ON GENERALIZATION OF CLASSIFICATION BASED SPEECH SEPARATION

Kun Han and *DeLiang Wang*[*]

Department of Computer Science and Engineering
& Center for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
{hank,dwang}@cse.ohio-state.edu

ABSTRACT

Monaural speech separation is a very challenging problem. Recent studies utilize supervised learning methods to estimate the ideal binary mask (IBM) to solve the problem. In a supervised learning framework, the issue of generalization to conditions different from those used in training is paramount. This paper describes methods that require only a small training corpus but can generalize to unseen conditions. The system utilizes support vector machines to learn classification cues and then employs a rethresholding method to estimate the IBM. A distribution fitting method is used to address unseen signal-to-noise ratio conditions and an iterative voice activity detection is used to address unseen noise conditions. Systematic evaluations show that the proposed approach produces high quality IBM estimates under unseen conditions.

Index Terms— Speech separation, Generalization, SVM, Rethresholding

1. INTRODUCTION

For speech separation, the ideal binary mask (IBM) has been suggested as a main goal for computational auditory scene analysis (CASA) systems [1]. The IBM is defined in terms of premixed target and interference. Specifically, with a time-frequency (T-F) representation of a sound mixture, the IBM is a binary matrix along time and frequency where a matrix element is 1 if the signal-to-noise ratio (SNR) within the corresponding T-F unit is greater than a local SNR criterion (LC) and is 0 otherwise. A series of studies shows that IBM separation produces large speech intelligibility improvements in noise [2, 3, 4].

Recent studies have utilized supervised classification based systems for IBM estimation [5, 6, 7]. Typical supervised learning requires that the distribution of the training set match that of the test set. For speech separation, if the input SNRs or background noises contained in the test mixtures are not seen in the training set, there is no guarantee that the system will achieve good classification results. Previous systems have avoided this problem by either training and testing on the same SNR and noise conditions, or training on a large variety of SNRs or noises, which requires substantial computational resources. To minimize the need for such expensive training and because one cannot expect to train on all possible conditions that will be seen in testing, it is important to design a system that is able to generalize to unseen conditions.

In this work, we are concerned with speech separation from non-speech interference. We propose a system that aims to estimate the IBM under unseen SNR or noise conditions. Our system includes an SVM based supervised learning stage following by a rethresholding step. We utilize SVMs to produce initial separation cues and then calculate new thresholds to classify T-F units. The new thresholds are adaptively computed based on the characteristics of a test mixture, which are expected to generalize to unseen SNR or noise conditions.

The paper is organized as follows. In the next section, we present an overview of the proposed system. Section 3 describes how to generalize to unseen SNR and noise conditions. The systematic evaluation results are given in Section 4. The last section concludes the paper.

2. SYSTEM OVERVIEW

The system consists of several stages. A 16000 Hz input mixture signal $s(t)$ is analyzed by a 64-channel gammatone filterbank, with center frequencies distributed from 50 Hz to 8000 Hz. In each channel, the output is divided into 20-ms time frames with 10-ms overlap between consecutive frames. This processing produces a decomposition of the input signal into a two-dimensional T-F representation, or *cochleagram* [8].

Given the filtered subband mixture, we extract two types of features from each T-F unit: pitch-based features and relative spectral transform-perceptual linear prediction (RASTA-PLP) features [9]. For pitch-based features, the autocorrelation function (ACF) $A(c, m)$ for channel c and frame m

[*] We thank John Woodruff for critically reading an earlier draft. This research is supported by an AFOSR grant (FA9550-08-1-0155) and an STTR subcontract from Kuzer.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

is computed at the pitch lag [8]. Similarly, we compute the envelope ACF, $A_E(c,m)$, which captures the amplitude modulation information in high frequency channels. In order to encode variations, we also calculate delta features. Specifically, the time delta feature $\Delta A^T(c,m)$ is the difference between $A(c,m)$ and $A(c,m-1)$ and the frequency delta feature $\Delta A^C(c,m)$ is computed in the same way. 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 performed on this filtered spectrum. This results in a 13-dimensional feature vector. We also calculate delta features for RASTA-PLP across frames and channels. Finally, both types of feature vectors are combined into a 45-dimensional feature vector for each T-F unit. This feature vector is used as the input to the SVMs.

