding to speech. The ROC curves for overall classification in test condition two are shown in Fig. 4(b).

Given their superior performance, in the following we only include 31-D results for MFCC and GFCC features.

4.3 Combining with Pitch-based Feature

Considering the excellent performance of some features in the matched-interference condition and the robustness of the pitch-based feature in the unmatched-interference condition, it seems

Table 3: HIT−FA results for pairwise combination of single features and pitch-based feature in test condition one

Feature	Overall			Voiced			Unvoiced		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
AMS+PITCH	73%	12%	61%	81%	10%	71%	41%	14%	27%
PLP+PITCH	78%	10%	68%	84%	9%	75%	58%	12%	46%
RASTA-PLP+PITCH	76%	10%	66%	83%	9%	74%	47%	11%	36%
GFCC31+PITCH	83%	9%	**74%**	88%	8%	**80%**	65%	11%	**54%**
MFCC31+PITCH	80%	10%	70%	85%	8%	77%	62%	12%	50%

Table 4: HIT−FA results for pairwise combination of single features and pitch-based feature in test condition two

Feature	Overall			Voiced			Unvoiced		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
AMS+PITCH	65%	12%	53%	73%	12%	61%	29%	11%	18%
PLP+PITCH	71%	12%	59%	76%	13%	63%	50%	9%	41%
RASTA-PLP+PITCH	72%	10%	**62%**	77%	12%	**65%**	50%	7%	43%
GFCC31+PITCH	75%	20%	55%	79%	19%	60%	62%	23%	39%
MFCC31+PITCH	73%	13%	60%	77%	13%	64%	55%	11%	**44%**

sensible to combine the single features with the pitch-based feature. Table 3 lists the HIT−FA results for pairwise combinations in test condition one. Due to pitch estimation errors, the combination does not improve the performance in this test condition. However, we find that the combination using the ground truth pitch significantly improves the performance for all the features. Results for the second test condition are listed in Table 4. Even with estimated pitch, the performance of all the features is significantly boosted by the combination, demonstrating the role of the pitch-based feature in generalization to unseen noises. As before, RASTA-PLP leads the overall performance in this combination.

4.4 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 [19]. Table 5 and Table 6 show the HIT−FA results by adding first-order delta features (denoted by Δ) along time in two test conditions. We can clearly see improvements in both test conditions. Two observations are in order. First, adding deltas is helpful for unvoiced speech segregation. Second, almost all features benefit from adding deltas in the unmatched condition, indicating their effect in improving generalization.

We have also experimented with adding deltas along frequency channel as suggested in [19]. However this yields only small improvements (approximately 1% to 3%) at the expense of added dimensionality. As a trade-off, we suggest adding frequency deltas only for the pitch-based feature which has a low dimensionality, producing a 18-D feature denoted by PITCH$\Delta\Delta$.

Table 5: HIT−FA results by including first-order delta features in test condition one

Feature	Overall			Voiced			Unvoiced		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
AMSΔ	77%	7%	70%	82%	8%	74%	55%	4%	51%
PLPΔ	83%	7%	76%	86%	9%	77%	71%	4%	67%
RASTA-PLPΔ	80%	7%	73%	83%	9%	74%	69%	5%	64%
GFCC31Δ	87%	5%	**82%**	89%	7%	**82%**	77%	3%	**74%**
MFCC31Δ	85%	6%	79%	88%	7%	81%	72%	4%	68%
PITCHΔ	N/A	N/A	N/A	76%	15%	61%	N/A	N/A	N/A

Table 6: HIT−FA results by including first-order delta features in test condition two

Feature	Overall			Voiced			Unvoiced		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
AMSΔ	67%	26%	41%	71%	26%	45%	51%	28%	23%
PLPΔ	73%	22%	51%	74%	23%	51%	67%	20%	47%
RASTA-PLPΔ	73%	13%	**60%**	73%	14%	**59%**	69%	11%	**58%**
GFCC31Δ	75%	28%	47%	76%	28%	48%	75%	28%	47%
MFCC31Δ	75%	19%	56%	76%	20%	56%	68%	16%	52%
PITCHΔ	N/A	N/A	N/A	76%	19%	57%	N/A	N/A	N/A

