A Supervised Learning Approach to Monaural Segregation of Reverberant Speech

Zhaozhang Jin, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—A major source of signal degradation in real environments is room reverberation. Monaural speech segregation in reverberant environments is a particularly challenging problem. Although inverse filtering has been proposed to partially restore the harmonicity of reverberant speech before segregation, this approach is sensitive to specific source/receiver and room configurations. This paper proposes a supervised learning approach to monaural segregation of reverberant voiced speech, which learns to map from a set of pitch-based auditory features to a grouping cue encoding the posterior probability of a time–frequency (T–F) unit being target dominant given observed features. We devise a novel objective function for the learning process, which directly relates to the goal of maximizing signal-to-noise ratio. The models trained using this objective function yield significantly better T–F unit labeling. A segmentation and grouping framework is utilized to form reliable segments under reverberant conditions and organize them into streams. Systematic evaluations show that our approach produces very promising results under various reverberant conditions and generalizes well to new utterances and new speakers.**

***Index Terms*—Computational auditory scene analysis (CASA), monaural segregation, room reverberation, speech separation, supervised learning.**

I. Introduction

ROOM reverberation happens in everyday listening and it creates a considerable challenge to speech separation. While humans excel in "hearing out" a target source from sound mixtures in noisy and reverberant conditions, simulating this perceptual ability remains a fundamental challenge [57]. This paper is concerned with monaural separation of reverberant voiced speech, in which only monaural recordings are available. Monaural speech separation has many applications including hearing aid design and noise removal for automatic speech recognition.

Various methods have been proposed for monaural speech enhancement or separation, including spectral subtraction [8], Wiener filtering [14], [36], minimum mean square error (MMSE) estimation [19], and subspace analysis [20]. However, strong assumptions about the interference (e.g., quasi-stationarity) in these methods limit their application in dealing with a general acoustic background. A class of speech separation algorithms models the underlying sources and fits the learned models to the observations. The essence of such algorithms is that the expected patterns of the sources are extracted through training and then those patterns whose combinations best match the observed signal are selected to estimate the sources. In [49], speaker dependent Hidden markov models (HMMs) are trained and combined into a factorial HMM architecture for computing a masking function for separation. A modeling technique based on composite source modeling is proposed in [44] to model each source using a set of Gaussian subsources. A soft mask filter is then derived using MMSE estimation for separating the sources. These approaches can offer satisfactory solutions if extracted source characteristics match the statistical properties of mixed signals. This is, however, not always true leading to some adaptation schemes for adjusting source models [59]. Rather than modeling each individual source, the relationships between sources can also be learned using a discriminant method. One example is the spectral learning approach which is based on parameterized affinity matrices built from low-level features and solves the separation problem by formulating it as segmentation in a time–frequency (T–F) plane [4]. The performance of these methods is unclear in reverberant conditions.

Inspired by human auditory perception [10], computational auditory scene analysis (CASA) aims to separate a mixture of sources into different auditory streams based on perceptual principles [57]. CASA systems have significantly advanced the state-of-the-art performance in monaural separation [11], [25], [56], [58]. An ideal binary T–F mask has been proposed as a computational goal of CASA [55]. Such a mask can be constructed from prior knowledge of target and interference; specifically, a value of 1 in the mask indicates that the target is stronger than interference and 0 otherwise. Studies show that speech reconstructed from the idea binary mask produces large improvement in human speech intelligibility [3], [13], [35]. Such a goal has been shown to still be reasonable when room reverberation is present [42], [47].

Pitch, or harmonic structure, has long been studied as a prominent characteristic of speech signals and offers a major cue for a listener to separate target speech from other sounds [10]. The pitch cue has been applied successfully in monaural CASA algorithms under anechoic conditions (e.g., in [25]). However, the harmonic structure is distorted by reverberation as reflections of each harmonic combine with the direct sound. As a result, the performance of pitch-based CASA systems suffers in room reverberation [12], [48]. To tackle this problem, either the distorted harmonicity of the speech signal should be restored, or the low-level cues and the means by which they

Manuscript received April 28, 2008; revised November 17, 2008. Current version published March 20, 2009. This work was supported in part by the AFOSR under Grant FA9550-08-1-0155 and in part by the NSF under Grant IIS-0534707. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tim Fingscheidt.

