Towards Generalizing Classification Based Speech Separation

Kun Han, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Monaural speech separation is a well-recognized challenge. Recent studies utilize supervised classification methods to estimate the ideal binary mask (IBM) to address the problem. In a supervised learning framework, the issue of generalization to conditions different from those in training is very important. This paper presents methods that require only a small training corpus and can generalize to unseen conditions. The system utilizes support vector machines to learn classification cues and then employs a rethresholding technique to estimate the IBM. A distribution fitting method is used to generalize to unseen signal-to-noise ratio conditions and voice activity detection based adaptation is used to generalize to unseen noise conditions. Systematic evaluation and comparison show that the proposed approach produces high quality IBM estimates under unseen conditions.**

***Index Terms*—Generalization, rethresholding, speech separation, support vector machine (SVM).**

I. Introduction

SPEECH communication usually takes place in complex acoustic environments. The human auditory system is adept in separating target sound from background interference. Despite decades of effort, monaural speech separation remains one of the most difficult problems in speech processing. Various approaches have been proposed for monaural speech separation. Speech enhancement approaches [8], [11], [16] utilize the statistical properties of the signal to enhance speech that has been degraded by additive non-speech noise. Model based approaches [25], [26], [31], [43] use trained models to capture the characteristics of individual signals for separation. On the other hand, computational auditory scene analysis (CASA) [38] aims to separate a sound mixture based on perceptual principles [5].

For sound separation, the ideal binary mask (IBM) has been recently proposed as a main goal of CASA [37]. The IBM can be constructed from premixed target and interference. Specifically, with a time-frequency (T-F) representation of a sound mixture, the IBM is a binary T-F matrix where 1 indicates that the signal-to-noise ratio (SNR) within the corresponding T-F unit is greater than a local SNR criterion (LC) and 0 otherwise. In this work, we are concerned with monaural speech separation from nonspeech interference and we use -5 dB as LC in all experiments. A series of perceptual studies have shown that IBM separation produces large speech intelligibility improvements in noise for both normal-hearing and hearing-impaired listeners [1], [6], [21], [39].

Adopting the IBM as the computational goal, we can formulate sound separation as binary classification. Roman *et al.* [29] proposed an early supervised classification method for IBM estimation although the method used binaural features for speech separation. Several studies employ binary classification for IBM estimation in the monaural domain. Seltzer *et al.* [32] treated the identification of noise components in a spectrogram as a Bayesian classification problem for robust automatic speech recognition. Weiss and Ellis [42] utilized relevant vector machines to classify T-F units. Jin and Wang [17] trained multilayer perceptrons (MLP) to classify T-F units using pitch-based features. Their system obtains good separation results in reverberant conditions. Kim *et al.* [20] used Gaussian mixture models (GMM) to learn the distribution of amplitude modulation spectrum (AMS) features for target-dominant and interference-dominant units and then classified T-F units by Bayesian classification. Their classifier led to speech intelligibility improvements for normal-hearing listeners. Kim and Loizou [19] further proposed an incremental training procedure to improve speech intelligibility, which starts from a small initial model and updates the model parameters as more data become available. In [10], we proposed a support vector machine (SVM) based system and utilized both pitch-based and AMS features to classify T-F units. Comparisons show that the SVM system yields more accurate classification than the GMM classifier by Kim *et al.*

For supervised learning to be effective, the distribution of the training set needs to match that of the test set. For speech separation, if input SNRs or background noises in test mixtures are not seen in the training set, the trained classifier will unlikely achieve good classification results. Previous systems have avoided this issue by testing on SNR and noise conditions similar to those in training. Hence, it is important to investigate the generalization capability of such classifiers.

In this study, we propose an approach to estimate the IBM under unseen SNR or noise conditions. The proposed approach consists of an SVM training stage followed by a rethresholding step. SVM is a state-of-the-art learning machine which is widely used for classification problems [36]. SVM maximizes the margin of separation between different classes of training data, and as a result it shows good generalization. We utilize SVMs to produce initial classification boundaries and then

Manuscript received January 26, 2012; revised May 11, 2012 and August 15, 2012; accepted August 19, 2012. Date of publication August 28, 2012; date of current version October 23, 2012. This work was supported in part by an AFOSR grant (FA9550-08-1-0155) and an STTR subcontract from Kuzer. A preliminary and different version of this work was accepted for presentation in ICASSP 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Man-Wai Mak.

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

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

Digital Object Identifier 10.1109/TASL.2012.2215596

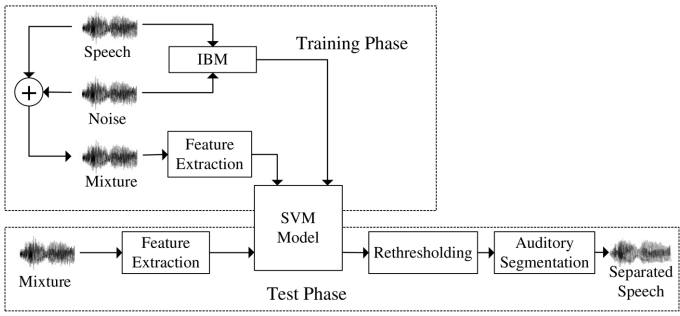

Fig. 1. Diagram of the proposed system.

derive new thresholds to classify T-F units in unseen acoustic environments. The new thresholds are adaptively computed based on the characteristics of test mixtures, and they are expected to generalize to new SNR or noise conditions. For unseen SNRs, by analyzing statistical properties, we determine the new thresholds by fitting the distribution of SVM outputs. For unseen noises, a voice activity detector is incorporated to construct a development set and then derive the thresholds.

Although our study treats the separation problem as binary masking, soft masking is also a common approach to separation (e.g. [41], [16]). The paper is organized as follows. In the next section, we present an overview of the proposed system. Sections III and IV describe how to generalize the SVM system to unseen SNR and noise conditions, respectively. Systematic evaluation and comparison are given in Section V. We discuss related issues and conclude the paper in Section VI.

II. System Overview

Fig. 1 shows the diagram of the proposed system, which consists of a training phase and a test phase. In the training phase, the speech and the noise are used to create the IBM, which provides the desired output for training. The features in each T-F unit are extracted from the mixture and then used to train an SVM model in each frequency channel. In the test phase, we first use the trained SVM to initially classify T-F units, and then utilize a rethresholding technique to generalize the system under different test conditions. Auditory segmentation is used to improve the estimated mask and separated speech is finally resynthesized by using the estimated IBM.

A. Feature Extraction

An input mixture $s(t)$ is first fed into a 64-channel gammatone filterbank whose center frequencies are distributed from 50 Hz to 8000 Hz [38]. This filterbank is derived from psychophysical studies of auditory periphery and is a standard model of cochlear filtering [27]. In each channel, the output is windowed into 20-ms time frames with 10-ms frame shift. This processing decomposes the input signal into a two-dimensional T-F representation called a *cochleagram* [38]. We use $u_{c,m}$ to denote a T-F unit in the cochleagram, which corresponds to frequency channel c and time frame m.

