A classification based approach to speech segregation

Kun Han[a] and DeLiang Wang
Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

(Received 24 November 2010; revised 14 August 2012; accepted 4 September 2012)

A key problem in computational auditory scene analysis (CASA) is monaural speech segregation, which has proven to be very challenging. For monaural mixtures, one can only utilize the intrinsic properties of speech or interference to segregate target speech from background noise. Ideal binary mask (IBM) has been proposed as a main goal of sound segregation in CASA and has led to substantial improvements of human speech intelligibility in noise. This study proposes a classification approach to estimate the IBM and employs support vector machines to classify time-frequency units as either target- or interference-dominant. A re-thresholding method is incorporated to improve classification results and maximize hit minus false alarm rates. An auditory segmentation stage is utilized to further improve estimated masks. Systematic evaluations show that the proposed approach produces high quality estimated IBMs and outperforms a recent system in terms of classification accuracy. © *2012 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4754541]

PACS number(s): 43.72.Dv [CYE] Pages: 3475–3483

I. INTRODUCTION

Monaural speech segregation is the task of segregating a speech signal from its background interference from a monaural recording. For this task, the information regarding sound directions is not available; and one can only make use of the intrinsic acoustic properties of speech and interference. The task has proven to be extremely challenging (Wang and Brown, 2006). In this work, we are concerned with monaural segregation of speech from non-speech interference.

Psychoacoustic research in *auditory scene analysis* (ASA) (Bregman, 1990) has inspired considerable work in developing computational auditory scene analysis (CASA) systems for speech segregation (Wang and Brown, 2006). The ideal binary mask (IBM) has been suggested as a main goal for CASA systems (Wang, 2005). 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 recent studies shows that IBM segregation produces substantial speech intelligibility improvements in noise for both normal-hearing and hearing-impaired listeners (Anzalone *et al.*, 2006; Brungart *et al.*, 2006; Li and Loizou, 2008; Wang *et al.*, 2009).

Because the IBM is a matrix of binary values, IBM estimation is a form of binary classification. To our knowledge, the first attempt to treat speech segregation as binary classification was made in the binaural domain (Roman *et al.*, 2003). Recently, several studies have utilized supervised classification to deal with monaural speech segregation (Jin and Wang, 2009; Kim *et al.*, 2009). More specifically, Jin and Wang (2009) employed multilayer perceptron (MLP) based classifiers and trained these classifiers to classify T-F units using pitch-based features. Their system obtains promising separation results in various reverberant conditions and generalizes well to new utterances and new speakers. Kim *et al.* (2009) used Gaussian mixture models (GMMs) to learn the distribution of amplitude modulation spectrum (AMS) features for target- and interference-dominant classes and then classified T-F units by Bayesian classification. Their system represents the first monaural segregation algorithm with demonstrated speech intelligibility improvement.

From the classification point of view, the first issue to address is feature extraction. The features used should distinguish target-dominant units from interference-dominant units. Pitch, or harmonic structure, is a prominent feature in voiced speech. Some previous studies show that pitch-based features are very effective for IBM estimation and robust to various forms of signal corruption (Hu and Wang, 2004; Jin and Wang, 2009). However, pitch-based features cannot address unvoiced speech segregation because unvoiced speech lacks harmonic structure. On the other hand, AMS contains information for discriminating both voiced and unvoiced speech from nonspeech intrusions (Tchorz and Kollmeier, 2003; Kim *et al.*, 2009). We propose to combine these two types of features and construct a larger feature set for classification that is expected to be discriminative in both voiced and unvoiced speech and generalize to different noise types.

Another important issue for classification is classifier design. Previously, MLPs (Hu and Wang, 2008; Jin and Wang, 2009) and GMMs (Kim *et al.*, 2009) have been explored. In this study, we propose using support vector machines (SVMs), which find an optimal (i.e., largest margin) hyperplane to classify data (Vapnik, 2000). Typically, the output of the discriminant function of an SVM is a real number, the absolute value of which indicates the distance

[a]Author to whom correspondence should be addressed. Electronic mail: hank@cse.ohio-state.edu

from the optimal hyperplane. The threshold of 0 is commonly used to binarize the output to calculate the label of each datum. In this study, we introduce a re-thresholding technique to improve classification results and maximize the hit rates minus false-alarm rates. In addition, we incorporate an auditory segmentation method to group more target-dominant units and remove interference-dominant units (Jin and Wang, 2009).

The paper is organized as follows. In the next section, we present an overview of the proposed system. Section III describes how to extract auditory features. A detailed description of SVM classification is presented in Sec. IV. Section V describes the auditory segmentation stage. The systematic evaluation results and comparison are given in Sec. VI. We discuss related issues and conclude the paper in Sec. VII.

II. SYSTEM OVERVIEW

Figure 1 shows the diagram of the proposed system, which consists of several stages. The first stage of the system is auditory peripheral analysis. An input mixture signal $x(t)$ is resampled to 16 000 Hz and analyzed by a 64-channel gammatone filterbank with their center frequencies distributed from 50 to 8000 Hz (Wang and Brown, 2006). This filterbank is a standard model of cochlear filtering and is derived from psychophysical studies of the auditory periphery (Patterson *et al.*, 1988). In each channel, the output is divided into 20-ms time frames with 10-ms overlapping between consecutive frames. This processing produces a decomposition of the input signal into a two-dimensional T-F representation or *cochleagram* (Wang and Brown, 2006). Each T-F unit in the cochleagram corresponds to a frequency channel and a time frame.