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

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

Digital Object Identifier 10.1109/TASL.2008.2010633

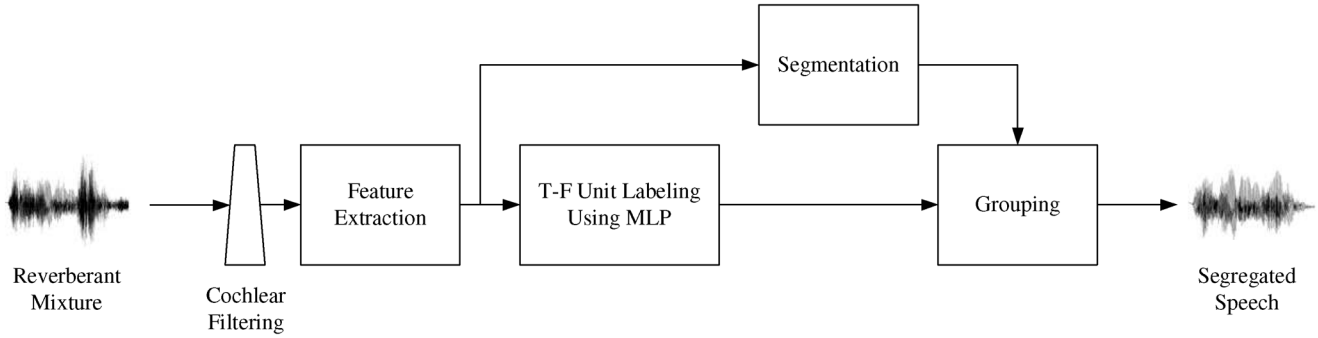


Fig. 1. Schematic diagram of the proposed system. A reverberant mixture is processed in a three-stage system. The first stage analyzes the input signal by an auditory filterbank in successive time frames and extracts pitch-based features within each time–frequency unit. In the second stage, multilayer perceptron (MLP) is trained in every channel to associate those features with the grouping cues. T–F units are then labeled according to a criterion based on the MLP output. The last stage performs segmentation and grouping. A target stream together with its background is formed.

are utilized should be improved. To restore speech harmonicity, a commonly used method is to estimate and apply an inverse filter of the room impulse response (RIR) corresponding to the target and microphone locations [22], [48], [60]. Although inverse filtering can partially counteract the smearing effect of reverberation on the speech spectrum, it assumes that a room configuration, e.g., room dimensions, wall reflection coefficients, source and microphone locations, etc., is stationary. Any change in room configuration, like a source movement, degrades the performance significantly [9], [45]. To quantify such effects, we systematically evaluate the sensitivity of inverse filtering to a number of room configurations and different reverberation times (T_{60}). As another drawback of inverse filtering, filter estimation requires the absence of interference [48], which is unrealistic for real-world application.

We focus on pitch-based features, and propose a supervised learning approach to achieve robust performance against reverberation effects for voiced speech segregation. Based on cochlear filtering, we extract a set of pitch-based features within each T–F unit from a reverberant signal. In low-frequency channels, harmonics are resolved since a filter does not respond to more than one harmonic due to its narrow bandwidth. In high-frequency channels, harmonics are unresolved since, with a wider bandwidth, a filter responds to multiple harmonics. Therefore, the feature set should contain two subsets to be sensitive to resolved and unresolved harmonics, respectively. Each subset includes several features in order to account for variations brought about by reverberation. To collectively utilize the discriminative power of the feature set in a reverberant environment, we train a multilayer perceptron (MLP) for each frequency channel in order to estimate a grouping cue within each T–F unit. The grouping cue encodes the posterior probability of a T–F unit being target dominant given observed features. By analyzing the goal of maximizing signal-to-noise ratio (SNR) in segregation, we formulate an objective function for MLP training which takes into account of unit-wise errors in a generalized form of mean squared error (MSE). Since it is a continuous function of model parameters, an error backpropagation technique can be devised in order to maximize SNR. In addition, we employ a new segmentation method to more reliably compute auditory segments in reverberant environments. Specifically, we use cross-channel correlation and temporal continuity for segmentation in the low-frequency range because they are observed to be relatively robust to reverberation [51]. In the high-frequency range, we apply onset-offset detection [27] to capture intensity variation and form segments by matching pairs of detected onsets and offsets. The motivation behind this is that auditory onsets are relatively unaffected by reverberation because the direct path from a source is the shortest path [37]. Finally, the grouping stage organizes segments into streams.

The paper is organized as follows. In the next section, we present an overview of the proposed system. Section III describes how to extract pitch-based features and perform MLP learning. A detailed description of the segmentation and grouping stage is presented in Section IV. Section V provides experimental results and comparisons. Section VI analyzes system robustness quantitatively at the feature level. We discuss related issues and conclude the paper in Section VII.

II. System Overview

As shown in Fig. 1, the proposed system consists of three stages. The first stage analyzes the input signal in the time–frequency domain using an auditory periphery model. A T–F unit corresponds to a certain channel in the filterbank at a certain time frame. Normalized correlograms are then computed. In order to detect both resolved and unresolved harmonics, auditory features are extracted based on both the filter response and the response envelope within each of the units. Section III-A gives the detail of this stage.

The next stage is to label each of the T–F units using MLP. Previous studies [25], [48] treat a single pitch-based feature as the grouping cue and rely on thresholding for unit labeling. Under reverberant conditions, such pitch-based features are no longer reliable due to smeared harmonicity. Therefore, we use multiple features to capture harmonicity within a unit and label it using a trained classifier. Our training objective is to maximize the SNR performance instead of minimizing unit labeling errors. This objective function makes the learning process cost sensitive and leads to good segregation performance. Labeling a unit based on the MLP output is treated from a probabilistic perspective. Essentially, the MLP output is translated into the posterior probability of a T–F unit belonging to the target and a labeling criterion is consequently derived. This part is described in Sections III-B and III-C.

Segmentation and grouping are two integral parts of CASA and we describe them together in the third stage. In segmentation, the input is decomposed into T–F segments, each of which is a contiguous region deemed to mainly originate from a single source. In grouping, those segments that likely come from the same source are grouped into a stream by using pitch and other grouping cues. To improve segmentation in reverberant conditions, we apply different strategies in different frequency ranges. Specifically, segmentation in low frequency merges T–F units using cross-channel correlation and temporal continuity. In high frequency, onset and offset detection is utilized. It is expected that onset cues are robust to room reverberation in the light of the precedence effect [37], which refers to the perceptual importance of a direct sound or signal onset. Once segments are formed, they are organized into the target or the interference stream resulting in a binary mask where all T–F units are labeled 1 for the target stream and 0 elsewise. This mask gives an estimate of the ideal binary mask and is used to resynthesize segregated target speech. Details are presented in Section IV.