Given the cochleagram of the mixture, we extract acoustic features from each T-F unit. In [10], a combination of pitch-based features and amplitude modulation spectrum (AMS) features [35] is used to effectively classify T-F units under the noise matched condition. For SNR generalization, since we only consider the matched noises, it is reasonable to adopt the same combined features into our system.

For noise generalization, AMS features may not be an appropriate choice, because they do not show good performance under unseen noise conditions [10], [40]. According to a recent comparison of features [40], we use relative spectral transform-perceptual linear prediction (RASTA-PLP) features [12] to perform classification under unseen noise. With the pitch based features, the combined features are expected to perform good discriminative capacity on various noises.

Delta features are found to be helpful in speech separation as they encode feature variations [10], [20]. We concatenate the original features with their time delta features and frequency delta features into a combined feature vector for classification. In Section V, we discuss feature extraction in details.

B. SVM and Rethresholding

With the extracted features, we use SVM to classify T-F units to target-dominant or interference-dominant classes. We use probabilistic SVMs to model the posterior probability that a T-F unit label Y is assigned 1 given the feature vector, denoted as $P(Y = 1|\mathbf{x})$. A separate SVM is trained for each frequency channel because the characteristics of the speech signal in different channels can be very different. In the training phase, we use the radial basis function kernel, $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ and the parameters are chosen by 5-fold cross-validation. To obtain a probabilistic representation, we use a sigmoid function to map an SVM decision value to a number between 0 and 1, which is then interpreted as the posterior probability of the target [28]. With this compact representation, one can derive new thresholds within [0,1] instead of $[-\infty, \infty]$. The SVM library LIBSVM [7] is used in our experiments.

In the test phase, the decision value for each T-F unit is calculated from the discriminant function as follow:

$$f(\mathbf{x}) = \sum_{i \in \mathrm{SV}} a_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{1}$$

where SV denotes the set of support vector indices in training data and y_i is the label corresponding to $\mathbf{x}_i$. a_i is a Lagrange multiplier and b is the bias, both of which can be determined in the training phase. The decision value $f(\mathbf{x})$ is a real number between $(-\infty, +\infty)$, which is then mapped to a number within [0, 1] representing the posterior probability of the unit being target-dominant [28]:

$$P(Y = 1|\mathbf{x}) = \frac{1}{1 + \exp\left(\alpha f(\mathbf{x}) + \beta\right)} \tag{2}$$

where the parameters α and β denote the shape of the sigmoid function, which are fixed in the training phase. Note that (2) is a monotonic bijective function but the original threshold $f(\mathbf{x}) = 0$ does not necessarily correspond to $P(Y = 1|\mathbf{x}) = 0.5$.

Generally speaking, standard probabilistic SVMs use $P = 0.5$ as the threshold to perform classification. In this study we train with a fixed input SNR or using a small number of noise types and wish to generalize to a variety of unseen conditions. In this case, we do not expect the trained SVMs to produce good classification results in unseen conditions.

In [10], we proposed a rethresholding technique to improve SVM classification results, which has been successfully used for text classification [4], [34]. One reason for the use of rethresholding is that there exists a mismatch between the training set

and the test set. Under unmatched conditions, the optimal hyperplane for the training set likely deviates from the optimal hyperplane for the test set. In this study, we propose to use rethresholding to adjust trained hyperplanes in order to generalize to unseen SNR or noise conditions.

Specifically, we first need to find a channel-specific threshold θ_c that maximizes the classification accuracy in channel c, and then use the new threshold to binarize $P(Y=1|\mathbf{x})$:

$$Y = \begin{cases} 1, & \text{if } P(Y=1|\mathbf{x}) > \theta_c \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Our experiments show that with properly chosen thresholds, the system can significantly improve classification. For SNR generalization, the system is trained on 0 dB. Therefore, the key problem is how to determine the new threshold θ_c for each channel c. In [10], a small validation set is used to determine the thresholds. But this strategy cannot be used in this study, because the statistical properties of the test set are very different from those of the training set and unknown. Thus, we need to develop new strategies for rethresholding under unseen SNR or noise conditions, which are described in detail in Section III and Section IV, respectively.

C. Auditory Segmentation

The rethresholded mask gives a good estimate of the IBM, but it still misses some target-dominant units and contains some interference-dominant units. To further improve the rethresholded mask we utilize auditory segmentation which takes into consideration contextual information beyond individual T-F units. The goal of auditory segmentation is to form contiguous T-F segments and each segment is supposed to primarily originate from the same sound source [38]. In this study, we adopt the same segmentation method as in [10], [17]. Specifically, for voiced intervals, we compute cross-channel correlation of filter response in low frequency channels (below 2000 Hz) and cross-channel correlation of envelope response in high frequency channels (above 2000 Hz).

To perform segmentation, only those units with sufficiently high ($\geq$0.95) cross-channel or envelope cross-channel correlation are selected. Selected neighboring units are iteratively merged into segments across frequency channels and time frames. For unvoiced intervals, segments are formed by matching pairs of onset and offset fronts and a multiscale analysis is applied to integrate segments at several scales [15]. With the rethresholded mask, we label all units in a segment as the target if the energy included in the target-dominant units is greater than the energy included in the interference-dominant units. Then, we treat all the segments shorter than 50 ms as the interference and obtain the final estimated IBM.

To summarize, given a noisy speech signal, we first extract features in the T-F domain and then use SVM to produce initial classification for each T-F unit. Then, we use rethretholding to adapt SVM output under different conditions. Finally, auditory segmentation is used to improve the estimated IBM. The following sections describe how to apply rethresholding under different conditions.

III. Generalization to Different Input SNRs

For SNR generalization, the training set contains mixtures at a single input SNR and the system will be tested on mixtures at different input SNRs. In this case, if we directly use $\theta = 0.5$ as the threshold, the system does not generalize well to unseen SNRs. We refer to the threshold that maximizes some classification accuracy as the optimal threshold. We observe that in unmatched SNR conditions, the use of the optimal threshold in each channel can substantially improve the classification result relative to the default threshold of 0.5. In other words, if we can find thresholds close to the optimal one, the system is expected to generalize well under unseen SNR conditions.

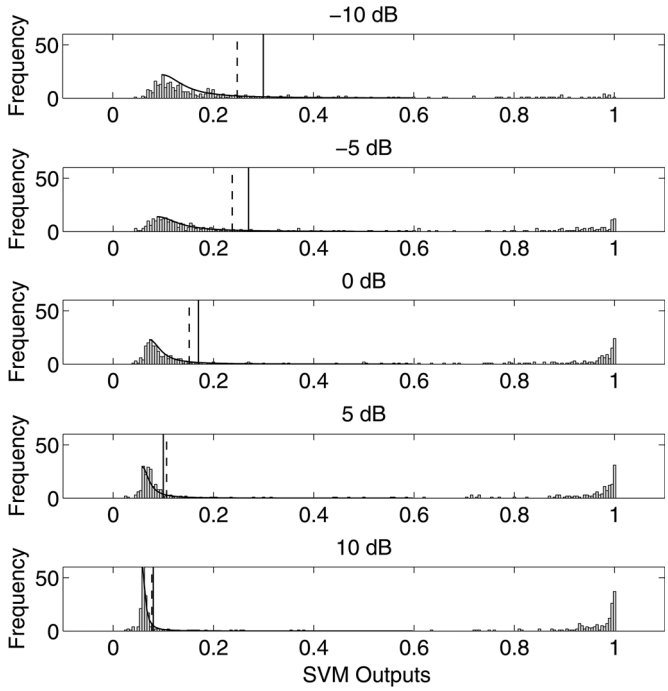

Fig. 2. Histograms of the SVM outputs in the 18th channel with different input SNRs. The solid curve denotes the half-Cauchy distribution used to fit the SVM outputs. A vertical line indicates the optimal threshold and a dashed vertical line the estimated optimal threshold by using distribution fitting.

