AN SVM BASED CLASSIFICATION APPROACH TO 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 task. CASA-based systems utilize acoustic features to produce a time-frequency (T-F) mask. In this study, we propose a classification approach to monaural separation problem. Our feature set consists of pitch-based features and amplitude modulation spectrum features, which can discriminate both voiced and unvoiced speech from nonspeech interference. We employ support vector machines (SVMs) followed by a re-thresholding method to classify each T-F unit as either target-dominated or interference-dominated. An auditory segmentation stage is then utilized to improve SVM-generated results. Systematic evaluations show that our approach produces high quality binary masks and outperforms a previous system in terms of classification accuracy.

Index Terms— Speech separation, IBM, SVM, Re-thresholding, Segmentation

1. INTRODUCTION

A key problem in speech processing is speech separation from a monaural recording, that is, to separate a speech signal from its background interference. In this setting, one can only consider the intrinsic properties of speech or interference to distinguish and separate them. This problem has proven to be very challenging.

Inspired by human auditory perception, CASA (computational auditory scene analysis) aims to separate a sound mixture into different auditory streams based on perceptual principles [1]. An ideal binary mask (IBM) has been proposed as a main computational goal of CASA [2]. The IBM is defined in terms of premixed target and interference. Specifically, for a time-frequency (T-F) unit, if the signal-to-noise ratio (SNR) within the unit is greater than a local criterion (LC), it will be labeled as 1 and otherwise it will be labeled as 0. Previous studies show that IBM separation produces large improvements in human speech intelligibility [3, 4, 5]. Therefore, one way to approach monaural speech separation is to estimate the IBM.

IBM estimation can be viewed as binary classification. To our knowledge, the first attempt to treat speech separation as binary classification was made in the binaural domain [6]. Supervised classification has also been recently studied for monaural speech separation [7, 8].

From the classification point of view, the first issue to address is feature extraction. Pitch, or harmonic structure, is a prominent characteristic of speech signals and has been proven to be effective for separating voiced speech from other sounds. Pitch-based features are robust to various forms of signal corruption but they cannot address separation of unvoiced speech which lacks harmonic structure. On the other hand, amplitude modulation spectrum (AMS) has been used as a feature for discriminating both voiced and unvoiced speech from nonspeech intrusions [8]. In this study, we propose to combine these two types of features and construct a larger feature set for classification.

Obviously, classifier design is also important for successful classification. The task here is to classify T-F units in terms of extracted features. In this study, we propose to employ support vector machines (SVMs), which are largest-margin classifiers with good generalizability. A typical output from the discriminant function of an SVM is a real number indicating the distance from the decision boundary, and the threshold of 0 is commonly used to binarize the output for classification. In this study, we introduce a re-thresholding strategy in order to maximize the hit rate minus false-alarm (FA) rate. In addition, we employ an auditory segmentation method to further improve the classification results.

The paper is organized as follows. In the next section, we present the proposed system in detail. The experimental results and comparisons are given in Section 3. The last section concludes the paper.

2. SYSTEM DESCRIPTION

As shown in Fig.1 the proposed system consists of four stages. The first stage is the auditory peripheral analysis. An input mixture $x(t)$ is analyzed by a bank of 64 gammatone filters from 50 Hz to 8000 Hz [1]. In each channel, the output is divided into 20-ms time frames with 10-ms overlapping between consecutive frames. This processing produces a two-dimensional time-frequency representation, called *cochleagram*. In the next stage, features are extracted from each unit in the cochleagram. Then, a trained SVM classifies T-F units to 1 or 0 in each channel. By combining these classification results we obtain an estimated IBM. An auditory segmentation stage utilizes cross-channel correlation and onset/offset information to improve the mask. Details are presented as follows.