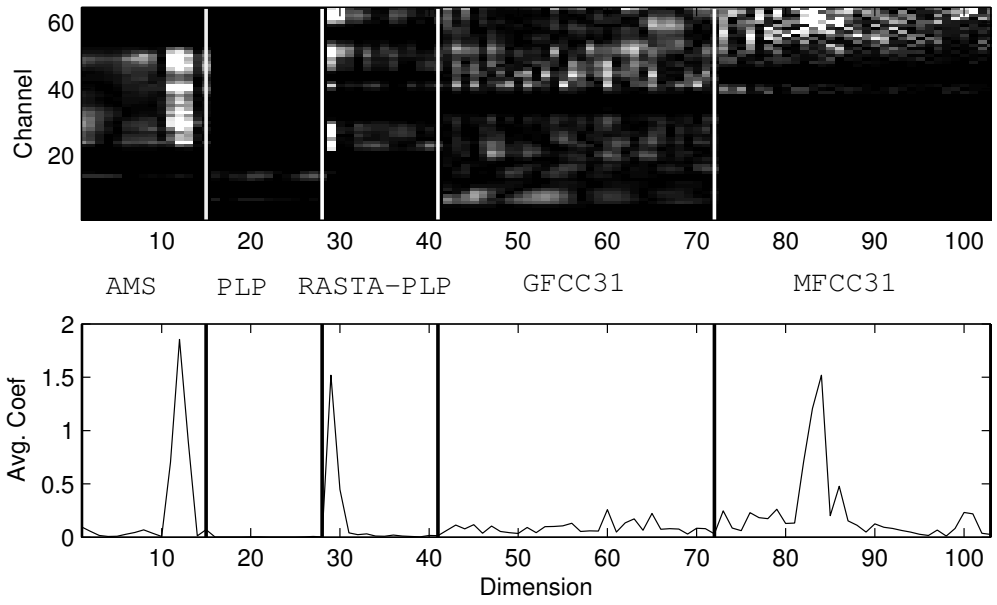


Figure 5: Regression coefficients. The top panel shows the magnitudes of regression coefficients for each channel, where brighter color indicates higher value. The bottom panel shows the averages across 64 channels.

4.5 Feature Combination

In this subsection, we evaluate feature combination as described in Section 3. Since we have shown the importance of adding the pitch-based feature, we focus on selecting complementary features from the rest. We empirically found that $\gamma = 0.2$ offers a good trade-off between