Furthermore, we observe that, although the optimal threshold varies in different SNR conditions, SVM outputs have similar distribution shapes and the optimal thresholds are located at similar positions relative to the distribution shapes. As a typical example, Fig. 2 shows the histograms of the SVM outputs in the 18th channel for a female utterance, "A man in a blue sweater sat at the desk," from the IEEE corpus [30] mixed with speech-shaped noise. The system is trained on 100 IEEE sentences mixed with speech-shaped noise, factory noise and babble noise at 0 dB and SVM outputs are generated at -10, -5, 0, 5 and 10 dB input SNRs. The figure shows that there exists a peak K on the left side ($P < 0.6$) of each histogram. Also, the SVM ouputs on the left side for different SNRs have similar distribution shapes which gradually become sharper as the input SNR increases. Further, the optimal threshold θ shown as the solid vertical line in each histogram is increasingly close to the peak K as the distribution becomes sharper. If we only consider the SVM outputs on the left side of the histogram, the optimal threshold always occurs at the tail end of the distribution under each input SNR condition. This motivates us to use the same distribution function to fit SVM outputs at different SNRs with different parameter values.

One can perform distribution fitting in two ways: fit all SVM outputs less than 0.6 or SVM outputs between K and 0.6. We

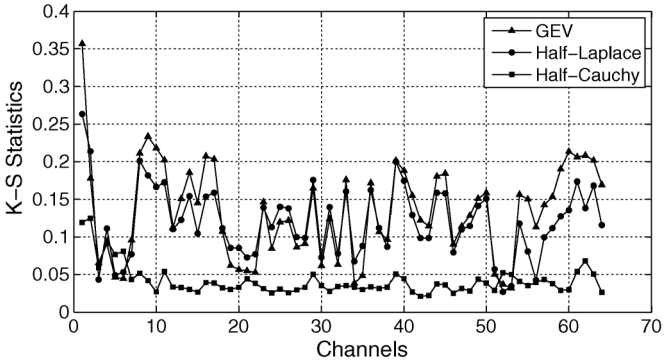

Fig. 3. Kolmogorov-Smirnov statistics for three distributions.

have explored several reasonable distributions as the candidates to fit SVM outputs and use the Kolmogorov-Smirnov (K-S) statistics [3] to test the goodness of fit. Three distributions are tested: the generalized extreme value distribution (GEV) is used to fit all SVM outputs less than 0.6, whereas the half-Cauchy and the half-Laplace distributions are used to fit the SVM outputs within $[K, 0.6]$. The probability density functions are:

$$\begin{aligned}\text{GEV}: \ & f(x;\mu,\sigma,\xi) \\ &= \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{\frac{-1-\xi}{\xi}} \\ &\quad \times \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \qquad (4)\end{aligned}$$

$$\begin{aligned}\text{Half}-\text{Cauchy}: \ & f(x;\mu,\sigma) \\ &= \begin{cases} \frac{2}{\pi\sigma\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]}, & \text{if } x \geq \mu \\ 0, & \text{otherwise} \end{cases} \qquad (5)\end{aligned}$$

$$\begin{aligned}\text{Half}-\text{Laplace}: \ & f(x;\mu,\sigma) \\ &= \begin{cases} \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right), & \text{if } x \geq \mu \\ 0, & \text{otherwise} \end{cases} \qquad (6)\end{aligned}$$

where, μ, σ, ξ are parameters determined by maximal likelihood estimation.

Fig. 3 shows the K-S statistic test results in each channel, averaging over 10 IEEE sentences mixed with the three noises at five SNR levels. From the figure, all three distributions achieve relatively low K-S statistics, meaning that the candidate distributions fit the data well. The best one is the half-Cauchy distribution which has the lowest K-S statistics in most channels. Consequently, we use the half-Cauchy distribution in our method. As shown in Fig. 2, under each SNR condition, the solid curve denotes the probability density function of a half-Cauchy distribution, which well fits the SVM outputs between K and 0.6.

Therefore, given SVM outputs in one channel, we estimate parameters of a half-Cauchy distribution to fit the outputs by maximal likelihood estimation. Based on the fitted distribution function F and the observation that the optimal threshold θ is located at the tail end of the distribution, the corresponding cumulative probability $\rho = F(\theta)$ should be close to 1. This turns optimal threshold estimation to another problem: given F with unknown parameters $\Omega = \{\mu, \sigma\}$ and a predetermined cumulative probability ρ, we can first estimate Ω based on SVM outputs and then approximate the optimal threshold by calculating the inverse cumulative distribution function $\theta = F^{-1}(\rho; \Omega)$. Here, ρ is set to 0.9, which is chosen from a validation set. Incidentally, we choose 0.6 instead of 0.5 as the upper bound of the SVM outputs to fit the distribution because we want to include more samples for the fitting. The number of SVM outputs less than 0.6 is very unlikely too small (i.e., less than 5% of the total SVM outputs) to well fit a distribution function, because human speech contains pauses which should produce sufficient interference-dominant units (i.e., SVM outputs less than 0.6) in the mixture. In the case that few SVM outputs can be used to fit the distribution, we simply set the threshold to the original value 0.5. Note that, although we only use those SVM outputs less than 0.6 to fit a distribution function, it does not mean that the estimated optimal threshold is less than 0.6. The threshold only depends on the parameters of the fitted distribution function and can be any value in [0, 1].

To summarize, we use the following algorithm to estimate the optimal threshold θ in each channel:

1) Given the SVM outputs, we uniformly divide [0, 1] into 100 bins and derive the histogram of SVM outputs. For those bins less than 0.6, we choose the bin with the highest frequency as the peak K.
2) We use the half-Cauchy distribution F with unknown parameters Ω to fit the SVM outputs within $[K, 0.6]$ and use maximal likelihood to estimate Ω;
3) We estimate the optimal threshold using inverse cumulative distribution function $\theta = F^{-1}(\rho; \Omega)$.

The dashed line shown for each histogram in Fig. 2 denotes the estimated optimal threshold based on distribution fitting, which is close to the optimal threshold. Finally, we use the threshold calculated from the algorithm to binarize the SVM outputs in each channel and obtain a rethresholded mask. This mask is further improved by an auditory segmentation procedure and form an estimated IBM. It is worth emphasizing that this method estimates optimal thresholds only based on SVM outputs of the mixture without the knowledge of the input SNR.