III. Learning Grouping Cues

Our goal is to learn from a reverberant mixture detectable cues that indicate whether target speech dominates in each T–F unit. Specifically, we learn a mapping from a set of pitch-based features to a grouping cue, which encodes the posterior probability of a T–F unit being target dominant.

A. Feature Extraction

To extract pitch-based features, an input mixture $x(t)$ is first decomposed into the time–frequency domain using a gammatone filterbank and time windowing. This filterbank is a standard model of cochlear filtering and is derived from psychophysical studies of the auditory periphery [43]. We use a 128-channel filterbank whose center frequencies are quasi-logarithmically spaced from 50 to 8000 Hz. The response of a filter channel is further transduced by the Meddis model of auditory nerve transduction [39], which simulates the nonlinear properties of inner hair cells and produces the firing rate of an auditory nerve fiber, denoted by $h(c,t)$. Note that $h(c,t)$ retains the original sampling frequency. In each channel, the output is divided into 20-ms time frames with 10-ms overlapping between consecutive frames. The resulting time–frequency representation is called a cochleagram with u_{cm} denoting a T–F unit for frequency channel c and time frame m. This is a standard procedure of peripheral analysis in CASA and implementation details are presented in [57, ch. 1].

Then, the normalized correlogram $A(c,m,\tau)$ for u_{cm} for time lag τ is computed by the following autocorrelation:

$$A(c,m,\tau) = \frac{\sum_{n=-N/2}^{N/2} h(c,mN/2+n)h(c,mN/2+n+\tau)}{\sqrt{\sum_{n=-N/2}^{N/2} h^2(c,mN/2+n)}\sqrt{\sum_{n=-N/2}^{N/2} h^2(c,mN/2+n+\tau)}} \quad (1)$$

where N denotes the frame size. Since we use input mixtures sampled at 16 kHz, $N = 320$. The range of the normalized correlogram is [0,1] with value 1 at zero time lag.

Following [24, ch. 5], we construct a feature vector for each u_{cm}. The first subset of three features is derived from the hair cell output $h(c,t)$, suitable for detecting resolved harmonics in low-frequency channels. Given the pitch period τ_m at frame m, $A(c,m,\tau_m)$ is a quantitative measure of how the observed signal in u_{cm} is consistent with τ_m. This measure has already been used and proven to be effective under anechoic conditions, and we consider it as a primary feature in the set. The average instantaneous frequency $\bar{f}(c,m)$ can be estimated from the zero-crossing rate of $A(c,m,\tau)$. When multiplying $\bar{f}(c,m)$ with τ_m, the product provides an alternative way of periodicity comparison and supplements the autocorrelation measure in the feature set. So, the next two features are extracted out of this product: the second feature, the nearest integer $[\cdot]$ to the product, indicates a harmonic number, and the third feature, the distance $|\cdot|$ between the product and the nearest integer, represents the deviation between the two periods. As mentioned in Section I, to detect unresolved harmonics, the second subset of three features is based on the envelope of the hair cell output $h_E(c,t)$ and the corresponding normalized correlogram $A_E(c,m,\tau)$. Here, the purpose is to extract amplitude modulation (AM) for high-frequency channels and $h_E(c,t)$ better reveals the periodicities of these harmonics. To extract AM, we perform bandpass filtering with the passband from 50 to 550 Hz, which corresponds to the plausible pitch range of the target speech. The feature set is given as

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

These features together form the basis for the grouping cue which is robust to reverberation effects. When extracting the harmonic features, the pitch period τ_m needs to be specified. To remove the influence of pitch errors on the segregation system, we obtain *a priori* pitch contours from the premixed reverberant target speech using Praat [7].

The desired value of the grouping cue $C_g(c,m)$ is defined to be 1 if u_{cm} is dominated by the target stream and 0 otherwise, consistent with the notion of the ideal binary mask [55] which labels a T–F unit as target if and only if target energy is greater than interference energy within that unit. Thus, the ideal binary mask provides the desired values of $C_g(c,m)$.

B. MLP Training

We use an MLP to learn the grouping cue $C_g(c,m)$ from the pitch-based features $\mathbf{x}_{c,m}$. One MLP is trained for each channel. Training usually minimizes an objective function (i.e., error function) defined as the square distance between desired and actual outputs. Our previous study [31] uses a conventional MSE objective function, defined as

$$J_c = \frac{1}{M}\sum_m (d_c(m) - y_c(m))^2 \quad (3)$$

where $d_c(m)$ and $y_c(m)$ are desired (binary) and actual outputs, m frame index, M the total number of frames, and c channel index. The model using the above objective function performs reasonably well [31]. However J_c treats all T–F units equally. Such treatment may not be optimal—a T–F unit with higher energy contributes more to the overall SNR than a unit with lower energy. In other words, minimizing J_c does not necessarily lead best SNR performance.