2.1. Feature Extraction

For pitch-based features, the autocorrelation function (ACF) $A(c, m, \tau_m)$ for channel c and frame m is computed at the pitch lag τ_m [1]. The range for time delays in ACF corresponds to the plausible pitch range of 80 to 500 Hz. Similarly, we compute the envelope ACF, $A_E(c, m, \tau_m)$, which captures the amplitude modulation information in high frequency channels. In order to encode variations, we also calculate delta features. Specifically, for $m \geq 2$, time delta

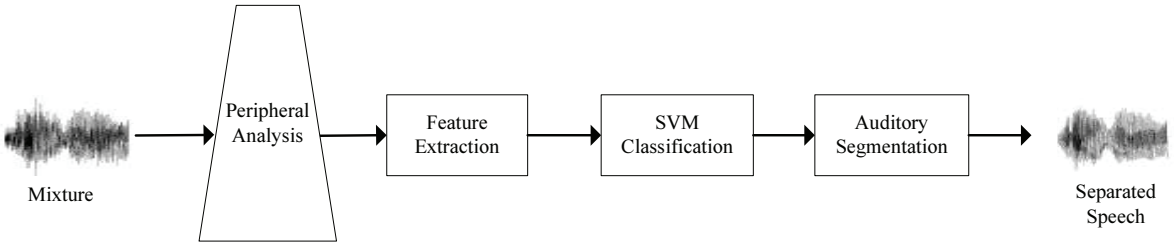

Fig. 1. Diagram of the proposed system

feature $\Delta A^T(c, m, \tau_m)$ is the difference between $A(c, m, \tau_m)$ and $A(c, m-1, \tau_m)$, and $\Delta A^T(c, 1, \tau_m)$ is set to $\Delta A^T(c, 2, \tau_m)$. We compute frequency delta feature $\Delta A^C(c, m, \tau_m)$ in the same way. The pitch-based feature vector is then given by:

$$\mathbf{x}_A(c,m) = \begin{pmatrix} A(c,m,\tau_m) \\ A_E(c,m,\tau_m) \\ \Delta A^T(c,m,\tau_m) \\ \Delta A_E^T(c,m,\tau_m) \\ \Delta A^C(c,m,\tau_m) \\ \Delta A_E^C(c,m,\tau_m) \end{pmatrix}$$

Here, the pitch period τ_m needs to be specified during the calculation. In the training stage, we use *Praat* [9] to extract the ground-truth pitch. In the test stage, we use a pitch detector [10] to estimate pitch from the mixture. Note that, unvoiced frames do not include pitch information, so we simply put 0 as the value of the corresponding feature. Classification in unvoiced intervals mainly relies on AMS features.

To extract AMS features, the envelope extracted from each T-F unit is Hanning windowed with zero-padding, and a 256-point fast Fourier transform (FFT) is computed. Then, the FFT magnitudes are multiplied by 15 triangular-shaped windows and summed to produce 15 modulation spectrum amplitudes, which represent the AMS feature vector [8]. We denote them by $M_1(c, m), ..., M_{15}(c, m)$. Similarly, we calculate delta features ΔM^T and ΔM^C across frames and channels, respectively, and the overall AMS feature vector is given by:

$$\mathbf{x}_M(c,m) = \begin{pmatrix} M_1(c,m) \\ \cdots \\ M_{15}(c,m) \\ \Delta M_1^T(c,m) \\ \cdots \\ \Delta M_{15}^T(c,m) \\ \Delta M_1^C(c,m) \\ \cdots \\ \Delta M_{15}^C(c,m) \end{pmatrix}$$

The total dimension of the AMS feature vector $\mathbf{x}_M(c, m)$ is $3 \times 15 = 45$. Finally, we combine the pitch-based feature vector and AMS feature vector into a 51-dimensional feature vector for each T-F unit.

2.2. SVMs Classification

SVMs are used to estimate the IBM in each channel. We choose the radial basis function $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2)$ as the kernel and the parameters are tuned by 5-fold cross-validation. The SVM library LIBSVM [11] is used in our experiments.