IV. Generalization to Different Noises

Another important issue is generalization to unseen noises. We also use rethresholding to generalize the system to unseen noises as we have observed that optimal thresholds significantly improve the classification results.

Although distribution fitting is able to generalize the trained models to unseen SNR conditions, it does not work well for unseen noise conditions because the characteristics of noises can be very different and no pattern of the histograms appears to fit all noises. Fig. 4 shows histograms of SVM outputs in the 18th channel corresponding to four female utterances mixed at 0 dB with (a) speech-shaped noise and (b) rock music. Both noises are not seen in the trained SVM model (see Section V-B for more details). The solid vertical line indicates the optimal threshold in each histogram. We can compare the histograms in (a) and (b) in each row. Although the same sentence is used to generate the SVM outputs, they have very different distributions as the noises are different. On the other hand, for those mixtures with the same type of noise, the optimal thresholds have close values: around 0.5 for speech-shaped noise and 0.8 for rock music. The histograms for speech-shaped noise in this figure are quite different from those in Fig. 2 for two reasons: (1) speech-shaped noise is contained in the training set in Fig. 2 but not in the training set in Fig. 4; (2) the system in Fig. 2 uses

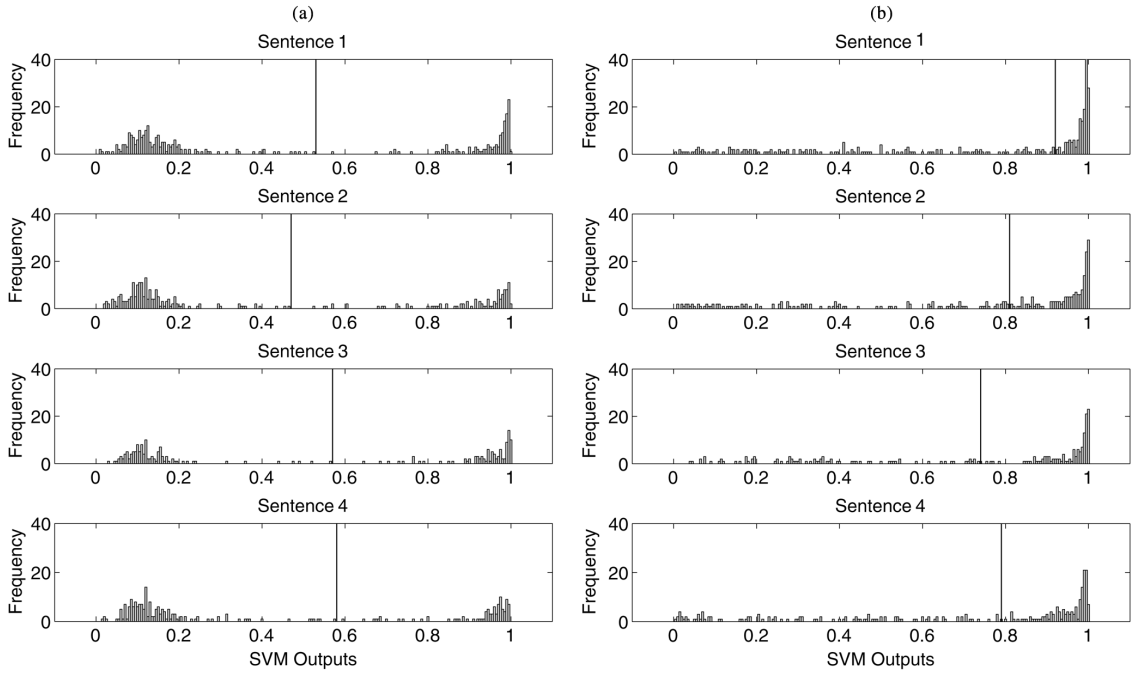

Fig. 4. Histograms of the SVM outputs in the 18th channel. Four different utterances mixed with (a) speech-shaped noise and (b) rock music. The solid vertical line in each panel denotes the optimal threshold.

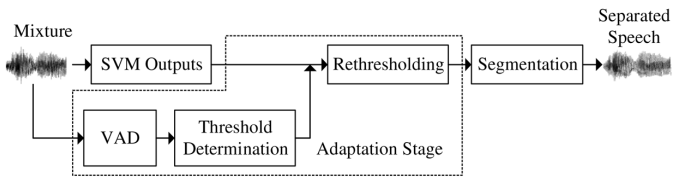

Fig. 5. Diagram of VAD based rethresholding for generalization to unseen noises.

AMS and pitch-based features for SNR generalization but the system in Fig. 4 uses RASTA-PLP and pitch-based features for noise generalization.

The above analysis suggests that, if mixtures come from the same kind of noise, it is reasonable to apply the same threshold to all these mixtures in each channel. In other words, although it is impossible to directly obtain the optimal thresholds for a test mixture as the IBM is not accessible, if we can somehow access part of the noise, we can use the noise part to construct a development set including a reference mixture and the corresponding IBM to calculate the optimal thresholds. The optimal thresholds obtained from the development set are expected to perform well on the test mixture because the same type of noise is used in both mixtures. Obviously, to construct the development set clean speech is needed, which can be an arbitrary utterance. We randomly choose a single utterance, "Shake the dust from your shoes, stranger," from the IEEE corpus and use this one to construct the development sets for all test mixtures.

To obtain noise portions from a test mixture, we propose to apply voice activity detection (VAD) in an adaptation stage to perform rethresholding. VAD is used to identify noise-only frames which are then mixed with the above clean speech to construct a development set. The thresholds chosen from the development set are used to produce a binary mask. Fig. 5 illustrates the computational flow.

As shown in the figure, given a test mixture, we use the trained SVMs to output the posterior probability of speech dominance for each T-F unit. In parallel, we use Sohn *et al.*'s VAD algorithm [33] to detect noise frames. This standard VAD algorithm uses a statistical model-based method to produce the likelihood of speech presence for each frame. In our corpus, speech pause accounts for around 30% of frames, so we select 30% of the frames with the lowest likelihoods as the candidates of noise frames. To avoid spurious noise frames caused by VAD errors, we further use detected pitch in the feature extraction stage in Section II-A to improve the VAD results: a candidate of noise frame is removed if a pitch is detected in this frame. In addition, since very short noise sections are not useful for constructing a development set, we exclude those noise sections whose lengths are shorter than 50 ms (or 5 frames).

With detected noise frames and the one clean utterance, we mix them into a reference mixture. This mixing, however, requires that the noise frames and the clean utterance have the same length. In this study, both the test mixture and the clean utterance last around 2 seconds, and as a result the total length of detected noise frames is usually significantly shorter than the length of the clean utterance. To match the utterance length, we first concatenate detected noise frames to a noise section and then repeatedly duplicate the noise section until the total length is equal to that of the utterance. The resulting noise section and the clean utterance are used to construct a development set. We find that, although a longer test mixture (>10 seconds) can provide more noise frames without duplication, it does not give better results than a 2-second mixture.