The next stage, feature extraction, extracts two types of features from each T-F unit: Pitch-based features and AMS features. After the feature extraction stage, we train SVMs to classify T-F units as either target-dominant or interference-dominant. Due to frequency specific characteristics of the input signal, one SVM is trained for each channel independently. Finally, in the auditory segmentation stage, we perform cross-channel correlation and onset/offset analysis to generate T-F segments. The T-F units in a segment primarily originate from the same sound source, and therefore we group them into either the target or interference stream based on unit classification results. The final binary mask represents an estimate of the IBM and is used to resynthesize segregated target speech. The resynthesis is basically performed by summing the filter responses in target-dominant units and compensating for phase shifts across the filterbank (Wang and Brown, 2006).

III. FEATURE EXTRACTION

A. Pitch-based features

Let $u_{c,m}$ denote a T-F unit for channel c and frame m and $x(c, t)$ denote the filter response for channel c at time t. To extract pitch-based features for $u_{c,m}$, the normalized autocorrelation function (ACF), $A(c, m, \tau)$, is computed at each lag τ (Wang and Brown, 2006):

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

Here n denotes discrete time, $T_m = 10$ ms is the frame shift, and T_n is the sampling time. We use input mixtures sampled at 16 kHz in this study; this gives $T_n = 0.0625$ ms. The preceding summation is over 20 ms, the length of a time frame. We also compute envelope ACF, $A_E(c, m, \tau)$, similar to Eq. (1), which captures amplitude modulation information in high frequency channels.

For voiced speech, $u_{c,m}$ is considered target-dominant if the corresponding response or response envelope has a period close to that of the target speech, i.e., pitch period $\tau_S(m)$ (Hu and Wang, 2004). In this case, $A(c, m, \tau)$ will have a peak close to $\tau_S(m)$. Therefore we can use the ACF and the envelope ACF at the pitch lag, $A(c, m, \tau_S(m))$ and $A_E(c, m, \tau_S(m))$, to construct pitch-based features. These two features have been demonstrated to be effective for discriminating voiced speech (Hu and Wang, 2010).

As commonly done in automatic speech recognition, we calculate delta features to encode feature variations. Specifically, for $m \geq 2$, time delta feature $\Delta A^M(c, m, \tau_S(m))$ is simply set to $A(c, m, \tau_S(m)) - A(c, m-1, \tau_S(m))$; and $\Delta A^M(c, 1, \tau_S(m))$ is set to $\Delta A^M(c, 2, \tau_S(m))$ for convenience. We compute frequency delta feature $\Delta A^C(c, m, \tau_S(m))$ in the same way. The pitch-based feature vector is then given by

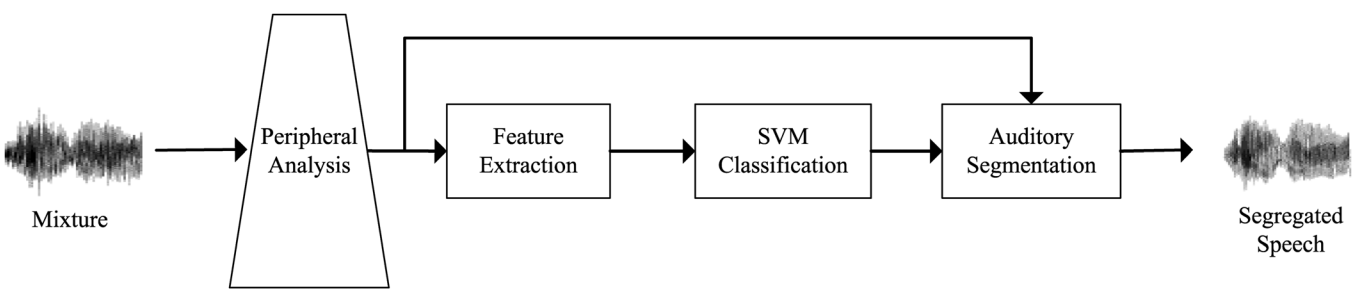

FIG. 1. Diagram of the proposed speech segregation system.

$$x_{ACF}(c,m) = \begin{pmatrix} A(c,m,\tau_S(m)) \\ A_E(c,m,\tau_S(m)) \\ \Delta A^T(c,m,\tau_S(m)) \\ \Delta A_E^T(c,m,\tau_S(m)) \\ \Delta A^C(c,m,\tau_S(m)) \\ \Delta A_E^C(c,m,\tau_S(m)) \end{pmatrix}. \quad (2)$$

When we extract the pitch-based features, the pitch period $\tau_S(m)$ needs to be specified. To remove the influence of pitch errors on the speech segregation system, we use PRAAT (Boersma and Weenink, 2007) to extract the ground-truth pitch from the premixed speech in the training phase.