We use probabilistic SVMs to model the posterior probability that a T-F unit is assigned a 1 by the IBM given the feature vector, denoted $P(y=1|\mathbf{x})$. A separate SVM is trained for each frequency channel. We use the radial basis function kernel, $K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\gamma||\mathbf{x}_i-\mathbf{x}_j||^2)$, where parameters are chosen by 5-fold cross-validation. A sigmoid function is used to map a decision value to a number between 0 and 1, which we then interpret as a posterior probabilitiy [10]. The SVM library LIBSVM [11] is used in our experiments.

Generally speaking, the standard SVMs use $\theta=0.5$ as the threshold to perform classification. However, in this study we train using a small number of noise types and with a fixed input SNR and wish to generalize to a large variety of unseen conditions. In this case, we do not expect the trained SVMs to produce good results directly. Motivated by [7], we incorporate a rethresholding stage to address the unmatched situation. That is, we select the threshold θ_c that maximizes the classification accuracy in channel c, and then use the new threshold to binarize the SVM outputs, i.e., $P(y|\mathbf{x})$:

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

We discuss how to determine the new thresholds θ_c in the next section.

3. GENERALIZATION TO UNSEEN CONDITIONS

3.1. Unseen SNR conditions

For the SNR generalization problem, we refer to the threshold that maximizes the classification accuracy as the optimal threshold. We observe that in unmatched SNR conditions, the optimal threshold in each channel can substantially improve the classification result relative to a threshold of 0.5. Further, although optimal thresholds vary in different SNR conditions, the SVM outputs have similar distributions, and the optimal thresholds are located at similar positions of the distributions. Fig. 1 shows histograms of the SVM outputs in the 10th channel. The system is trained on factory noise at 0 dB and SVM outputs are generated for the same noise condition at -10, -5, 0, 5 and 10 dB SNRs. The figure shows that each histogram has a peak P_k on the left side ($P<0.6$) and SVM outputs between P_k and $P=0.6$ for each histogram can be fitted by the same distribution (but with different parameter value Ω). The vertical line in each histogram indicates the optimal threshold which always locates at the tail end of each distribution.

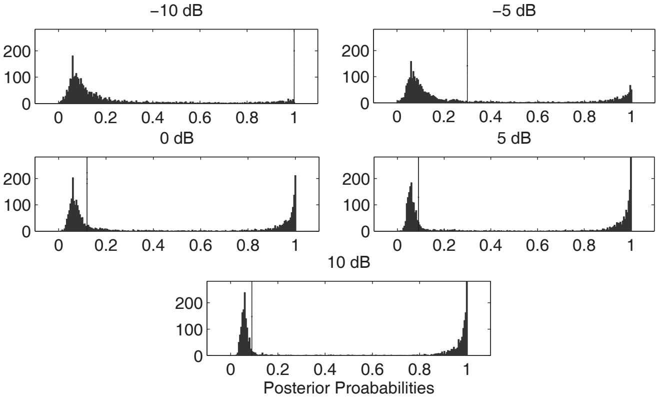

Fig. 1. Histograms of the posterior probabilities in the 10th channel with different input SNRs.

To determine the threshold in channel c, we first find the peak P_k on the left side of the histogram and then fit a predetermined distribution function $f(x;\Omega)$ to the SVM outputs. We limit the range of the distribution fit to $[P_k, 0.6]$ and use a half-Cauchy distribution in this study:

$$f(x;\mu,\sigma)=\begin{cases}\dfrac{2}{\pi\sigma[1+(\frac{x-\mu}{\sigma})^2]}, \text{if } x\geq\mu\\ 0, \text{ otherwise}\end{cases} \tag{2}$$

where, μ, σ are parameters determined by maximal likelihood estimation. Once the parameters of the distribution are fixed, the new threshold θ_c is calculated using the inverse probability density function $\theta_c=f^{-1}(\alpha;\Omega)$, where α corresponds to the tail end of the distribution and is empirically set to 0.9. This method estimates IBMs under unseen SNR conditions without the knowledge of the input SNR.