After we construct a development set containing a single mixture, we calculate optimal thresholds θ based on the reference mixture and its IBM. That is, we apply our trained models to the reference mixture to calculate SVM outputs and use the corresponding IBM to choose the optimal threshold θ in terms of accuracy in each frequency channel. With the obtained θ and SVM outputs of the test mixture, it is straightforward to use (3) to produce a rethresholded mask. Finally, we employ a segmentation step to further improve IBM estimation.

V. Evaluation and Comparison

A. Generalization Results for Unseen SNRs

We first evaluate the capacity of our system to generalize to unseen SNRs. As we mentioned above, we utilize pitch-based features, AMS features and their delta features for SNR generalization. For pitch-based features, we calculate the normalized autocorrelation function $A(c, m, \tau)$ at pitch period $\tau_S(m)$. For voiced speech, $A(c, m, \tau_S(m))$ measures how well the unit response is consistent with the target pitch, which has been proven to be an effective feature for speech separation [14], [17]. To remove the influence of pitch errors in the training phase, we use *Praat* [2] to extract the ground-truth pitch from the premixed speech in the training phase, and use the pitch tracker in [18] to extract the estimated pitch from the mixture in the test phase. Similarly, we also compute autocorrelation from the envelope of the response to obtain $A_E(c, m, \tau_S(m))$ as a feature to capture amplitude modulation information.

We calculate delta features in the following manner: in the time dimension, for $m \geq 2$, the time delta feature $\Delta A^M(c, m, \tau_S(m)) = A(c, m, \tau_S(m)) - A(c, m-1, \tau_S(m))$; $\Delta A^M(c, 1, \tau_S(m))$ is simply set to $\Delta A^M(c, 2, \tau_S(m))$ for convenience. We compute the frequency delta feature $\Delta A^C(c, m, \tau_S(m))$ in the same way. Therefore, between response and envelope autocorrelation we get a 6-dimensional pitch-based features $\mathbf{x}_P$.

We use the same method as in [10] to extract AMS features. Specifically, the envelope from the filter response within each T-F unit is extracted. The envelope is Hanning windowed and zero-padded for a 256-point FFT. The resulting FFT magnitudes are integrated by 15 triangular windows, generating a 15-dimensional AMS feature. Similarly, we calculate delta features across time frames and frequency channels. In each T-F unit, the pitch-based feature vector $\mathbf{x}_P$ and the AMS feature vector $\mathbf{x}_A$ are combined into a feature vector and used for the classification under different SNR conditions.

The features are extracted from the IEEE corpus [30]. Similar to Kim *et al.* [20], the training set consists of 100 female utterances mixed with three types of noise: speech-shape noise, factory noise and babble noise at 0 dB. For the test set, we choose 10 new utterances mixed with the same three types of noise at -10, -5, 0, 5 and 10 dB.

In order to quantify the performance of our system, we compute the HIT rate which is the percent of the target-dominant units in the IBM correctly classified, and the false-alarm (FA) rate which is the percent of the interference-dominant units in the IBM wrongly classified. We use the difference between HIT and FA, HIT-FA, as an evaluation criterion since it has been shown to be correlated to human speech intelligibility [20], [21] and has been adopted in earlier studies [10], [20].

Fig. 6 shows the average HIT-FA results over the three noises under each input SNR condition. The triangle line indicates the original HIT-FA rates without rethresholding. With the optimal thresholds, the HIT-FA rates are boosted by 10% absolute on average, which clearly shows the advantage of rethresholding. By using distribution fitting based rethresholding, we improve the HIT-FA results by 9% for low input SNR conditions (-10 and -5 dB) and 10% for high SNR conditions (5 and 10 dB). The result in the matched SNR condition is also improved, probably because the ground-truth pitch is used in the training phase but the estimated pitch is used in the test phase. This pitch discrepancy would lead to an optimal threshold different from the original threshold 0.5 (as shown in Fig. 2), so the HIT-FA rate could be improved by rethresholding even under matched SNR conditions. No segmentation is used in this comparison. It is interesting to note that, the distribution fitting based rethresholding outperforms the optimal rethresholding under the -10 dB condition. This is because the optimal threshold is chosen to maximize the accuracy in each channel, which does not necessarily maximize the corresponding HIT-FA rate for the whole mask (see [10]).

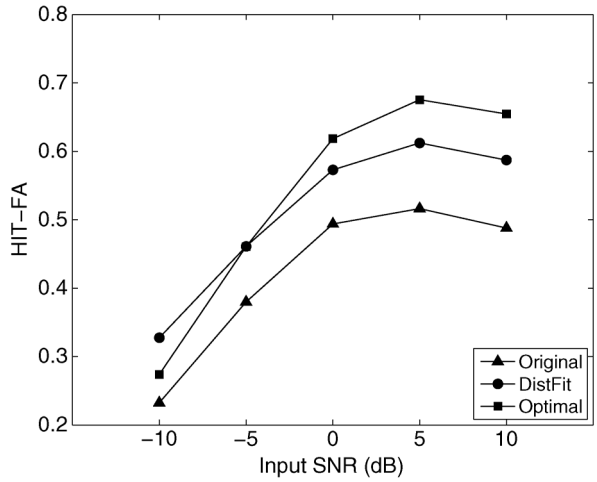

Fig. 6. Distribution fitting based SNR generalization results in terms of HIT-FA. The line with triangles denotes the original SVM results, the line with circles the distribution fitting based rethresholding results, and the line with squares the results using optimal thresholds.

The above results show the advantage of rethresholding in our system. We now compare our system with three recent speech separation systems. The first one is an IBM estimation system proposed by Kim *et al.* [20]. As mentioned in Section I, this system extracts AMS features and utilizes GMM classifiers to estimate the IBM, and it has been demonstrated to improve speech intelligibility in human listening tests. Their system is trained on the same 100 utterances mixed with the same three noises, but three SNR levels of at -5, 0 and 5 dB SNR as reported in [20]. We train a 256-component GMM for each class in each channel. The second one is a state-of-art speech enhancement system based on noise tracking proposed by Hendriks *et al.* [11]. This system assumes that both the speech and noise DFT coefficients have a complex-Gaussian distribution and utilizes a minimum mean-squared error (MMSE) estimator of the noise magnitude-squared DFT coefficients to estimate noise power spectral density. The clean speech DFT coefficients are estimated from a magnitude-DFT MMSE estimator presented in [9]. With these estimates, one can calculate the speech and noise energy within a time-frequency unit in the linear DFT domain. Since our IBM is defined in the gammatone filterbank domain, we need to convert the speech and noise energy in the linear DFT domain to the corresponding energy estimates in the gammatone filterbank domain [23]. Without loss of generality, we