In the test phase, we extract pitch from mixtures by a pitch tracker. Specifically, we use the recently proposed tandem algorithm (Hu and Wang, 2010), which iteratively estimates pitch and computes a binary mask. To further improve pitch tracking results, we generate the initial pitch estimate for the tandem algorithm by utilizing the multipitch tracker of Jin and Wang (2010) that works well when more than one voiced sound is present. The tandem algorithm produces accurate pitch estimation results under most conditions, but for some mixtures, the generated pitch contours overlap in the time domain. So we need to further group pitch contours into the target track. We first remove those pitch contours shorter than 50 ms or out of the plausible pitch range for the specific speaker; the plausible ranges of the female and male speakers are set to [150, 400 Hz] and [80, 300 Hz], respectively. For two overlapping pitch contours, we retain the one closer to the average pitch frequency (250 Hz for the female speaker and 130 Hz for the male speaker). To exclude residual interference pitch contours, we first employ a simple energy-based method to detect voiced frames. Specifically, we label a frame as strongly voiced if the normalized log energy of the frame is greater than 0.6, voiced if the energy is between 0.4 and 0.6, and unvoiced otherwise. Then a pitch contour is selected if more than 15% frames of this contour are strongly voiced or 35% frames are either voiced or strongly voiced. This simple selection method eliminates most interference pitch contours and produces the final pitch estimation result.

Note that because unvoiced frames lack harmonic structure, we simply put 0 as the values of the corresponding vector. In this way, the pitch-based features will not play a role in unvoiced frames, and classification in those frames will instead rely on AMS features.

B. AMS features

AMS features exist in both voiced and unvoiced speech, which contain information on both center frequencies and modulation frequencies within each analysis frame (Tchorz and Kollmeier, 2003). We use the same method of AMS extraction described in Kim *et al.* (2009). Specifically, we first extract the envelope from the filter response within each T-F unit. The envelopes are computed by full-wave rectification and then decimated by a factor of 4. The decimated envelope is then Hanning windowed with zero-padding, and a 256-point fast Fourier transform (FFT) is computed. The FFT computes the modulation spectrum in each T-F unit with a frequency resolution of 15.6 Hz. Next, the FFT magnitudes are multiplied by 15 triangular-shaped windows spaced uniformly across the 15.6-400 Hz range and summed to produce 15 modulation spectrum amplitudes, which represent the AMS feature vector. We denote them by $M_1(c, m),\ldots, M_{15}(c, m)$. Similarly, we calculate delta features ΔM^T and ΔM^C across frames and channels respectively, as in Kim *et al.* (2009).

The AMS feature vector is given by:

$$x_{AMS}(c,m) = \begin{pmatrix} M_1(c,m) \\ \cdots \\ M_{15}(c,m) \\ \Delta M_1^T(c,m) \\ \cdots \\ \Delta M_{15}^T(c,m) \\ \Delta M_1^C(c,m) \\ \cdots \\ \Delta M_{15}^C(c,m) \end{pmatrix}. \quad (3)$$

The total dimensionality of the AMS feature vector $\mathbf{x}_{AMS}(c, m)$ is $3 \times 15 = 45$. Finally, the pitch-based feature vector and the AMS feature vector are combined into a 51-dimensional feature vector for each T-F unit. The combined features are used as the input to the classifier.

IV. SVM CLASSIFICATION

Given the extracted features, the task now is to classify T-F units to either target-dominant or interference-dominant. As mentioned earlier, one SVM is trained for each filter channel. By applying a kernel trick, an SVM maps a feature vector $\mathbf{x}_i$ into a higher dimensional feature space where a hyperplane is derived to maximize the margin of class separation. In this study, we choose the radial basis function kernel, $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$.

In the training phase, given a set of pairs $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is a feature vector and y_i is the corresponding binary label, the SVM requires a solution to the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{w},\xi} && \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi i \\ &\text{subject to} && y_i(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathrm{x}_i) + b) \geq 1 - \xi_i, \\ &&& \xi_i \geq 0 \end{aligned} \quad (4)$$

where $\mathbf{w}$ is the weight vector of the hyperplane. ξ is a nonnegative variable measuring the deviation of a data point from the hyperplane. C controls the trade-off between complexity of the SVM and the number of nonseparable points. $\boldsymbol{\phi}$ is the vector of a set of nonlinear functions that transform the input space to a feature space of higher dimensionality. b is the bias. The parameters C and γ must be specified, and we choose them using fivefold cross-validation in each channel separately. The SVM library LIBSVM (Chang and Lin, 2001) is used in our experiments.

Once the SVM training is completed, we use the trained models to classify T-F units. The discriminant function for classification is given as follow:

$$f(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b = \sum_{i \in \mathrm{SV}} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b, \tag{5}$$

where SV denotes the set of support vector indices in training data. α_i is a Lagrange multiplier that can be determined in the training phase. For a textbook treatment of SVM, the reader is referred to Haykin (2009).

The output of the discriminant function is a real number and the binary label of each datum is typically given by the sign of this output. We find that this standard method tends to under-label target-dominant units for several reasons. First, with unbalanced training samples, the SVM hyperplane is often skewed to the minority, i.e., the class with fewer data (Akbani *et al*. 2004; Wu and Chang, 2005). For typical IBM estimation, the input SNR is around 0 dB and the interference is broadband noise. In this situation, target-dominant units are much fewer than interference-dominant units because the speech energy is more concentrated in the cochleagram than that of noise. The unbalanced data likely cause the trained SVMs to misclassify some 1 s to 0 s. The second reason is that we use different pitch trackers to extract pitch-based features in the training and test phases, which makes the hyperplane obtained from the training data not exactly match that of the test data. More discussion on this point will be given in Sec. VII. Additionally, the standard SVM aims to minimize the classification error, but one of the goals of this study is to maximize the hit rate (HIT) minus false-alarm rate (FA), or HIT-FA.