In order to derive an objective function that directly relates to the goal of maximizing SNR, we start by analyzing the SNR definition. Since the computational goal of our proposed segregation system is to identify T–F regions that are target dominant, we use the same SNR measure in [25], which regards the resynthesized signal from the ideal binary mask as ground truth

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_t s_I^2(t)}{\sum_t \left(s_I(t) - s_E(t)\right)^2}. \tag{4}$$

Here, $s_I(t)$ and $s_E(t)$ are signals resynthesized from the ideal binary mask and an estimated mask, respectively. Consider the SNR in a single channel as training is independently conducted within individual channels. To maximize the overall SNR, we maximize SNR in each channel. Rewrite (4) for a single channel as

$$\mathrm{SNR}_c = 10\log_{10}\frac{\sum_m d_c(m)\cdot E_c(m)}{\sum_m \left(d_c(m) - Y_c(m)\right)^2 \cdot E_c(m)} \tag{5}$$

where $E_c(m)$ represents the mixture energy within u_{cm}, calculated as the sum of squares of the unit response. $Y_c(m)$ is an actual binary label, binarized from $y_c(m)$. From (5), it is intuitively clear that minimizing the denominator maximizes SNR_c. Therefore, we define the new objective function J_c' as

$$J_c' = \sum_m \left(d_c(m) - y_c(m)\right)^2 \cdot E_c(m) \Big/ \sum_m E_c(m). \tag{6}$$

Note that the function J_c' is modified from the denominator in (5) in order to make it differentiable, needed for applying gradient descent learning. The denominator in (6) is added for the purpose of normalization [cf. (3)]. It is worth mentioning that J_c' is a generalized form of MSE, with each squared error weighted by normalized energy within the corresponding T–F unit.

From the machine learning point of view, the inclusion of weights in a classification task is known as cost-sensitive learning. It is optimal learning when different misclassification errors incur different penalties, which is exactly the situation we face. The backpropagation algorithm is adapted to learn MLP parameters. In theory, each of the weights in (6) acts as a constant factor in the partial derivative J_c'. So the delta rule can be easily rewritten. It should be noted that the normalization term in (6) is necessary to ensure the convergence of the modified backpropagation algorithm [32]. In implementation, we train one MLP for each channel. Each MLP has the same network topology with six input nodes, 20 hidden nodes, and one output node. The number of hidden nodes is chosen based on tenfold cross-validation. The transfer function of the hidden and output layers are both hyperbolic tangent sigmoid. During training, we use J_c' in conjunction with a generalized Levenberg–Marquardt backpropagation algorithm [23] which achieves fast convergence by avoiding the computation of the Hessian Matrix.

C. MLP-Based Unit Labeling

For each T–F unit u_{cm}, we apply the trained MLP to feature vector $\mathbf{x}_{c,m}$ yielding $C_g(c,m)$. It should be noted that each channel has a separately trained MLP, as illustrated in Fig. 2. We then use this grouping cue to label u_{cm}. Formally speaking, the trained MLP estimates the posterior probability directly [6], [41], therefore the grouping cue can be described as

$$C_g(c,m) = P(H_1|\mathbf{x}_{c,m}) \tag{7}$$

where H_1 is the hypothesis of u_{cm} being target dominant. Let H_0 be the hypothesis of u_{cm} being interference dominant. Consequently, we define the unit labeling criterion: A T–F unit u_{cm} is labeled as target speech if $P(H_1|\mathbf{x}_{c,m}) > P(H_0|\mathbf{x}_{c,m})$. Due to the fact that $P(H_1|\mathbf{x}_{c,m})$ and $P(H_0|\mathbf{x}_{c,m})$ sum to one, the above inequality can be written as $P(H_1|\mathbf{x}_{c,m}) > 1 - P(H_1|\mathbf{x}_{c,m})$. Hence, this criterion can be simplified as

$$C_g(c,m) > 1/2. \tag{8}$$

Note that the above criterion is based on the assumption that the priors $P(H_0)$ and $P(H_1)$ remain unchanged during training and labeling phases. When we interpret the MLP output as an estimate of the posterior probability, this estimate encapsulates prior information according to the Bayes rule. In other words, the decision rule is optimal only if there is no mismatch between training and test priors. When this condition is violated, the decision rule becomes suboptimal and possibly unacceptable [1]. Although this is not a concern in our paper, we discuss in Section VII circumstances in which such a mismatch may occur and possible solutions to compensate the classifier for more reliable performance.

IV. Segmentation and Grouping

The segmentation and grouping stage segregates a reverberant mixture into a target and an interference stream. T–F unit labeling gives one way of segregation; however, it is error-prone because a local unit is too small for robust decisions in the presence of interference and room reverberation. This is supported by a comparison between the segregation results at unit and segment levels [25]. To utilize more global information of the source that is missing from individual units, we adopt the stage of segmentation in CASA (see [57, ch. 1]) and form segments on the T–F plane based on auditory cues. A segment is a contiguous region of T–F units and segment-level information is expected to provide a more robust foundation for grouping.