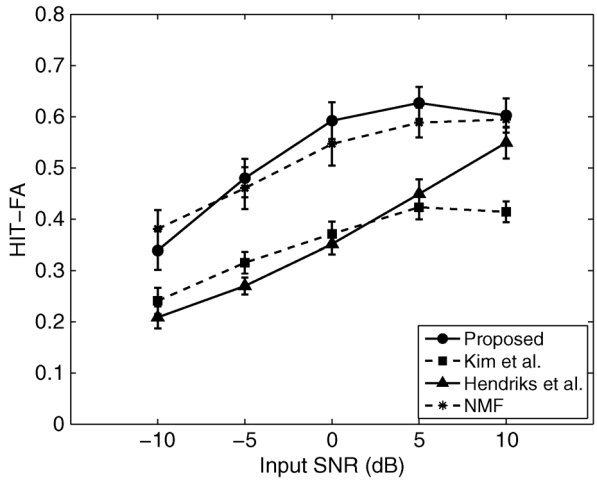

Fig. 7. HIT-FA rates with respect to input SNR levels. The error bars indicate 95% confidence intervals of the means.

consider the energy E of a T-F unit $u_{c,m}$ in the gammatone filterbank domain:

$$E(c,m) = \sum_n |y_c[n]|^2 = \frac{1}{K}\sum_{k=0}^{K-1} |Y_c[k]|^2$$
$$= \frac{1}{K}\sum_{k=0}^{K-1} |X[k]|^2 \cdot |G_c(k)|^2 \quad (7)$$

where $y_c[n]$ denotes a filtered time domain signal in frequency channel c and frame m, and $Y_c[k]$ are the DFT coefficients of $y_c[n]$, where K is set to 512 in our experiments. The second equation is due to Parseval's theorem [24]. G_c is the frequency response function of the gammatone filter in channel c. $X[k]$ is a DFT coefficient of the original signal, which can be estimated by Hendriks *et al.*'s system. For each T-F unit in the gammatone filterbank domain, we use (7) to calculate the speech and noise energy respectively, and then compute the local SNR to generate the binary mask. The third method is a model-based system using a general framework proposed by Ozerov *et al.* [26]. This method utilizes nonnegative matrix factorization (NMF) to perform separation. We use 10 IEEE sentences to train a 64-component speaker NMF model and the same three noises to train a 16-component noise NMF models. Since the NMF-based method produces the separated speech signal and noise signal in the time domain directly, we decompose these two signals to the T-F domain and calculate local SNRs to form a binary mask for comparison.

As shown in Fig. 7, the proposed system slightly outperforms the NMF-based method (by around 4% on average) in terms of HIT-FA rates. The other two systems perform considerably worse. To indicate statistical significance, we also show 95% confidence intervals in the figure, which are calculated from a normal distribution fitted by obtained results. Note that, Kim *et al.*'s system is trained on -5, 0 and 5 dB input SNRs, and it is supposed to achieve good performance at the three trained input SNRs.

We should point out that Hendriks *et al.*'s system is not designed to estimate the IBM. We have also implemented a binary masking system proposed by Jensen and Hendriks [16] for comparison. Their system derives a gain function based on the same spectral magnitude MMSE as in Hendriks *et al.* but generates an optimal binary mask in the MMSE sense, which is a binarization based on gain thresholds. We first calculate a gain threshold for each T-F unit and convert it to an energy threshold in the DFT domain. Eq. (7) is then used to calculate the corresponding energy threshold for each T-F unit in the gammatone filterbank domain. With the cochleagram of the mixture and the calculated energy thresholds, we can generate an optimal binary mask in the gammatone filterbank domain. However, their system achieves lower HIT-FA rates than the one based on Hendriks *et al.* described above. One important reason is that Jensen and Hendriks aim to obtain the optimal binary mask in the MMSE sense rather than the ideal binary mask used in our study. This suggests that there are differences between an optimal-binary-mask estimator and an ideal-binary-mask estimator. Even though Jensen and Hendriks [16] reported that estimated optimal binary masks do not lead to significant improvements of speech intelligibility, the same cannot be said of estimated IBMs [20].

The comparisons above focus on unit classification accuracy, where we need to convert the energy estimates from Hendriks *et al.* in the DFT domain and the separated signals from the NMF-based method in the time domain to the gammatone filterbank domain. To eliminate the effects of conversion, we use inverse FFT to resynthesize estimated speech energy in the DFT domain to the waveform. We also resynthesize from the estimated IBMs of Kim *et al.* and the proposed system to waveform [38]. With the resynthesized signal, we measure the output SNR of the separated speech as follows [14]:

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_n s_I^2(n)}{\sum_n \left[s_I(n) - s_E(n)\right]^2} \quad (8)$$

For Kim *et al.* and the proposed system, $s_I(n)$ and $s_E(n)$ indicate the signals resynthesized using the IBM and the estimated IBM, respectively. For Hendriks *et al.*'s system, $s_I(n)$ and $s_E(n)$ indicate the clean speech and the signal resynthesized using the estimated speech energy, respectively. For the NMF-based method, $s_I(n)$ and $s_E(n)$ indicate the clean speech and the separated speech signal, respectively. To quantitatively evaluate the performance, an SNR gain is computed by subtracting the output SNR of separated speech by the input SNR before separation. Fig. 8 shows the SNR gains. The proposed system achieves considerable SNR gains at all input SNRs. Although the SNR gains of all systems decrease gradually as the input SNR increases, the other three systems have more significant degradation at higher input SNRs.

B. Generalization Results for Unseen Noises

We utilize pitch-based features, RASTA-PLP features and their delta features for SNR generalization. To get RASTA-PLP features, after the power spectrum is warped to the Bark scale, we log-compress the resulting auditory spectrum, filter it by the RASTA filter, and expand it by an exponential function. Subsequently, PLP analysis is taken on this filtered spectrum. The original RASTA-PLP feature is a 13-dimensional vector and we also calculate the delta features for RASTA-PLP across time frames and frequency channels to generate a 39-dimensional RASTA-PLP feature vector $\mathbf{x}_R$. The pitch-based feature vector $\mathbf{x}_P$ and the RASTA-PLP feature vector $\mathbf{x}_R$ are finally combined

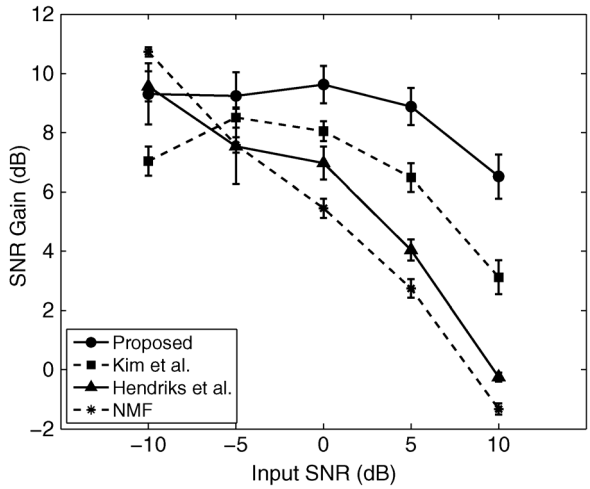

Fig. 8. SNR gains with respect to input SNR levels. The error bars indicate 95% confidence intervals of the means.

and a 45-dimensional feature vector for each T-F unit is used as the input to the classifier for noise generalization.