For the preceding reasons, we propose to apply re-thresholding as a *post-training* strategy, which is used in the decision phase without affecting the training phase. This technique has been successfully used in some other applications (Brank *et al*., 2003; Sun *et al*., 2009). Given a feature vector **x,** the discriminant function gives an algebraic distance from **x** to the optimal hyperplane (Haykin, 2009):

$$r = \frac{f(\mathrm{x})}{\|\mathbf{w}\|}. \tag{6}$$

Therefore those data with small $|f(\mathbf{x})|$ are close to the trained hyperplane and thus easy to be misclassified if the hyperplane is skewed. We adopt a channel-specific threshold to label $f(\mathbf{x})$. Specifically, we select the threshold θ_c that maximizes the HIT-FA rate in channel c in a validation set with 10 sentences and then use the new threshold to binarize $f(\mathbf{x})$:

$$y(\mathbf{x}) = \begin{cases} 1, & \text{if } f(\mathbf{x}) > \theta_c \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Other approaches can be used to adapt the hyperplane. For example, one can use $f(\mathbf{x})$ to estimate the *posteriori* probability $P(y=1|f(\mathbf{x}))$ and use $P(y=1|f(\mathbf{x})) = 0.5$ as a criterion to classify data (Platt, 1999). Another method is to find the threshold that makes the percentage of each class matches the percentage in the training data. We have tried both methods, but they do not perform better than the simple cross-validation method.

With SVM classification, our system generates an estimated IBM by combining the classification results in all the channels. As an example, Fig. 2 illustrates the segregation results for a noisy speech signal. Figure 2(a) shows the cochleagram of a female utterance, "A man in a blue sweater sat at the desk," from the IEEE corpus (Rothauser *et al*., 1969). Figure 2(b) shows the cochleagram of a factory noise. The cochleagram of their mixture at 0 dB is shown in Fig. 2(c). By comparing the energy of each T-F unit in Figs. 2(a) and 2(b), we obtain the IBM shown in Fig. 2(d) where 1 is indicated by white and 0 by black and LC is −5 dB. Figure 2(e) shows the binary mask generated by the standard SVMs without re-thresholding. The SVMs correctly classify most T-F units in both voiced and unvoiced speech intervals but miss some target-dominant units. By applying re-thresholding, the system recovers many target-dominant units as shown in Fig. 2(f). This recovery comes at the expense of adding some scattered interference-dominant units.

V. AUDITORY SEGMENTATION

As shown in Fig. 2, an SVM-generated mask is close to the IBM but still misses some target-dominant units and contains some interference-dominant units. We further improve estimated IBMs by auditory segmentation, which refers to a stage of processing that breaks the auditory scene into contiguous T-F regions each of which contains acoustic energy mainly from a single sound source (see also Jin and Wang, 2009; Hu and Wang, 2010).

With the voicing of a frame determined as described in Sec. III A, we utilize cross-channel correlation to segment T-F units for voiced intervals (Wang and Brown, 2006). The cross-channel correlation measures the similarity between the responses of two adjacent filters. The units with high cross-channel correlation indicate that they are likely from the same sound source. We calculate the cross-channel correlation of $u(c, m)$ as follows:

$$C(c,m) = \frac{1}{L}\sum_{\tau=0}^{L-1} \hat{A}(c,m,\tau)\hat{A}(c+1,m,\tau), \tag{8}$$

where $\hat{A}(c,m,\tau)$ denotes a normalized autocorrelation function with zero mean and unit variance, and L is the maximum delay for the plausible pitch frequency range from 70 to 400 Hz. For low frequency channels (below 2000 Hz), only units with sufficiently high cross-channel correlation ($\geq$0.95) are iteratively merged into segments. We use a similar way to calculate the cross-channel correlation of envelope response $C_E(c, m)$ and use it to segment units in high frequency channels (above 2000 Hz).

Because unvoiced speech lacks harmonic structure, we utilize onset/offset analysis (Hu and Wang, 2007) to segment T-F units within unvoiced intervals. Onsets and offsets correspond to sudden acoustic energy increases and decreases, respectively. Segments are formed by matching pairs of onset and offset fronts. In addition, a multiscale analysis is applied to integrate segments at several time-frequency scales (Hu and Wang, 2007).

With obtained segments, we first treat all the segments shorter than 50 ms (or five frames) as the interference. We then label each remaining segment wholly as the target (i.e.,