To segment reverberant mixtures, we apply two different strategies in different frequency ranges. Specifically, in low frequency (below 800 Hz), we merge T–F units into segments based on cross-channel correlation and temporal continuity [56]. The first cue arises from the fact that a single harmonic or formant activates a number of adjacent channels due to their overlapping bandwidths and their responses are highly correlated. In addition, a signal usually lasts for some time, which

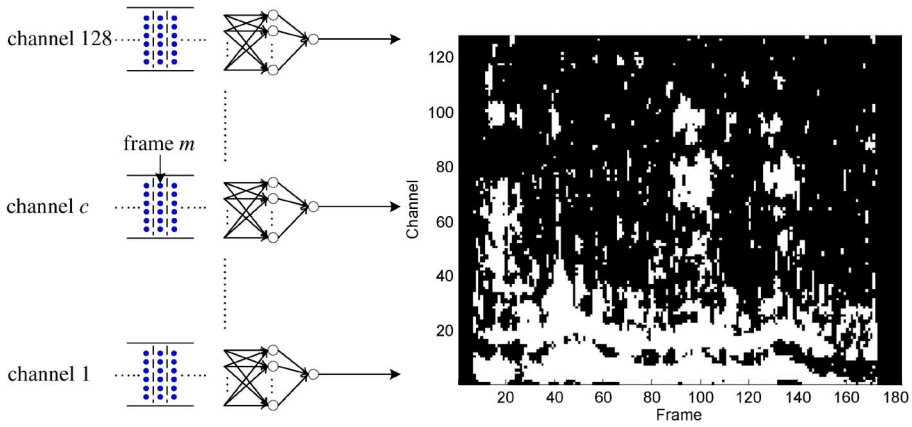


Fig. 2. Applying multilayer perceptron (MLP) in unit labeling, where a binary decision, target or interference dominant, is made in each T–F unit. The MLP for each channel is shown as a two-layer diagram, with six dimensional features (vertical dots) as inputs. In the above cochleagram, the units labeled as target dominant are indicated by white and those labeled otherwise are indicated as black.

implies temporal continuity. The cross-channel correlation is calculated as

$$C(c,m) = \frac{1}{L}\sum_{\tau=0}^{L-1} \hat{A}(c,m,\tau)\hat{A}(c+1,m,\tau) \qquad (9)$$

where $\hat{A}(c,m,\tau)$ is a normalized autocorrelation function with zero mean and unit variance and L is the maximum delay for the plausible pitch range. Only units with sufficiently high cross-channel correlation—greater than 0.99 [26]—are selected and iteratively merged into segments.

It is widely known that amplitude modulation effects of unresolved harmonics occur in high-frequency channels. In segmentation, the cross-channel correlation $C_E(c,m)$, which is calculated from $\hat{A}_E(c,m,\tau)$ (cf. (9)), has been proven to be useful [26]. However, since $h_E(c,t)$ is susceptible to room reverberation [51], $C_E(c,m)$ is sensitive to reverberation, which hinders its application (see Section V). Signal onsets, on the other hand, are largely unaltered by room reverberation because the direct sound arrives earlier than its echoes. Once onsets are detected, offsets are determined by searching for the highest intensity drop between two consecutive onsets. Therefore, we propose that high-frequency regions be segmented using onset and offset detection [27]. This method first smooths signal intensity over time in individual frequency channels to reduce insignificant fluctuations and then over frequency to enhance synchronized onsets and offsets. It then detects onsets and offsets from smoothed intensity in each channel. Segments are formed by matching pairs of onset and offset fronts, which are the vertical contours connecting onset and offset candidates across frequency. In order to achieve a compromise between over- and under-segmentation, a multiscale integration is applied from a coarse scale to a fine scale. Along the scale change, new segments are created and existing segments are better localized.

The segments obtained from the above two methods are combined to form complete segmentation. Specifically, cross-channel correlation based segments in the low-frequency range are first kept. If a segment crosses the low- and high-frequency ranges, its high-frequency portion is also maintained. Onset/offset based segments are then included; if some part of a new segment is covered by the existing segments, this part is removed before the segment is added. Fig. 3 compares segmentation with and without using onset and offset cues. As can be seen in Fig. 3(b), more significant segments in high frequency are detected, indicating more effective segmentation using onset/offset analysis. Unlike [26] which only uses onset/offset analysis for segmenting unvoiced speech (their voiced speech segmentation is based entirely on cross-channel correlation), our use of onset/offset based segmentation is limited to the high-frequency range and deals with voiced speech.

With unit labels obtained in Section III-C together with T–F segments, we group each segment into the target stream if the energy corresponding to its T–F units with target labels (1 s) dominates, i.e., greater than the energy of the T–F units with non-target labels (0 s). Finally, to group more units into the target stream, we expand each segment in the target stream by iteratively recruiting its neighboring units that are labeled as target and do not belong to any segment [26]. Consequently, a binary mask is formed and the segregated target speech can be resynthesized from this mask for performance evaluation [57, ch. 1].

Before presenting evaluation results, it may be useful to briefly contrast our system with that of Hu and Wang [26] (see also [25]) for segregating voiced speech. Despite the similarity in final grouping (as pointed out above), these two systems differ in major ways. The most important difference is that we employ supervised learning for T–F unit labeling using a set of pitch-based features, while their unit labeling approach involves no learning and uses two features. Our model uses two methods for segmentation in the low-frequency and high-frequency ranges, whereas their model uses cross-channel correlation for segmentation at all frequencies. These differences lead to substantially better performance in our system when dealing with reverberant speech, as reported in the next section.