Table 7: HIT−FA results for feature combination in test condition one

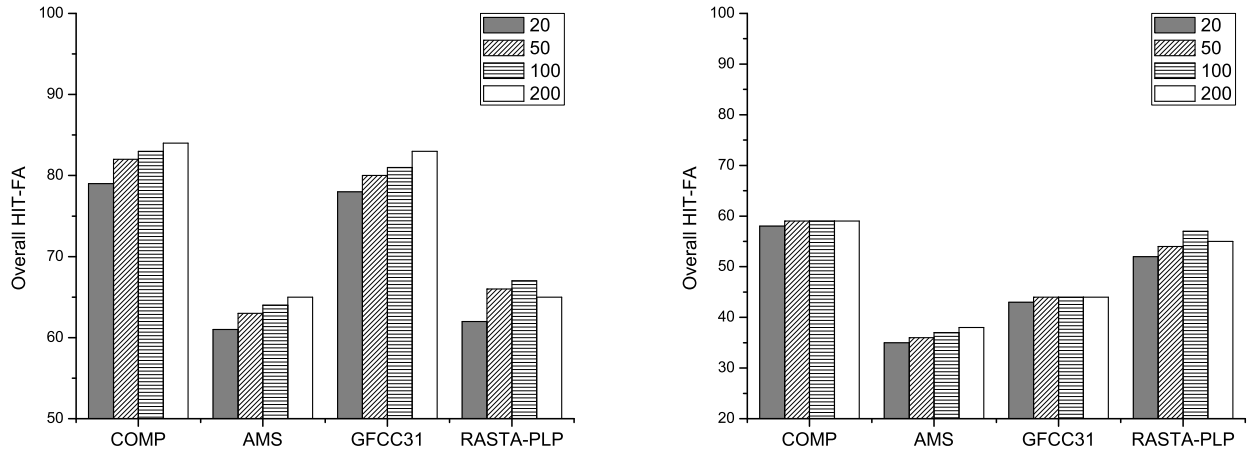
Feature Combination	Overall			Voiced			Unvoiced		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
AMS+RASTA-PLP +MFCC31	87%	5%	**82%**	89%	6%	**83%**	77%	4%	**73%**
AMS +PITCHΔΔ	73%	11%	62%	83%	9%	74%	36%	13%	23%
RASTA-PLP +PITCHΔΔ	72%	12%	60%	82%	13%	69%	39%	12%	27%
GFCC31 +PITCHΔΔ	82%	9%	73%	87%	8%	79%	63%	11%	52%
MFCC31 +PITCHΔΔ	80%	9%	71%	86%	8%	78%	58%	11%	47%
AMS+RASTA-PLPΔ +MFCC31+PITCHΔΔ	85%	7%	78%	89%	7%	82%	71%	9%	62%
PITCHΔΔ	N/A	N/A	N/A	80%	17%	63%	N/A	N/A	N/A

Table 8: HIT−FA results for feature combination in test condition two

Feature Combination	Overall			Voiced			Unvoiced		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
AMS+RASTA-PLP +MFCC31	81%	22%	59%	81%	22%	59%	81%	22%	59%
AMS +PITCHΔΔ	66%	10%	56%	75%	11%	64%	27%	8%	19%
RASTA-PLP +PITCHΔΔ	73%	10%	63%	79%	12%	67%	48%	8%	40%
GFCC31 +PITCHΔΔ	75%	19%	56%	78%	17%	61%	62%	24%	38%
MFCC31 +PITCHΔΔ	74%	12%	62%	79%	12%	67%	53%	11%	42%
AMS+RASTA-PLPΔ +MFCC31+PITCHΔΔ	81%	11%	**70%**	83%	12%	**71%**	72%	10%	**62%**
PITCHΔΔ	N/A	N/A	N/A	80%	24%	56%	N/A	N/A	N/A

model complexity and cross-validation accuracy. For $\lambda = 0.2\lambda_{\max}$, we plot the magnitudes of regression coefficients for each channel in the top panel of Fig. 5 and their averages across 64 channels in the bottom panel. It is clear that AMS, RASTA-PLP and MFCC31 are associated with large regression coefficients, while the coefficients of PLP are zero in almost all channels. GFCC31's contribution to model fitting is relatively weak, making it almost redundant given AMS, RASTA-PLP and MFCC31. Therefore, we choose AMS+RASTA-PLP+MFCC31 as our complementary feature set. Leveraging deltas, we set the final combined feature to AMS+RASTA-PLPΔ+MFCC31+PITCHΔΔ, resulting in a 90-D feature vector.

The HIT−FA results in the two test conditions are shown in Table 7 and Table 8. As comparison, we also include results from AMS+RASTA-PLP+MFCC31, AMS+PITCHΔΔ, RASTA-PLP+PITCHΔΔ, MFCC31+PITCHΔΔ, GFCC31+PITCHΔΔ, and PITCHΔΔ. Com-



(a) Test condition one

(b) Test condition two

Figure 6: Overall HIT−FA rates of representative features with number of training utterances. "COMP" stands for AMS+RASTA-PLP+MFCC31.

paring with Table 1 and Table 2, we see that the complementary feature set AMS+RASTA-PLP+MFCC31 performs the best in test condition one, equaling GFCC31Δ. The final combined feature generalizes well to the unmatched-interference condition as shown in Table 8. For reference, the final combined feature using ground truth pitch achieves 84% and 76% HIT−FA rates in the two test conditions, respectively.