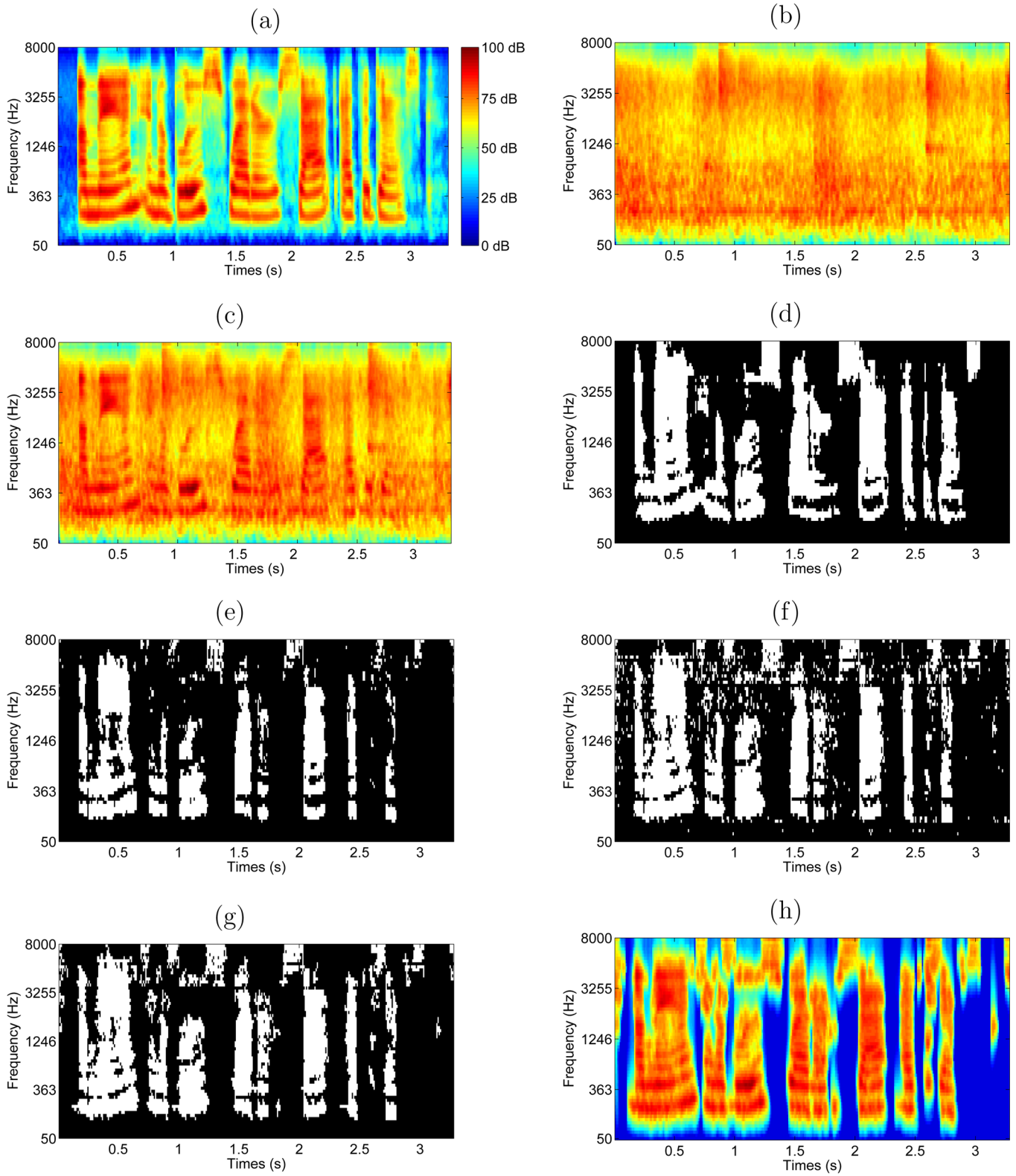

FIG. 2. (Color online) IBM estimation. (a) Cochleagram of a female utterance. (b) Cochleagram of a factory noise. (c) Cochleagram of the mixture at 0 dB. (d) IBM for the mixture. (e) SVM-generated mask without re-thresholding. (f) SVM-generated mask with re-thresholding. (g) Estimated IBM after auditory segmentation. (h) Cochleagram of the masked mixture by the estimated IBM.

mask value 1) if more than half of the segment energy is included in the classified target units in Sec. IV. If a segment fails to be labeled as the target in this way, the individually classified T-F units within the segment are still included in the target stream. This results in the final estimated IBM, and the segregated target speech can be resynthesized from this mask (Wang and Brown, 2006). Figure 2(g) shows a binary mask after auditory segmentation. We can see that most isolated interference-dominant units are removed from the mask and some missed target-dominant units are grouped at

the same time. The cochleagram of the masked mixture by the estimated IBM is shown in Fig. 2(h). Note the similarity of Figs. 2(a) and 2(h).

VI. EVALUATION AND COMPARISON

A. Systematic evaluation

We evaluate the performance of our system by using the IEEE corpus (Rothauser *et al.*, 1969), which contains 720 sentences spoken by two speakers, one male and one female. All utterances are downsampled from 25 to 16 kHz. For the training set, we choose 100 utterances mixed with three types of noise—N1: Speech-shaped noise, N2: Factory noise, N3: 20-talker babble noise—at −5, 0, and 5 dB SNR. The test set consists of 60 utterances mixed with the three types of noise at −5 and 0 dB. There is no overlap between the training and the test utterances. Each utterance is mixed with a noise sample randomly cut out from the original noise recording. The LC is set to −5 dB for all 64 channels to generate IBMs. These choices are motivated by those in Kim *et al.* (2009) where the same speech corpus and noises were used.

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 FA rate, which is the percent of the interference-dominant units in the IBM wrongly classified. It has been shown that HIT-FA is highly correlated to human speech intelligibility (Li and Loizou, 2008; Kim *et al.*, 2009). We also compute the classification accuracy, which is the percent of misclassified units.

Tables I and II show the average results for the female and male utterances, respectively. As shown in the tables, our system achieves relatively high HIT rates and relatively low FA rates even at these low input SNRs. Under all conditions, the accuracy results are greater than 75% for the female utterances and 70% for the male utterances. These results demonstrate that our system produces high quality estimated IBMs. Here, the babble noise results are relatively lower than others, mainly because it is more difficult to group pitch contours under these conditions. We also observe that the pitch determination performance of the male utterances is slightly lower than that of the female utterances, causing the classification results for the male utterances not as good as those for the female utterances. We note that, without auditory segmentation, the average HIT-FA results in Tables I and II are lower by 2% for the female utterances and 5% for the male utterances.