3.2. Unseen noise conditions

Another important issue is generalization to unseen noises. We find that with even a small amount of the novel noise intrusion, one can construct a development set to choose thresholds that perform well under the given noise condition. Motivated by this observation, we propose an approach that couples voice activity detection (VAD) with an iterative procedure to perform rethresholding.

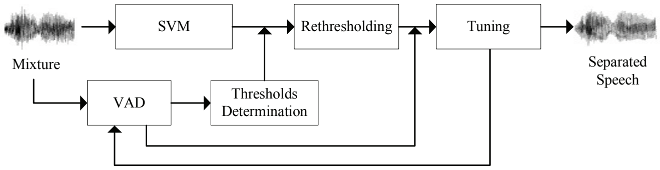

Fig. 2. Diagram of the iterative VAD based rethresholding system.

As shown in Fig. 2, 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 [12] to detect the noise frames. This algorithm produces the likelihood of speech presence for each frame. We select 20% of the frames with the lowest likelihoods as the noise frames. In addition, to avoid potentially spurious noise frames, we exclude those noise segments whose lengths are less than 5 frames.

The detected noise frames are concatenated and then mixed with a stored utterance to form a reference mixture for which we can compute the corresponding IBM. The length of the detected noise frames is often shorter than that of the stored utterance, so we simply duplicate the noise frames until their length is equal to that of the utterance. We treat the reference mixture and its IBM as a development set to select thresholds $\hat{\theta}_c^0$. With $\hat{\theta}_c^0$ and the posterior probability of speech dominance for each T-F unit of the test mixture, it is easy to produce a rethresholding mask.

We further tune the rethresholding mask by means of the VAD results. If a frame is a noise frame, it is unlikely that a unit with strong energy in this frame is target-dominant. Therefore, we first calculate the mean log energy $\bar{E}$ of all T-F units in the mixture. For those units within the frames with the lowest 10% speech presence likelihoods, a 1-labeled unit is relabeled as 0, if the corresponding log energy is greater than $0.8\bar{E}$, because the energy probably comes from noise. After the VAD based tuning process, some false alarm units are corrected and a tuning mask is formed.

The above process generates good estimated IBMs, and we employ an iterative scheme to further improve the results. We first utilize the tuning mask to produce better VAD results. If most energy of a frame comes from noise, the frame is probably a noise frame. Thus, a frame is marked as a noise frame if the mean of the log energy in the 0-labeled units in this frame is more than $0.5\bar{E}$. These marked noise frames together with the noise frames detected from the VAD algorithm constitute a new noise frame set. With the new noise frame set, we generate a new reference mixture to choose the thresholds $\hat{\theta}_c^1$ and apply the same rethresholding and tuning stages as described above. In our experiments, two iterations are good enough to generate estimated IBMs and more iterations do not significantly contribute to the final results.

4. EVALUATION

4.1. Generalization results to unseen SNRs

We first evaluate the capacity of our system to generalize to unseen SNRs. The IEEE corpus [13] is used to train and test the system. For the training set, we choose 100 female utterances mixed with 3 types of noise: speech-shaped noise, factory noise and babble noise at 0 dB. The test set consists of 10 utterances mixed with the same 3 types of noise at -10, -5, 0, 5 and 10 dB. There is no overlap between the training and the test utterances. Each utterance is mixed with a noise segment selected randomly from the original noise recording. The LC is set to -5 dB for all 64 channels to generate IBMs. In order to quantify the performance of our system, we compute the HIT rate (the percent of the target-dominant units in the IBM correctly classified) and the FA rate (the percent of the interference-dominant units in the IBM wrongly classified). We give the difference between HIT and FA, HIT−FA, as it has been shown to be highly correlated with human speech intelligibility [6].

We compare the proposed system with the original SVM classification approach that does not incorporate rethresholding. As shown in Fig. 3, the proposed system achieves high HIT−FA rates and outperforms the original approach for all input SNR conditions. On average, the rethresholding method improves HIT−FA by 7% for five input SNRs.