V. Evaluation and Comparison

A. Corpus Generation

To simulate typical room acoustics, we use the image model which is commonly applied for efficient simulation of the acoustic properties of enclosures [2]. The basic idea of the

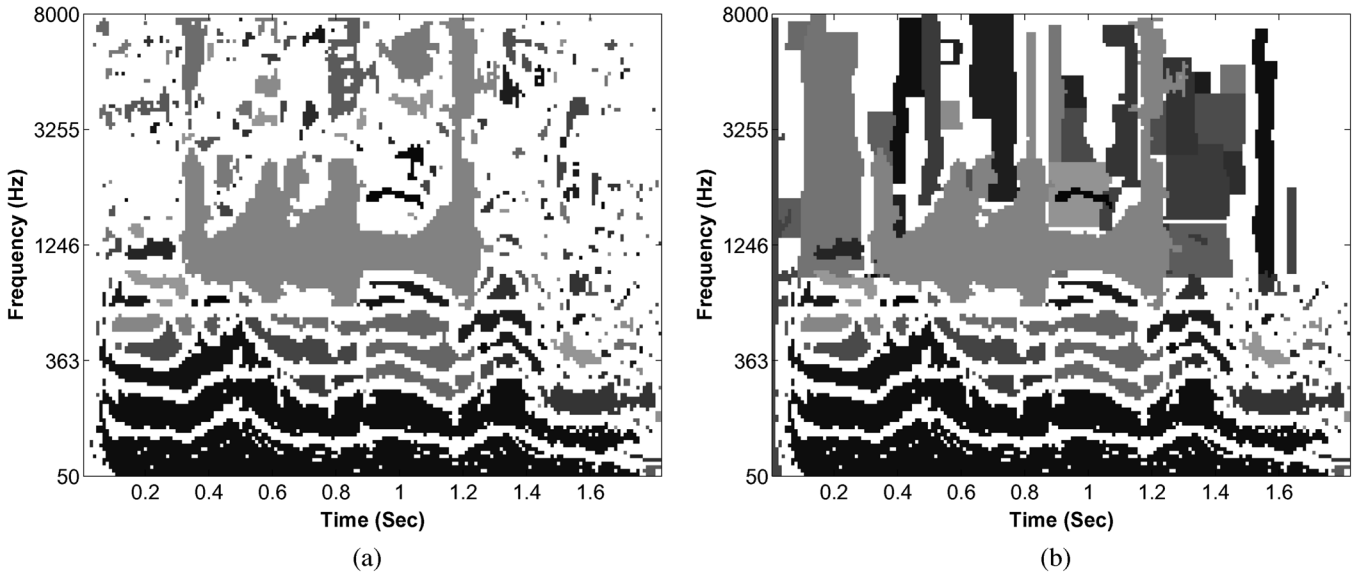


Fig. 3. Example of segmentation. (a) Segmentation using cross-channel correlation. (b) Segmentation using cross-channel correlation in low-frequency range and onset/offset analysis in high-frequency range. The input signal is the reverberant mixture of a voiced utterance and a pure tone in a simulated room whose $T_{60} = 0.3$ s. White indicates background. Regions of different gray levels indicate different T–F segments.

Allen–Berkley image model is that the room impulse response (RIR) can be represented as an infinite number of image sources that are created by reflecting an acoustic source in six room walls. In such a model, a pair of physical locations, corresponding to the source and the microphone, decide RIR in a fixed room. In order to simulate both convolutive and additive distortions, we specify the locations of the target and one interfering source and one more location for the microphone. More specifically, we start with anechoic target speech $s(t)$ and anechoic interference $n(t)$. We then generate a simulated room and randomly create a set, $\{\mathbf{r}_T, \mathbf{r}_I, \mathbf{r}_M\}$, representing locations of the target, the interference, and the microphone inside the room, respectively. From these locations, two RIRs are calculated by

$$h_T(t) = f(\mathbf{r}_T, \mathbf{r}_M) \tag{10}$$

and

$$h_I(t) = f(\mathbf{r}_I, \mathbf{r}_M). \tag{11}$$

$h_T(t)$ is the RIR corresponding to the recorded target at the microphone, and $h_I(t)$ corresponds to the recorded interference at the same microphone. Both $h_T(t)$ and $h_I(t)$ are causal finite-impulse response (FIR) filters. $f(\cdot)$ denotes the image model discussed above, which calculates the RIR with respect to the input location pair. Consequently, a reverberant mixture $r(t)$ is constructed by

$$r(t) = h_T(t) * s(t) + \alpha \cdot h_I(t) * n(t) \tag{12}$$

where "$*$" denotes convolution. We use α as a coefficient in order to set mixture SNR to 0 dB. The goal of our system is to segregate the reverberant target $h_T(t) * s(t)$ from the mixture $r(t)$.

In order to systematically evaluate the proposed system under different reverberant conditions, we simulate six acoustic

TABLE I
SETTINGS OF SIX ACOUSTIC ROOMS (L: LENGTH, W: WIDTH, H: HEIGHT)

Room No.	L × W × H (m)	Reflection Coeff.	T_{60} (s)
1	4 × 4 × 3	0.40	0.1
2	5 × 4 × 3	0.62	0.2
3	6 × 4 × 3	0.73	0.3
4	7 × 5 × 3	0.80	0.4
5	8 × 5 × 3	0.84	0.5
6	9 × 5 × 3	0.87	0.6

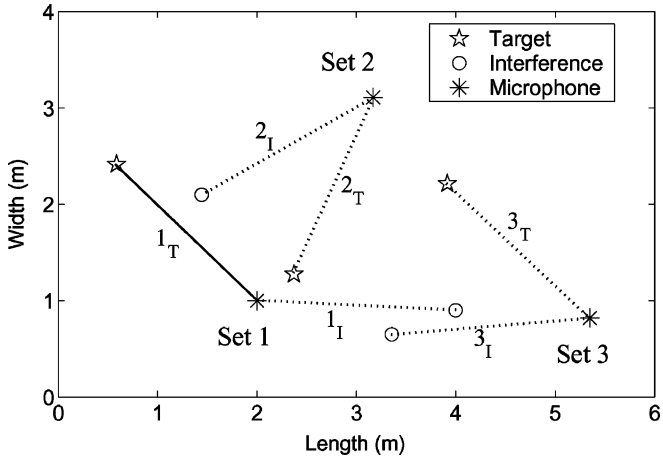


Fig. 4. Room configurations with three sets of locations $\{\mathbf{r}_T, \mathbf{r}_I, \mathbf{r}_M\}$ randomly created in an enclosure. Each line indicates the direct transmission path from the source to the microphone, corresponding to one RIR. For clarity, the example room is shown in two dimensions, though the simulations are performed in three dimensions. The solid line represents the RIR from which the inverse filter is estimated.