TABLE I. Classification results for female utterances mixed with different noises at different input SNRs.

		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 II. Classification results for male utterances mixed with different noises at different input SNRs.

		Speech-shaped		Factory		Babble	
		−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)
Proposed	**HIT**	54.05	67.41	56.07	66.72	58.98	66.01
	FA	9.83	8.06	15.31	12.51	23.44	19.17
	HIT-FA	44.22	59.34	40.76	54.21	35.54	46.84
	Accuracy	85.20	86.34	80.53	82.66	72.85	76.16
Kim *et al.*	**HIT**	57.34	57.62	54.53	55.76	46.03	49.07
	FA	19.48	14.91	24.10	20.36	25.08	22.45
	HIT-FA	37.86	42.71	30.43	35.40	20.95	26.61
	Accuracy	77.33	78.90	72.80	74.18	68.82	68.56

To provide an indication of generalizability, we also test our system on two unseen noises, N4: White noise and N5: Cocktail-party noise; different from the babble noise, the cocktail party noise mostly contains nonspeech background noise. Table III gives the results. From the table, one can see that our system achieves 58% HIT-FA rate for female speaker and 48% for male speaker on average; these are close to those with the noises in Tables I and II. We believe that the generalizability of our system mainly results from the use of pitch-based features (see the following discussion associated with Table VI and Sec. VI B).

The proposed system utilizes pitch-based features and AMS features to classify T-F units. To investigate the relative merit of each feature type, we use each type to train a classifier. The training and the test corpora are the same as those for combined features. As pitch exists only in voiced speech intervals, the system with pitch-based features is trained only during voiced intervals. Similar to the system with combined features, the ground-truth pitch is used in the training phase, and the estimated pitch is used in the test phase. For comparison, we evaluate HIT-FA results in voiced speech intervals that are determined by ground-truth pitch. Auditory segmentation is not included in all systems. Tables IV and V compare the HIT-FA results for individual feature types. On average, the system with combined features achieves the best HIT-FA rate, which outperforms the AMS features by 3.3% and pitch-based features by 2.2%. Table VI shows the comparison for new noises. In this case, the system with AMS features performs lower than that with combined features by around 20%. In contrast to AMS features, pitch-based features are robust to unseen noises and achieve comparable results with combined features. This comparison suggests that the capacity of generalization of the proposed system mainly derives from pitch-based features. AMS features capture mixture envelopes that tend to be sensitive to different noises.

Although our system is trained and tested on the IEEE corpus containing only one female and one male speaker, the classification system is expected to be speaker independent

TABLE III. Classification results for new noises.

		Female speaker				Male speaker			
		White		Cocktail party		White		Cocktail party	
		−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)
Proposed	**HIT**	69.44	72.55	54.31	66.29	71.82	77.14	46.02	61.61
	FA	7.25	8.32	7.02	6.27	16.69	17.65	14.62	15.21
	HIT-FA	62.19	64.23	47.29	60.02	55.13	59.49	31.40	46.39
	Accuracy	88.81	87.00	83.34	84.03	81.29	81.09	78.14	77.97
Kim *et al.*	**HIT**	48.32	56.40	55.43	58.54	46.60	54.24	49.78	53.82
	FA	25.80	25.61	29.13	24.36	16.82	14.84	35.58	32.62
	HIT-FA	22.52	30.78	26.31	34.17	29.78	39.40	14.20	21.21
	Accuracy	69.83	69.99	67.03	69.60	76.75	77.68	61.73	63.40

as the features used, i.e., AMS and pitch-based features, are not extracted in a speaker dependent way. To verify this, we directly use the trained models from the IEEE corpus, without change, to test on a new corpus from the TIMIT corpus (Garofolo *et al.*, 1993) which contains different speakers. Specifically, for a system of each gender, the training set contains only one speaker from the IEEE corpus, but the test set contains 10 different speakers from the TIMIT corpus, each of which produces one utterance mixed with the three noises at −5 and 0 dB SNR. The test results on TIMIT utterances are given in Tables VII and VIII. As shown in the tables, although the test set uses different speakers, the segregation results are only slightly lower than those shown in Tables I and II. On average, there is 3.4% degradation for female speakers and 2.9% for male speakers in terms of HIT-FA rates, demonstrating that the system can generalize to different speakers. On the other hand, there is some gender dependency as male and female voices show distinct feature values (particularly pitch values). Gender dependency, however, is not a big limitation as one can readily train a male model and a female model, and gender detection is a relatively easy task (Wu and Childers, 1991).

B. Comparison with the system of Kim *et al.*

Kim *et al.* (2009) proposed a speech segregation system that obtains high HIT-FA rates for noisy IEEE utterances and demonstrates improved speech intelligibility in listening tests. Here we compare our system with theirs in terms of HIT-FA. To implement their system, we use AMS features to train a 256-component GMM for each binary label in each channel and test their system on the same corpus as used in evaluating our system. The results from their system are given in Tables I–III.