In the test stage, the binary label of each datum is typically given by the sign of the real number computed from the discriminant function:

$$f(\mathbf{x}) = \sum_{i \in \text{SV}} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + \beta \tag{1}$$

where SV denotes the set of support vector indices in training data. $\mathbf{x}_i$ are support vectors and y_i are the corresponding labels. α_i are Lagrange multipliers and β is the bias, both of which can be determined in the training stage. The absolute value of $f(\mathbf{x})$ is proportional to the distance to the decision hyperplane.

We find that the trained SVMs tend to under-label target-dominated units for several reasons. The first reason is that, with unbalanced training samples, the SVM hyperplane is often skewed to the minority [12]. For IBM estimation, due to the concentration of speech energy, when the input SNR is around 0 dB and the interference is a nonspeech noise, target-dominated units are much fewer than interference-dominated units. Second, we use different pitch trackers in the training and test stages, which introduces some mismatch. In addition, standard SVM is designed to maximize the classification accuracy, and if we focus on other measurements, for example, the hit rate minus false-alarm rate (HIT-FA), we need to adapt the decision function in the test stage.

To rectify the under-labeling problem, we first choose a validation set with 10 sentences, and then find a new threshold θ_c that maximizes HIT-FA in channel c. The new thresholds are used to binarize $f(\mathbf{x})$:

$$y(\mathbf{x}) = \begin{cases} 1, \text{ if } f(\mathbf{x}) > \theta_c \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

2.3. Auditory Segmentation

In the segmentation stage, we utilize cross-channel correlation to segment units in voiced intervals [1]. Specifically, we compute cross-channel correlation $C(c, m)$ for each unit in low frequency channels and envelope cross-channel correlation $C_E(c, m)$ for each unit in high frequency channels [7]. Only those units with sufficiently high cross-channel correlation or envelope cross-channel correlation will be iteratively merged into segments. For unvoiced intervals, a multiscale onset/offset analysis [13] is applied to form segments, which detects sudden intensity increases (onsets) and decreases (offsets). This analysis is appropriate for segmenting unvoiced speech. Then, we group each segment into the target stream if the energy corresponding to the T-F units with the target label is greater than the energy corresponding to the T-F units with the interference label. Finally, those segments whose length is less than 50 ms are removed. By applying auditory segmentation, we obtain the final binary mask, and the separated speech can be resynthesized from this mask [1].

3. EVALUATION AND COMPARISON

Fig. 2 illustrates the results generated by our system for a 0 dB mixture of speech and factory noise. The LC is -5 dB. Fig. 2(a) shows the ideal binary mask where 1 is indicated by white and 0 by black. Fig. 2(b) shows the binary mask generated by the SVMs. Comparing it with Fig. 2(a), the SVMs correctly label most target-dominated units in both voiced and unvoiced speech intervals, but at the expense of adding a small amount of interference. As shown in Fig. 2(c), by incorporating segmentation, our the system is able to recover more target-dominated units and remove most isolated interference-dominated units.

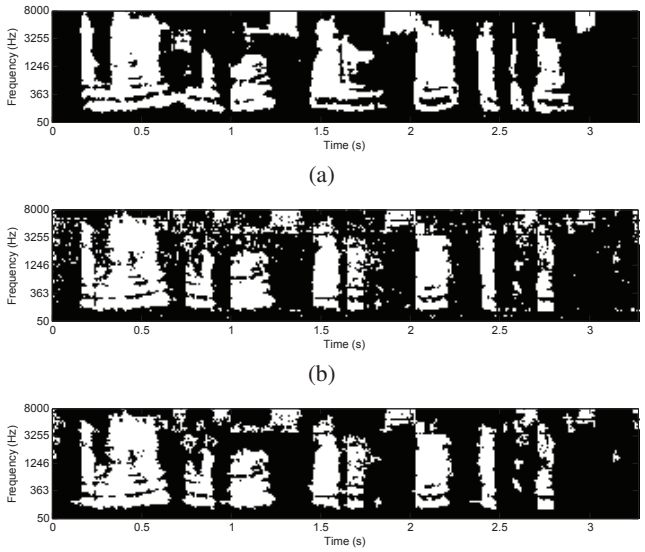

Fig. 2. Binary Masks Illustration. (a) IBM for a female utterance mixed with a factory noise. (b) SVM-generated mask. (c) Estimated IBM after auditory segmentation