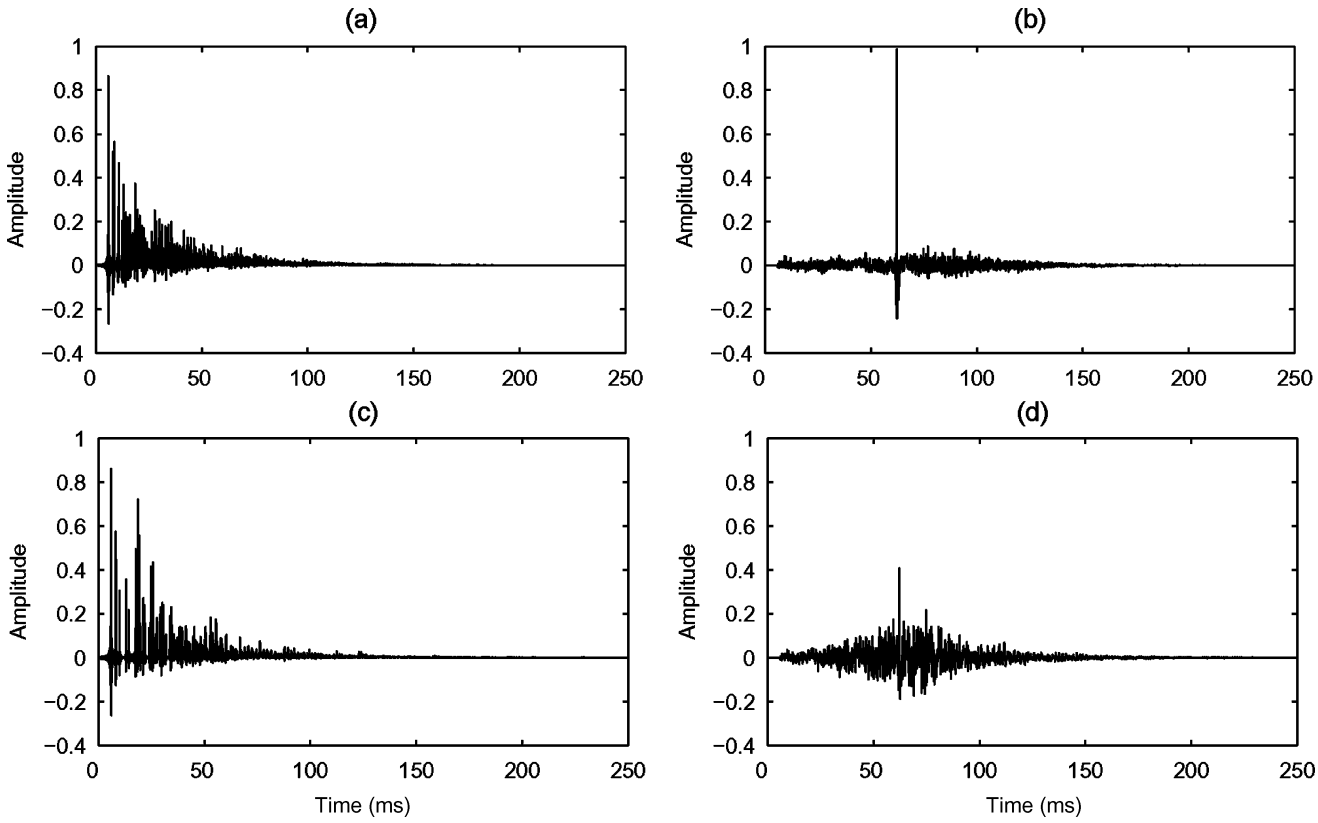


Fig. 5. Effects of inverse filtering on room impulse responses. (a) A RIR generated by the image model. The source and the microphone are at (4, 0.9, 1) and (2, 1, 1) respectively in Room 3, as listed in Table I. (b) The result of convolving the RIR in (a) with the estimated inverse filter. (c) A different RIR function in the same room but with the source location changed to (0.6, 2.4, 1). (d) The result of convolving the impulse response in (c) with the same estimated inverse filter.

rooms with different sizes and their reverberation times (T_{60}) range from 0.1 to 0.6 s in steps of 0.1 s. Table I shows detailed room specifications. In each room, we randomly create three sets of locations as mentioned above, resulting in three sets of $\{h_T(t), h_I(t)\}$ and three sets of reverberant mixtures created by (12). For example, Fig. 4 illustrates a simulated room with $T_{60} = 0.3$ s. The room size is $6 \times 4 \times 3$ m (length, width, height), but the figure only shows the first two dimensions for clarity. The pentagram, the circle and the asterisk display locations of $\{\mathbf{r}_T, \mathbf{r}_I, \mathbf{r}_M\}$ in each set respectively.