To evaluate generalization to unseen noises, we choose 30 female utterances from the IEEE corpus mixed with 5 types of noise out of a 100 nonspeech noise set [13] at 0 dB SNR to train the system. To construct a representative training set, we use a clustering based noise selection scheme to choose training noises. Intuitively, we want to include the most diverse noises as the training set, i.e., the distribution of features extracted from the training noises should cover the feature space as much as possible. For noise selection, we only consider RASTA-PLP features since pitch-based features do not exist in unvoiced speech. We first pass each noise waveform through a gammatone filterbank and then extract RASTA-PLP features from each T-F unit. Then, the mean of the RASTA-PLP features is calculated over all units for each type of noise. Therefore, each noise is represented by a 13-dimensional feature vector. We then apply the K-means ($\mathrm{K} = 5$ in this experiment) clustering to these 100 feature vectors and thus 100 noises are divided into 5 clusters. For each cluster, we select one noise that has the shortest distance to the cluster center as the representative. Therefore, 5 representative noises are used in the training set. Compared with random noise selection, this clustering-based noise selection produces 3% improvement in terms of HIT-FA.

To test our system, we use 10 new female utterances mixed with the 10 types of noise—N1: speech-shape noise, N2: factory noise, N3: fan noise, N4: bird chirp, N5: white noise, N6: cocktail party noise, N7: rain noise, N8: rock music, N9: wind noise, N10: clock alarm—at 0 dB. The test noises cover both stationary and nonstationary noises and have very different frequency characteristics, and none of them are in the training set.

Fig. 9 shows the HIT-FA results of the proposed system. For each noise, the left two bars show the original SVM results using a threshold of 0.5 and the rethresholding results using the optimal thresholds, respectively. The figure shows that the optimal rethresholding substantially improves HIT-FA and achieves an average improvement of 7.3%, which suggests the utility of rethresholding for generalization. The VAD based rethresholding improves HIT-FA rates under all unseen noise conditions and the average improvement is 5.9%. With

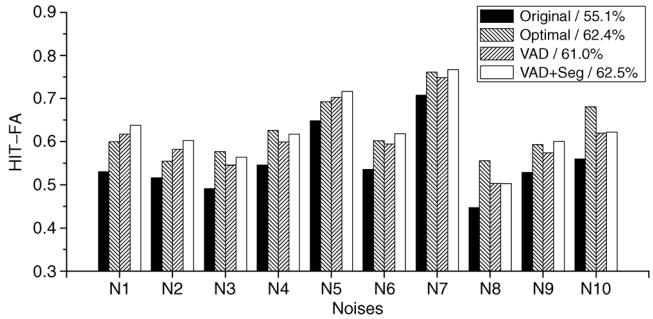

Fig. 9. Noise generalization in terms of HIT-FA. "Original" denotes the original SVM results without rethresholding, "Optimal" the rethresholding results using optimal thresholds, and VAD denotes the VAD based rethresholding results. VAD+Seg denotes the results using VAD based rethresholding followed by segmentation.

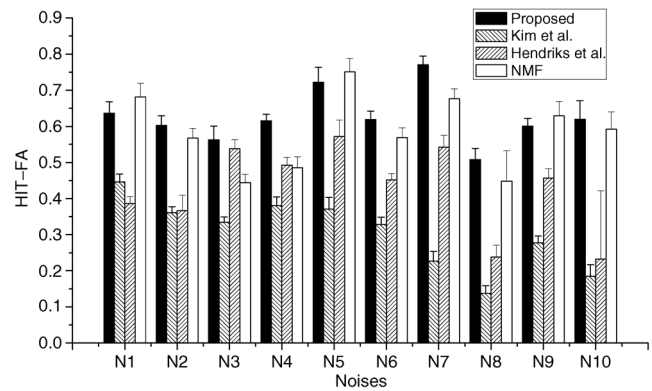

Fig. 10. Noise generalization comparisons in terms of HIT-FA. The proposed method denotes VAD based rethresholding followed by segmentation. The error bars indicate two-side 95% confidence intervals of the means, with only one side shown for clarity.

segmentation, the proposed system further improves IBM estimation, and it outperforms the original one by 7.4% making it comparable to the optimal rethresholding results. These results demonstrate that, with a little adaptation, our system generalizes well to different noise conditions.

Since our system utilizes nonspeech intervals detected by the VAD algorithm to adapt the thresholds, we also adopt a similar strategy in Ozerov *et al.* [25] for the model-based system where the noise model is adapted by the detected nonspeech intervals. In our experiment, we first train a 64-component speaker NMF model using 10 IEEE sentences (see Section V-A). In the test phase, we use the same VAD algorithm as in the proposed system to extract noise frames from the mixture, and then use these noise frames to train a 16-component noise NMF model. Finally, we use the obtained speaker model and noise model as priors to separate speech. In addition, we compare with the systems described in Section V-A. Fig. 10 shows the comparative results in terms of HIT-FA rates. As shown in the figure, the proposed system achieves the highest HIT-FA rates except for N1, N5 and N9 where NMF-based system performs slightly better. On average, the proposed system outperforms the NMF-based method by around 5%, which is statistically significant from confidence intervals.

As described in Section V-A, we can resynthesize waveform signals for the four systems and calculate SNR gains. Fig. 11 shows such results. Our system improves SNRs by 5 dB to 12 dB, depending on noise type, and it performs better than Kim

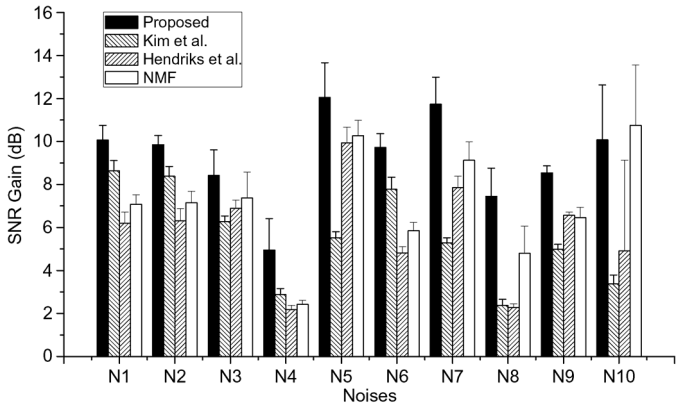

Fig. 11. Noise generalization comparisons in terms of SNR gain.

et al. by 3.7 dB, Hendriks *et al.* by 3.4 dB, and NMF-based system by 2.1 dB on average.

VI. Discussion

Monaural speech separation is a fundamental problem in speech processing. Supervised learning algorithms have been shown to be effective for speech separation, but a major issue for supervised learning is the capacity of generalization to unseen conditions, as the training set and the test set can have dissimilar properties. If this issue is not addressed, one cannot expect the trained model to perform well in unmatched conditions.