From Tables I and II, one can see that our system significantly outperforms theirs in terms of HIT-FA and accuracy. The average improvements are 17% for the HIT-FA rate and 9% for accuracy. Table III shows that their system does not generalize well to the two unseen noises, where the HIT-FA rates obtained are all lower than 40%. We have computed 95% confident intervals of HIT-FA means under all conditions, all of which are less than 2.5% for the proposed system and 2% for the system of Kim *et al.* These analyses show that the performance differences are statistically significant.

As we have seen in the preceding text, these comparisons show that our system significantly outperforms the system of Kim *et al.* We should point out that the amount of training data used in the preceding comparison may be inadequate for the GMM classifiers used in Kim *et al.,* which have more parameters than the SVM classifiers used in our system. In addition, their system uses a 25-channel frontend and the preceding comparison uses a 64-channel frontend. While the reliance on a large amount of training data should be considered as a limitation, these differences nonetheless may put the system of Kim *et al.* in an unfavorable situation. To rectify this situation, we perform a further comparison using exactly the same frontend processor, same features, and same training methodology as in Kim *et al.,* except for the classifiers. Specifically, we first downsample utterances from 25 to 12 kHz and then use the 25-channel mel-scale filterbank as in the system of Kim *et al.* Only AMS features are extracted from each T-F unit. The training set includes 390 IEEE sentences, each of which is mixed with the three noises at three input SNRs as described in the previous

TABLE IV. Comparison of systems with different features for female utterances.

	Speech-shaped		Factory		Babble	
HIT-FA	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)
Combined	51.83	63.17	50.37	60.28	43.60	46.24
AMS	50.58	59.98	43.30	52.71	41.16	46.22
Pitch-based	51.40	60.69	51.13	60.02	34.13	38.45

TABLE V. Comparison of systems with different features for male utterances.

	Speech-shaped		Factory		Babble	
HIT-FA	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)
Combined	36.87	53.46	37.64	51.14	34.61	43.01
AMS	38.79	44.57	37.29	42.61	36.27	41.24
Pitch-based	34.03	52.56	40.25	57.07	27.42	38.34

TABLE VI. Comparison of systems with different features for new noises.

	Female speaker				Male speaker			
	White		Cocktail party		White		Cocktail party	
	−5 dB (%)	0 dB(%)	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)
Combined	58.99	62.02	46.89	59.07	55.43	63.25	27.80	40.78
AMS	20.57	34.11	31.06	40.80	22.45	36.30	23.29	29.53
Pitch-based	60.33	64.16	38.18	55.20	60.84	67.44	25.84	39.71

TABLE VII. Classification results for female speakers on the TIMIT utterances.

		Speech-shaped		Factory		Babble	
		−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)
TIMIT	**HIT**	63.60	70.00	58.73	72.31	57.85	65.32
	FA	12.70	6.55	12.51	11.38	17.34	14.56
	HIT-FA	50.89	63.45	46.23	60.93	40.51	50.76
	Accuracy	83.51	87.49	82.43	84.52	77.25	78.78

TABLE VIII. Classification results for male speakers on the TIMIT utterances.

		Speech-shaped		Factory		Babble	
		−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)
TIMIT	**HIT**	61.05	67.94	59.40	66.65	57.18	64.07
	FA	19.10	10.40	19.53	15.56	24.47	23.43
	HIT-FA	41.94	57.55	39.87	51.09	32.71	40.65
	Accuracy	77.74	84.12	76.76	79.78	71.34	72.32

TABLE IX. Classification results with AMS features for female utterances.

		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

TABLE X. Classification results with AMS features for male utterances.

		Speech-shaped		Factory		Babble	
		−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)	−5 dB (%)	0 dB (%)
SVM	**HIT**	67.73	74.84	65.24	75.27	76.59	76.78
	FA	6.67	7.06	15.18	15.44	18.26	16.31
	HIT-FA	61.07	67.79	50.06	59.83	56.33	60.47
GMM	**HIT**	76.00	76.34	76.90	76.99	77.66	75.69
	FA	15.75	15.57	26.16	23.44	21.90	19.56
	HIT-FA	60.24	60.77	50.75	53.54	55.77	56.13

subsection. The test set includes 60 sentences mixed with three noises at −5 and 0 dB. The LC is set to −8 dB for the lower 15 frequency channels and −16 for the higher 10 frequency channels. No auditory segmentation is applied in our system. For a rigorous comparison, we train our SVM-based system and directly use the program code with trained GMMs provided by them to estimate the IBM.

Tables IX and X show the comparative results. Our system obtains greater than 60% HIT-FA rates for the female utterances and greater than 50% HIT-FA rates for the male utterances. Compared to GMMs, SVMs improve HIT-FA rates under most conditions except for the factory noise at −5 dB for the male utterances where results are comparable. Statistically, the 95% confident intervals of the HIT-FA means for the proposed system are around ±1.5%, while those for the GMM system are ±2% on average.

VII. DISCUSSION AND CONCLUSION

In this study, we have proposed SVM-based classification for IBM estimation. As a discriminative classifier, the SVM does not model the distribution of the observed features but directly gives a predictive model conditioned on the observed data. The SVM aims to not only minimize the classification error but find a hyperplane with the largest margin; this potentially improves generalizability. In contrast, the GMM specifies a joint probability density function over observed data and labels and tends to make more assumptions than discriminative classifiers. We also attempted to use MLPs as classifiers but observed that the performance is poorer than either that of SVMs or GMMs.