In [7], we quantitatively compared our original approach to Kim *et al.*'s system [6] which uses Gaussian mixture models for classification. Their system is trained under -5, 0 and 5 dB input SNRs. Although limited space does not permit a detailed comparison here, we point out that the proposed system achieves considerably higher HIT−FA rates than Kim *et al.*'s.

4.2. Generalization results to unseen noises

To evaluate generalization to unseen noises, we choose 30 female utterances mixed with 5 types of noise out of a corpus of 100 nonspeech noise types. We set the input SNR to 0 dB to train the system. To test our system, we use 10 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

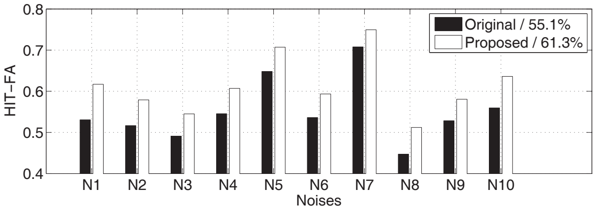

Fig. 4. Noise generalization results for 10 unseen noises.

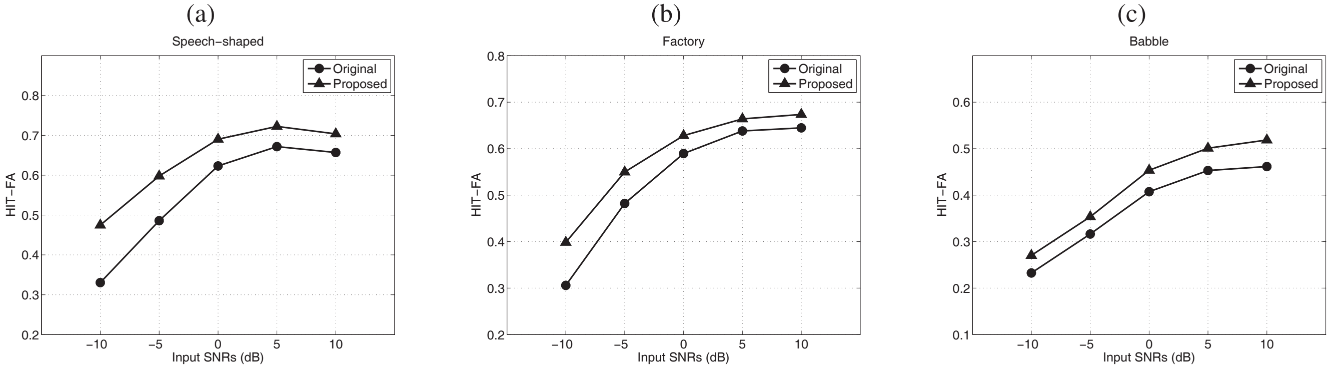

Fig. 3. SNR generalization results for (a) speech-shaped noise, (b) factory noise and (c) babble noise.

test noises cover both stationary and nonstationary noises and have very different frequency characteristics, none of which is seen in the training set.

Fig. 4 shows the comparison with the original SVM classification approach. The proposed system achieves higher HIT−FA rates under all unseen noise conditions. On average, our system outperforms the original approach by 6%, which demonstrates that with a small amount of training, our system can generalize to a large variety of unseen noise conditions. Compared with Kim *et al*.'s system, our system performs substantially better under unseen noise conditions.

5. CONCLUSION

This study aims to design a speech separation system that requires minimal training but is able to generalize to unseen conditions. The proposed system trains SVMs to model the posterior probability of each T-F unit, and then uses rethresholding to estimate the IBM. For unseen SNR conditions, we use an empirical method that does not require knowledge of the input SNR to determine thresholds. For unseen noise conditions, we propose an iterative scheme that incorporates a VAD algorithm to determine thresholds and estimate IBMs. The experiments show that the proposed approach achieves good generalization results for unseen conditions.

6. REFERENCES

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

[2] 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. Am.*, vol. 120, no. 6, pp. 4007–4018, 2006.

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

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

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

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

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

[8] D. L. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms and applications*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2006.

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

[10] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*, Cambridge, MA, USA, 1999, pp. 61–74, MIT Press.

[11] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001, http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

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

[13] 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 Electroacoustics*, vol. 17, pp. 227–246, 1969.