4.6 Training Corpus Size

As mentioned in Section 4.1, our training set is created from 50 clean utterances. In the following, we examine the dependence on the number of training utterances for each feature. We retrain SVM classifiers using 20, 100, and 200 utterances mixed with the same noises N1-N3, for representative features AMS+RASTA-PLP+MFCC31, AMS, GFCC31, and RASTA-PLP. The overall HIT−FA results are given in Fig. 6(a) and Fig. 6(b) for the two test conditions.

In the first condition, more utterances for training enable each feature 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 shows a 2% degradation from 100 to 200 utterances, indicative of overfitting. In the second condition, however, no significant performance gain is achieved beyond 50. For RASTA-PLP, a 5% gain is achieved by using 100 utterances compared to 20, but overfitting seems to occur when 200 utterances are used. It is worth noting that the performance of the combined feature using only 20 training utterances surpasses the other features using more training utterances. In summary, we do not observe strong dependence on the number of training utterances in condition two.

Table 9: Overall HIT−FA results in test condition one when tested on different SNR conditions

Feature	-5 dB			5 dB			10 dB		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
AMS	73%	12%	61%	65%	9%	56%	55%	15%	40%
RASTA-PLP	75%	7%	68%	71%	9%	62%	63%	8%	55%
GFCC31	92%	21%	71%	81%	11%	70%	75%	28%	47%
MFCC31	85%	9%	76%	80%	9%	71%	73%	10%	63%
AMS+RASTA-PLP +MFCC31	89%	9%	**80%**	84%	7%	**77%**	75%	8%	67%
AMS +PITCHΔΔ	72%	11%	61%	73%	12%	61%	65%	11%	54%
RASTA-PLP +PITCHΔΔ	74%	11%	63%	75%	10%	65%	66%	8%	58%
GFCC31 +PITCHΔΔ	83%	14%	69%	81%	12%	69%	74%	18%	56%
MFCC31 +PITCHΔΔ	80%	9%	71%	81%	11%	70%	73%	9%	64%
AMS+RASTA-PLPΔ +MFCC31+PITCHΔΔ	85%	8%	77%	85%	9%	76%	76%	8%	**68%**

Table 10: Overall HIT−FA results in test condition two when tested on different SNR conditions

Feature	-5 dB			5 dB			10 dB		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
AMS	58%	28%	30%	59%	20%	39%	53%	19%	34%
RASTA-PLP	67%	12%	55%	68%	13%	55%	60%	11%	49%
GFCC31	77%	41%	36%	73%	28%	45%	71%	32%	39%
MFCC31	71%	33%	38%	77%	27%	50%	60%	11%	49%
AMS+RASTA-PLP +MFCC31	78%	26%	52%	83%	21%	62%	76%	18%	58%
AMS +PITCHΔΔ	60%	11%	49%	68%	10%	58%	60%	9%	51%
RASTA-PLP +PITCHΔΔ	68%	11%	57%	74%	10%	64%	64%	8%	56%
GFCC31 +PITCHΔΔ	72%	25%	47%	77%	19%	58%	71%	21%	50%
MFCC31 +PITCHΔΔ	68%	13%	55%	77%	13%	64%	70%	12%	58%
AMS+RASTA-PLPΔ +MFCC31+PITCHΔΔ	75%	13%	**62%**	83%	12%	**71%**	74%	10%	**64%**

4.7 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 generalizes to different SNR conditions. To examine this question, we use the subband SVMs already trained on 0 dB mixtures described in Section 4.1 to segregate the same test mixtures at -5 dB, 5 dB, and 10 dB. Table 9 and Table 10 give the overall HIT−FA results for the two test conditions. All features are impacted in an unmatched-SNR condition. The reason for the performance degradation seems twofold. First, a change of SNR leads to a