This study builds on SVM classification. An SVM outputs binary labels according to decision values, which in essence give a distance measure to the decision hyperplane, corresponding to the confidence of classification. Under many unseen conditions, the trained SVM model does not completely fail, but the optimal hyperplane just skews from the trained hyperplane to some extent. Our analysis suggests that it is possible to improve classification results by adjusting the hyperplane, which is equivalent to using a new threshold to binarize output values. Therefore, the key idea of generalization in this study is to use rethresholding to adapt the trained model to unseen conditions and the generalization issue becomes how to find appropriate thresholds. Recent research on dataset shift in classification deals with the mismatch problem between the training data and the test data [22]. In our study, shifted data lead to changes of $P(Y|\mathbf{x})$, resulting in a shift of the optimal decision boundary. In this case, rethresholding is equivalent to adjusting SVM outputs $P(Y|\mathbf{x})$. It would be interesting to explore the formulation of rethresholding as dataset shift in future work.

In this study, we convert decision values to posterior probabilities. With the probabilistic interpretation of SVM outputs, a straightforward idea to deal with generalization is to perform probabilistic inference using prior knowledge. However, too many unpredictable variables affect the probabilistic inference, and it is very difficult to directly use the Bayesian formula to derive an appropriate threshold. Instead, we use probabilities to provide initial classification and incorporate statistical properties of the test mixture to classify T-F units. Here, we prefer probabilities to decision values because the probabilistic representation provides a uniform range of [0, 1] for rethresholding. We should state that rethresholding is not able to completely resolve the generalization issue, because even optimal thresholds may not be good enough, e.g., to achieve greater than 80% HIT-FA rates. However, as rethresholding directly focuses on the outputs of the trained model and does not require extra training, it is easy to incorporate into existing systems for improved generalization.

Under unseen SNR conditions, although the trained hyperplanes cannot be directly used to classify T-F units, the statistical properties of SVM outputs exhibit similarity at different SNRs, which provides a basis to adjust the hyperplanes. The distribution fitting based rethresholding determines the thresholds based on the test mixture and does not require any input SNR estimation.

This distribution fitting method does not work under unseen noise conditions, as no distribution is able to characterize the SVM outputs of various noises. Indeed, we tried a function approximation approach that learns a mapping from SVM outputs to optimal thresholds. However, such a mapping is not applicable to all noise types. Instead, we use VAD to detect a small amount of noise and construct a development set to choose thresholds. Obviously, the performance of our system depends on the VAD algorithm. To improve VAD results, we utilize detected pitch to remove spurious noise frames. This strategy provides a reliable set of noise frames. This is confirmed in our experiments where clean speech, rather than noisy speech, is used to produce ideal VAD results. The experiments do not show significantly better performance by using the ideal VAD results. Therefore, our pitch-improved VAD method is not a bottleneck of the proposed system.

Obviously, features play a crucial role in classification. We use pitch-based features and AMS features for unseen SNR generalization, as this combination has proven to be effective under matched noise conditions. For noise generalization, we use pitch-based features and RASTA-PLP features, both of which capture speech information and are robust to different noise conditions. Other features may also show robust performance under different noisy conditions, but here we are only concerned with generalization based on trained classifiers and do not focus on the selection of robust features (see [40]). We point out that, since AMS features and RASTA-PLP features are not able to distinguish different voices and the VAD algorithm can only detect nonspeech intervals in a noisy mixture, our system cannot be applied to separate multiple talkers.

In this study, we address the generalization problems to different SNRs and different noises separately. In practice, both situations may need to be considered simultaneously. In such situations, rethresholding may still be applicable. Future research is required to address this more challenging case, and may involve some form of SNR detection to jump start the separation process.

To conclude, we aim to design a speech separation system that requires minimal training but is generalizable to unseen conditions. The proposed system trains SVMs to provide initial classification and then uses the rethresholding technique to estimate the IBM. To determine the thresholds under unseen SNR conditions, we use a distribution fitting method. For unseen noise conditions, we use a VAD algorithm to produce noise-only frames and determine the thresholds from a small development set.

Auditory segmentation is incorporated to further improve the rethresholded mask. The experiments and comparisons show that the proposed approach achieves good generalization in unmatched conditions.

REFERENCES

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

[2] P. Boersma and D. Weenink, 2007, PRAAT: Doing Phonetics by Computer (Version 4.5) [Online]. Available: http://www.fon.hum.uva.nl/praat

[3] D. C. Boes, F. A. Graybill, and A. M. Mood, *Introduction to the Theory of Statistics, 3rd ed.*. New York: McGraw-Hill, 1974.

[4] J. Brank, M. Grobelnik, N. Milic-Frayling, and D. Mladenic, "Training text classifiers with SVM on very few positive examples," Microsoft Corp., Tech. Rep. MSR-TR-2003-34, 2003.

[5] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990, ch. 1.

[6] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[7] C. C. Chang and C. J. Lin, 2001, LIBSVM: A Library for Support Vector Machines, [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[9] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[10] K. Han and D. L. Wang, "An SVM based classification approach to speech separation," in *Proc. IEEE ICASSP*, 2011, pp. 5212–5215.

[11] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE ICASSP*, 2010, pp. 4266–4269.

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

[13] G. Hu, 100 Nonspeech Sounds, 2006 [Online]. Available: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html

[14] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[15] G. Hu and D. L. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 396–405, Feb. 2007.

[16] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 92–102, Jan. 2012.

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

[18] Z. Jin and D. L. 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.

[19] G. Kim and P. C. 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.

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

[21] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, pp. 1673–1682, 2008.

[22] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognit.*, vol. 45, no. 1, pp. 521–530, 2012.

[23] A. Narayanan and D. L. Wang, "A CASA based system for SNR estimation," Dept. of Comput. Sci. and Eng., The Ohio State Univ., Tech. Rep. TR36, 2011.

[24] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing (2nd ed.)*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[25] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[26] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, May 2012.

[27] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," MRC Applied Psychology Unit, Tech. Rep. No. 2341, 1988.

[28] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 1999, pp. 61–74.

[29] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[30] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 19, pp. 227–246, 1969.

[31] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, Sep. 1998.

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

[33] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[34] A. Sun, E. P. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: A comparative study," *Decision Support Syst.*, vol. 48, no. 1, pp. 191–201, 2009.

[35] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 184–192, Mar. 2003.

[36] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 2000.

[37] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA, USA: Kluwer, 2005, ch. 12, pp. 181–197.

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

[39] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. 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.

[40] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," Dept. of Comput. Sci. and Eng., The Ohio State Univ., Tech. Rep. TR37, 2011.

[41] Y. Wang and Z. Ou, "Combining HMM-based melody extraction and NMF-based soft masking for separating voice and accompaniment from monaural audio," in *Proc. IEEE ICASSP*, 2011, pp. 1–4.

[42] R. J. Weiss and D. P. W. Ellis, "Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking," in *Proc. Workshop Statist. Percept. Audition*, 2006, pp. 31–36.

[43] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 882–892, Mar. 2007.

Kun Han received the B.S. degree in electrical engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, and the M.S. degree in computer science from University of Science and Technology of China, Hefei, China. He is currently pursuing the Ph.D. degree in the Ohio State University.

His research interests include machine learning, computational auditory scene analysis, and speech processing.

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