By using re-thresholding, we obtain improved classification results. As standard SVMs tend to under-label T-F units, this method mainly increases HIT rates and hence improves HIT-FA rates. Although re-thresholding introduces some scattered interference-dominant units, it is easy to remove these units by auditory segmentation. Note that the setting of thresholds is application-dependent. In this study, we find that a small validation set is sufficient to find appropriate thresholds and they are robust to the choice of validation set.

Feature extraction plays an important role in classification. Pitch offers a major cue to segregate voiced speech from other sounds. However, determination of pitch in noisy conditions is a difficult task. Although we can use the ground-truth pitch to generate pitch-based features in the training phase, we have to estimate pitch from mixtures in the test phase. We have tried to use the same pitch tracker to estimate pitch in both training and test phases, which

generates matched features in the training and test phases. However, the models trained using the estimated pitch do not perform better than those using the ground-truth pitch extracted from clean speech. A pitch tracker has important influence on classification results. With better pitch estimation, our system should perform even better.

AMS features are easy to extract and exist in both voiced and unvoiced speech. As indicated in the results of Sec. VI, the generalizability of AMS features appears not as good as pitch-based features. Another limitation of AMS features is that they can only address nonspeech interference. For mixtures of two voices, AMS features are not able to distinguish them, but with multipitch tracking, pitch-based features are still discriminative even though this paper does not deal with segregation of two voices. The combination of two types of features constitutes a complementary feature set, which performs better than either type alone. In addition, as the extracted features capture speech characteristics rather than speaker characteristics, the system is speaker-independent as shown in the Sec. VI A.

In summary, we approach monaural speech segregation as binary classification. Our system extracts pitch-based and AMS features from T-F units and utilizes SVMs to classify them. An auditory segmentation stage further improves classification results. Systematic evaluations show that our system yields accurate classification results. As demonstrated in Li and Loizou (2008) and Kim *et al.* (2009), HIT-FA rates are correlated with speech intelligibility. Because our system achieves higher HIT-FA rates than the system of Kim *et al.*, it seems reasonable to expect that our system can lead to improved intelligibility. However, such expectation needs to be tested with human listeners, and we plan to conduct such tests in future research.

ACKNOWLEDGMENTS

This research was supported by an 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.

Akbani, R., Kwek, S., and Japkowicz, N. (**2004**). "Applying support vector machines to imbalanced datasets," in *Proceedings of the 15th European Conference on Machine Learning*, pp. 39–50.

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.

Boersma, P., andWeenink, D. (**2007**). "PRAAT: Doing phonetics by computer (version 4.5) [computer program]," http://www.fon.hum.uva.nl/praat (Last viewed November 2010).

Brank, J., Grobelnik, M., Milic-Frayling, N., and Mladenic, D. (**2003**). "Training text classifiers with SVM on very few positive examples," Technical Report MSR-TR-2003-34, Microsoft Corp.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (The MIT Press, Cambridge, MA), Chap. 1.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Chang, C. C., and Lin, C. J. (**2001**). "LIBSVM: A library for support vector machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm (Last viewed November 2010).

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (**1993**). "DARPA TIMIT acoustic-phonetic continuous speech corpus," National Institute of Standards and Technology, NISTIR 4930.

Haykin, S. S. (**2009**). *Neural Networks and Learning Machines* (Prentice Hall, New York), Chap. 6.

Hu, G., and Wang, D. L. (**2004**). "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Netw. **15**, 1135–1150.

Hu, G., and Wang, D. L. (**2007**). "Auditory segmentation based on onset and offset analysis," IEEE Trans. Audio, Speech, Lang. Process. **15**, 396–405.

Hu, G., and Wang, D. L. (**2008**). "Segregation of unvoiced speech from nonspeech interference," J. Acoust. Soc. Am. **124**, 1306–1319.

Hu, G., and Wang, D. L. (**2010**). "A tandem algorithm for pitch estimation and voiced speech segregation," IEEE Trans. Audio, Speech, Lang. Process. **18**, 2067–2079.

Jin, Z., and Wang, D. L. (**2009**). "A supervised learning approach to monaural segregation of reverberant speech," IEEE Trans. Audio, Speech, Lang. Process. **17**, 625–638.

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

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

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

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1988**). "An efficient auditory filterbank based on the gammatone function," Technical Report No. 2341, MRC Applied Psychology Unit.

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

Roman, N., Wang, D. L., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 227–246.

Sun, A., Lim, E. P., and Liu, Y. (**2009**). "On strategies for imbalanced text classification using SVM: A comparative study," Decision Support Syst. **48**, 191–201.

Tchorz, J., and Kollmeier, B. (**2003**). "SNR estimation based on amplitude modulation analysis with applications to noise suppression," IEEE Trans. Speech Audio Process. **11**, 184–192.

Vapnik, V. N. (**2000**). *The Nature of Statistical Learning Theory* (Springer-Verlag, New York), Chap. 5.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Dordrecht), pp. 181–197.

Wang, D. L., and Brown, G. J. (**2006**). "Fundamentals of computational auditory scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, edited by D. L. Wang and G. J. Brown (Wiley and Sons, Hoboken, NJ), Chap. 1, pp. 1–37.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336–2347.

Wu, G., and Chang, E. (**2005**). "KBA: Kernel boundary alignment considering imbalanced data distribution," IEEE Trans. Knowl. Data Eng. **17**, 786–795.

Wu, K., and Childers, D. G. (**1991**). "Gender recognition from speech. Part I: Coarse analysis," J. Acoust. Soc. Am. **90**, 1828–1840.