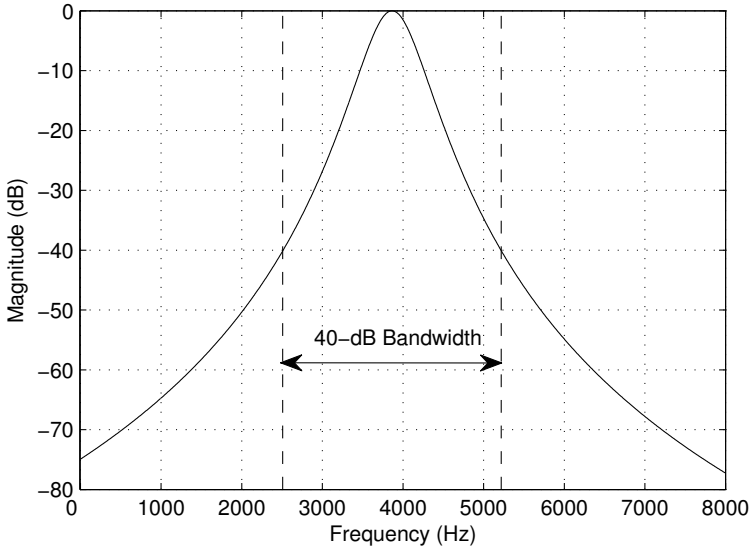


Figure 7: Magnitude response of a gammatone filter with center frequency $f_c = 3864$ Hz and the corresponding 40-dB bandwidth.

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 the HIT rate decreases significantly. Relatively speaking, MFCC31 and RASTA-PLP hold up well, especially at the lower SNR level. Again, the inclusion of the pitch-based feature clearly helps each feature to stabilize the labeling performance. The final combined feature significantly outperforms the other features in each SNR condition. When ground truth pitch is used, it achieves 86%, 81%, and 72% HIT−FA in test condition one, and 75%, 75%, and 68% in test condition two, at -5, 5 and 10 dB SNR respectively. These results are comparable to the matched-SNR scenario.

5 Dimension Reduction

Unit level features described in Section 2 contain redundancy since a subband signal contains information only within a certain bandwidth. The removal of redundancy could increase efficiency without hurting performance. We propose to reduce feature dimensionality based on bandwidth analysis of the gammatone filterbank. The impulse response function of a fourth-order gammatone filter with center frequency f_c is:

$$g_c(t) = t^3 \exp\left(-2\pi B_c t\right) \cos\left(2\pi f_c t + \phi\right), \; t > 0 \tag{4}$$

Table 11: HIT−FA results for reduced dimension GFCC feature in test condition one

Feature	Overall			Voiced			Unvoiced		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
GFCC31	86%	6%	80%	89%	7%	82%	76%	4%	72%
GFCC13	84%	6%	78%	87%	8%	79%	74%	4%	70%
GFCC_DR	83%	6%	77%	87%	7%	80%	69%	4%	65%

Table 12: HIT−FA results for reduced dimension GFCC feature in test condition two

Feature	Overall			Voiced			Unvoiced		
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA
GFCC31	74%	30%	44%	74%	29%	45%	74%	32%	42%
GFCC13	68%	31%	37%	68%	30%	38%	70%	31%	39%
GFCC_DR	75%	31%	44%	77%	30%	47%	67%	33%	34%

where ϕ is the phase, and B_c indicates bandwidth. A good approximation to its frequency response [32] is