To systematically evaluate the performance of our system, we choose for the training set 100 female utterances from the IEEE corpus [14] mixed with 3 types of noise (speech-shaped, factory, 20-talker babble noise) at -5, 0 and 5 dB SNR. The test set consists of 60 female IEEE sentences which do not appear in the training set. Each test utterance is mixed with the 3 types of noise at -5 and 0 dB. The LC is set to -5 dB for all channels. These choices were motivated by those in [8]. In order to quantify the performance of our system, we compute the HIT rate (the percent of the target-dominated units in the IBM correctly classified) and the FA rate (the percent of the interference-dominated units in the IBM wrongly classified). We also give the difference HIT-FA as it has been shown to be highly correlated to human speech intelligibility [8]. In addition, we show classification accuracy: the percent of misclassified units with respect to the IBM.

As shown in Table 1, even with low input SNRs our system achieves relatively high HIT rates and relatively low FA rates. That HIT rates at -5 dB are lower than those at 0 dB is partly caused by more unbalanced data at the lower SNR condition. Under most conditions, HIT-FA rates are greater than 50% at -5 dB and 60% at 0 dB. Here, the babble noise results are not as good as others because it is difficult to group pitch contours under this noise condition.

Table 1 also shows a comparison with Kim *et al.*'s system [8] which utilizes Gaussian mixture models (GMMs) to learn the distribution of AMS features and uses a Bayesian classifier to label T-F units. Their system is chosen for comparison because it obtains good HIT-FA results and demonstrates improved speech intelligibility in listening tests. In addition, the AMS feature is used in both systems. For comparison, we train a 256-mixture Gaussian for each binary label and each channel. The training and test sets are the same as in our system. We can see from Table 1 that our system performs consistently better than theirs. On average, the proposed system is about 20% better than Kim *et al.*'s system in terms of HIT-FA and more than 10% better in terms of accuracy.

We also test our system on two unseen noises (white and cocktail party noise) to assess its generalizability, and the results are shown in Table 2. From Table 2, the average HIT-FA rate is about 59% which is close to those of trained noises. These results indicate good generalizability of our system to different types of noise. Table 2 also shows the corresponding results of Kim *et al.*'s system, which does not generalize well. We believe that the generalizability of our system mainly results from the use of pitch-based features.

The results in Tables 1 and 2 indicate that our system significantly ourperforms Kim *et al.*'s system. Since both the features and the classifiers used are different, it is not clear what contributes to the better performance of our system. To isolate different factors, we perform a further comparison using exactly the same frontend processor, same features, and same training methodology. Specifically, we evaluate both systems on the 25-channel mel-scale filterbank used in [8]. The training set includes 390 IEEE sentences mixed with the 3 types of noise at 3 input SNRs as in the previous experiment. The test set still includes 60 sentences mixed with the 3 noises at -5 and 0 dB. The LC is set to -8 dB for low frequency channels and -16 dB for high frequency channels. The auditory segmentation stage is excluded from our system in this experiment. In other words, the only difference is the classifier used. We should point out that we directly use the program code with trained GMMs provided by them. The comparative results are given in Table 3. As shown in the table, our system obtains HIT rates higher than 74% and FA rates lower than 17% under all conditions. Compared to Kim *et al.*, our system improves HIT-FA rates by about 3.5% at -5 dB and 8.9% at 0 dB. These improvements clearly demonstrate the advantage of SVM classification.