$$G_c(f) = \left(1 + \frac{j(f - f_c)}{B_c}\right)^{-4}. \tag{5}$$

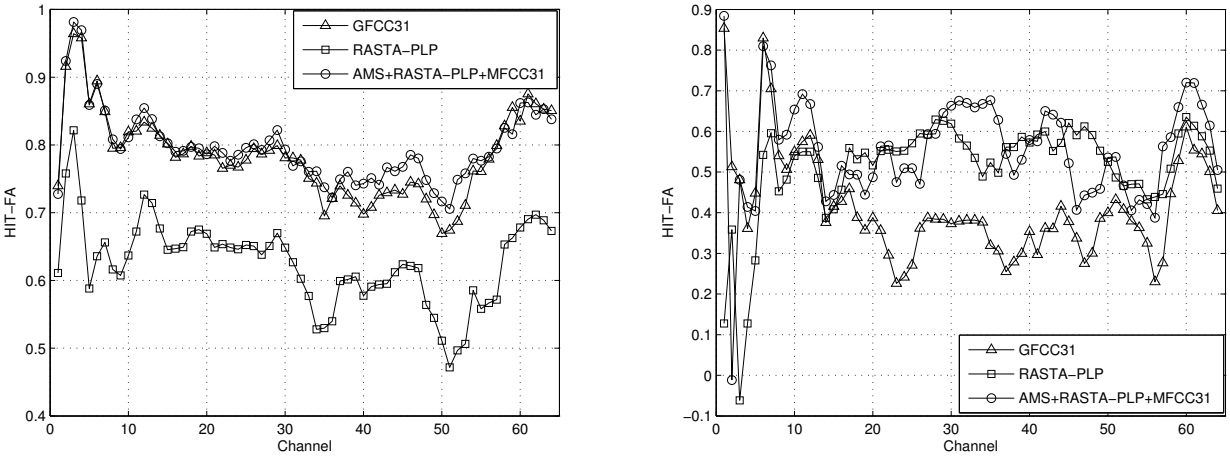
Based on its magnitude response $|G_c(f)|$, we can easily get the X-dB bandwidth as

$$\left[f_c - \sqrt{B_c^2(10^{X/40} - 1)}, f_c + \sqrt{B_c^2(10^{X/40} - 1)}\right], \tag{6}$$

which is defined as the frequency range between the lower and upper cutoff frequency at which the magnitude of the filter response is attenuated by X dB. Figure 7 illustrates the 40-dB bandwidth for a gammatone filter with center frequency $f_c = 3864$ Hz.

To extract unit level features, we first perform full spectral analysis on each subband signal, and retain only spectral contents within the X-dB bandwidth for further processing such as DCT. As an illustration, we apply this method to the GFCC feature, in which the 40-dB bandwidth is retained for decimation, cubic root compression and DCT (without truncation). This results in a new GFCC feature with an average dimension of 12[1], halving the original 31 dimensions. The HIT−FA results for this new feature, denoted by GFCC_DR, are listed in Table 11 and Table 12. Compared with the original GFCC31 feature, GFCC_DR achieves similar performance in both test conditions. In condition two, GFCC_DR which retains essential spectral information clearly outperforms GFCC13, a GFCC feature that reduces dimension by truncating higher DCT coefficients. Performance degradation is observed in unvoiced intervals. The reason is likely that the passband of a gammatone filter becomes wider at higher frequencies where unvoiced speech is prominent. Hence a fixed 40-dB bandwidth might not be sufficiently wide to discriminate unvoiced speech. A varied X could be used to alleviate this problem in the future. If we use a 60-dB bandwidth, for example, the average dimension increases to 23, but the HIT−FA rates become the same as GFCC31 in both test conditions.

[1]For a fixed X, the dimensionality of a feature varies in different channels with different widths of passbands.



(a) Test condition one

(b) Test condition two

Figure 8: HIT−FA rates of representative features in each frequency channel.

6 Discussion

Formulating monaural speech segregation as binary classification is an effective approach. So far, only pitch and AMS have been employed as T-F unit level features. In this paper, we have significantly expanded the 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 HIT−FA, a criterion that is highly correlated with human speech intelligibility. In terms of single features, GFCC shows excellent performance in the matched 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 important features contributing to target and interference discrimination. The selected complementary feature set AMS+RASTA-PLP+MFCC31 has shown stable performance in various test conditions and outperforms each of its components significantly.

We point out that HIT−FA results vary with respect to frequency channel, although classification is performed in a channel-independent way. The variations are shown in Fig. 8. In condition one, the performance clearly drops between channel 32 ($f_c = 1245$ Hz) and channel 58 ($f_c = 5732$ Hz). In condition two, the trend is less clear and it seems that combining outputs from classifiers trained on different features is helpful. Future effort is needed to understand the performance variation with respect to filter channel.

Generalization is a critical issue for classification-based speech segregation. We have examined the generalization performance of each feature in several unmatched conditions. These results point to the robustness of the pitch-based feature, which is parameterized by estimated pitch. Nevertheless, the pitch-based feature needs to be combined with general acoustic fea-

tures in order to segregate unvoiced speech and improve voiced speech segregation. The final combined feature AMS+RASTA-PLPΔ+MFCC31+PITCH$\Delta\Delta$ achieves very promising segregation results in various test conditions. These results are substantially better than those of earlier studies and are expected to further improve with better pitch tracking.

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 significantly 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.

Acknowledgement

This research was supported in part by an AFOSR grant (FA9550-08-1-0155) and an STTR grant from AFOSR. We thank Z. Jin for providing his pitch tracking code.

References

[1] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear and hearing*, vol. 27, no. 5, pp. 480–492, 2006.

[2] A.S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. The 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, Language Processing*, vol. 27, no. 2, pp. 113–120, 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," *Journal of the Acoustical Society of America*, 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, Language Processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[7] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[8] G. Garau and S. Renals, "Combining spectral representations for large-vocabulary continuous speech recognition," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 3, pp. 508–518, 2008.

[9] K. Han and D. Wang, "An SVM based classification approach to speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 5212–5215.

[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

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

[12] G. Hu, "Monaural speech organization and segregation," PhD Dissertation, The Ohio State University, Biophysics Program, 2006.

[13] G. Hu and D. Wang, "Segregation of unvoiced speech from nonspeech interference," *Journal of the Acoustical Society of America*, vol. 124, pp. 1306–1319, 2008.

[14] ——, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[15] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[16] G. Jang and T. Lee, "A maximum likelihood approach to single-channel source separation," *Journal of Machine Learning Research*, vol. 4, pp. 1365–1392, 2003.

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

[18] ——, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Language Processing*, vol. 19, pp. 1091–1102, 2011.

[19] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 126, pp. 1486–1494, 2009.

[20] N. Li and P. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.

[21] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009. [Online]. Available: http://www.public.asu.edu/~jye02/Software/SLEP

[22] L. Meier, S. V. D. Geer, and P. Bühlmann, "The group Lasso for logistic regression," *Journal of the Royal Statistical Society: Series B*, vol. 70, no. 1, pp. 53–71, 2008.

[23] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, pp. 2236–2252, 2003.

[24] S. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems*, 2001, pp. 793–799.

[25] M. Schmidt and R. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. 9th International Conference on Spoken Language Processing*, 2006.

[26] M. Seltzer, B. Raj, and R. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[27] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4625–4628.

[28] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 1589–1592.

[29] T. Takiguchi and Y. Ariki, "Robust feature extraction using kernel PCA," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[30] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[31] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Divenyi P., Ed. Kluwer Academic, Norwell MA., 2005, pp. 181–197.

[32] D. Wang and G. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.

[33] D. Wang, U. Kjems, M. Pedersen, J. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *Journal of the Acoustical Society of America*, vol. 125, pp. 2336–2347, 2009.

[34] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[35] A. Zolnay, D. Kocharov, R. Schlüter, and H. Ney, "Using multiple acoustic feature sets for speech recognition," *Speech Communication*, vol. 49, no. 6, pp. 514–525, 2007.