The above experiments show that the proposed system produces better estimated IBMs. Since we achieve higher HIT-FA rates than Kim *et al.*'s system, it is reasonable to project that our separation results will lead to significantly improved speech intelligibility in these noisy conditions for human listeners.

4. CONCLUSION

In this study, we treat speech separation as binary classification. We utilize SVMs to classify T-F units using pitch-based and AMS features. An auditory segmentation method further improves classification results. Systematic evaluations show the effectiveness of the proposed system for IBM estimation. Comparing with Kim *et al.*'s recent study [8], there is a strong prospect that speech separation by our system can lead to improved speech intelligibility in noise.

Acknowlegements

This research is supported by AFOSR grant (FA9550-08-1-0155). We thank G. Kim and P. Loizou for providing their system code and Z. Jin for providing his pitch tracking code.

Table 1. Classification results for different noises

		Speech-shaped		Factory		Babble	
		-5 dB	0 dB	-5 dB	0 dB	-5 dB	0 dB
Proposed	**HIT**	60.14%	69.89%	60.02%	70.52%	61.43%	69.00%
	FA	4.10%	3.89%	8.60%	7.09%	17.58%	16.12%
	HIT-FA	56.04%	66.00%	51.42%	63.43%	43.85%	52.88%
	Accuracy	90.33%	89.60%	86.09%	87.02%	77.52%	78.63%
Kim *et al.*	**HIT**	59.74%	61.02%	57.39%	60.38%	53.85%	56.30%
	FA	20.70%	16.20%	26.71%	22.43%	27.18%	24.60%
	HIT-FA	39.04%	44.82%	30.68%	37.95%	26.67%	31.71%
	Accuracy	76.25%	78.15%	70.60%	73.05%	68.40%	68.86%

Table 2. Classification results for new noises

		White		Cocktail-party	
		-5 dB	0 dB	-5 dB	0 dB
Proposed	**HIT**	69.44%	72.55%	54.31%	66.29%
	FA	7.25%	8.32%	7.02%	6.27%
	HIT-FA	62.19%	64.23%	47.29%	60.02%
	Accuracy	88.81%	87.00%	83.34%	84.03%
Kim *et al.*	**HIT**	48.32%	56.40%	55.43%	58.54%
	FA	25.80%	25.61%	29.13%	24.36%
	HIT-FA	22.52%	30.78%	26.31%	34.17%
	Accuracy	69.83%	69.99%	67.03%	69.60%

Table 3. Classification results with AMS features

		Speech-shaped		Factory		Babble	
		-5 dB	0 dB	-5 dB	0 dB	-5 dB	0 dB
SVM	**HIT**	77.51%	82.87%	74.26%	82.25%	80.84%	83.08%
	FA	8.43%	10.10%	12.89%	13.89%	15.80%	16.60%
	HIT-FA	69.08%	72.77%	61.37%	68.35%	65.04%	66.48%
GMM	**HIT**	80.84%	79.64%	81.91%	81.35%	81.36%	78.40%
	FA	13.27%	14.70%	24.01%	21.89%	16.57%	16.44%
	HIT-FA	67.57%	64.94%	57.90%	59.46%	64.79%	61.96%

5. REFERENCES

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

[2] 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.

[3] 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," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4007–4018, 2006.

[4] 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.

[5] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *The Journal of the Acoustical Society of America*, vol. 123, pp. 1673–1682, 2008.

[6] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

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

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

[9] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer (version 4.5)," 2007, Software available at http://www.fon.hum.uva.nl/praat.

[10] Z. Jin and D. L. Wang, "A multipitch tracking algorithms for noisy and reverberant speech," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010, pp. 4218–4221.

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

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

[13] G. Hu and D. L. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 396–405, 2007.

[14] 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 Transactions on Audio Electroacoustics*, vol. 17, pp. 227–246, 1969.