Our evaluation first uses Cooke's corpus [15], which contains 100 noisy utterances constructed by mixing ten anechoic voiced utterances (target speech) and ten different types of interference. In the aforementioned way, we generate a total of 1900 mixtures, with the original 100 mixtures in anechoic and $6 \times 3 \times 100$ mixtures in reverberant conditions. We further evaluate the proposed system using utterances from the TIMIT speech corpus [21] as target speech. Four speakers, two males and two females, are randomly selected from "DR1" through "DR4" dialect regions, respectively. For each speaker, we mix ten anechoic utterances with the same ten interferences to generate 1900 mixtures. The simulated rooms and source and microphone locations are the same as in the evaluation with Cooke's corpus.

B. Sensitivity of Inverse Filtering

As mentioned in Section I, one main problem with the inverse filtering approach is the sensitivity to even small changes in the acoustic environment. In other words, if an inverse filter is estimated from the same RIR used in segregation, i.e., matched inverse filtering, it is expected to enhance speech harmonicity; otherwise, it may further smear the harmonic structure. In this subsection, we quantitatively analyze such effects on Cooke's corpus under different conditions. It should not only clarify the problem but also offer a good connection to the results in Section V-D, where the proposed model is compared with an inverse filtering based system. Fig. 5 illustrates the effects of applying the same inverse filter to a matched RIR and a mismatched RIR with the source location moved. As can be seen in Fig. 5(b), the equalized response in the matched condition is much impulse-like, indicating the success of reverberation attenuation, while as shown in Fig. 5(d) the mismatched condition leads to further smearing.

We quantitatively evaluate the sensitivity of inverse filtering in terms of signal-to-reverberant energy ratio (SRR). SRR is an indicator of intelligibility of reverberant speech [30] and hence provides a measure of the effectiveness of inverse filtering. SRR is defined as

$$\mathrm{SRR} = 10 \log_{10} \left(\int_0^{t_1} p^2(t) dt \Big/ \int_{t_1}^{\infty} p^2(t) dt \right) \qquad (13)$$

where $p(t)$ is the instantaneous sound pressure of the RIR measured at time t, and t_1 is the arrival time of the first peak from the reflected impulses. A larger SRR value indicates better inverse filtering. Table II shows the SRR improvement after applying the inverse filter to RIR's in the six rooms in Table I.

TABLE II
SIGNAL-TO-REVERBERANT RATIO (SRR) CHANGE (IN dB) BY APPLYING AN ESTIMATED INVERSE FILTER TO EACH ROOM IMPULSE RESPONSE FUNCTION. THE MATCHED INVERSE FILTERING CONDITION IS SHOWN AS UNDERLINED BOLD

Room No.	1_T	1_I	2_T	2_I	3_T	3_I
1	-7.8	-6.8	-7.4	-6.0	-7.5	-8.4
2	-3.7	-2.8	-3.8	-4.5	-4.4	-2.7
3	**7.2**	-2.1	-0.2	-2.9	-2.8	-2.3
4	-1.5	-1.1	-0.3	-2.7	-0.8	-0.5
5	-2.1	-0.4	-2.2	-1.4	-1.4	-1.4
6	0.0	1.2	-0.3	0.2	1.3	-0.1

We have already created for each acoustic room six different RIRs from three sets of $\{\mathbf{r}_T, \mathbf{r}_I, \mathbf{r}_M\}$ described in Section V-A. These six RIRs are named $\{$"1_T", "1_I", "2_T", "2_I", "3_T", "3_I"$\}$ as presented in Fig. 4, where subscript t refers to the RIR from $\{\mathbf{r}_T, \mathbf{r}_M\}$ and i from $\{\mathbf{r}_I, \mathbf{r}_M\}$ in each set. The inverse filter is estimated to equalize the RIR ("1_T" in Room 3, the solid line shows the direct path in Fig. 4) using [22]. To examine the sensitivity of inverse filtering, this estimated inverse filter is used to convolve with all $6 \times 6 = 36$ RIRs and their resulting SRRs are calculated accordingly. It is evident from Table II that significant SRR improvement only occurs under the matched inverse filtering condition (shown as underlined bold). The SRR drops for almost all other cases, implying a further smearing effect caused by mismatched inverse filtering.

Our empirical results are in accordance with the observation reported in [40], where it is stated that inverse filtering increases the distortion when a response recorded at a different position is employed for dereverberation. Radlovic *et al.* [45] gave a theoretical analysis on the sensitivity of inverse filtering. In their paper, a quantitative distortion measure of frequency responses is used based on the difference between the two transfer functions from the source to the reference and to the displacement point, respectively. The measure calculated in simulation is in good agreement with their theoretical derivation and shows that small changes in the source or microphone position on the order of one-tenth of the acoustic wavelength cause significant degradations in the equalized room response. It is also pointed out that greater distortion is expected for high frequencies, indicating higher sensitivity to position changes in the high-frequency range.

C. SNR Results

Given that the computational objective of our segregation system is to identify T–F regions that are target dominant, we adopt the same SNR measure in [25] to assess the segregation performance using the resynthesized speech from the ideal binary mask as the ground truth. Equation (4) gives this measure. Considering that these harmonic features likely vary with changing acoustic environments, we evaluate the proposed system in three different scenarios which place different levels of demand on generalization.

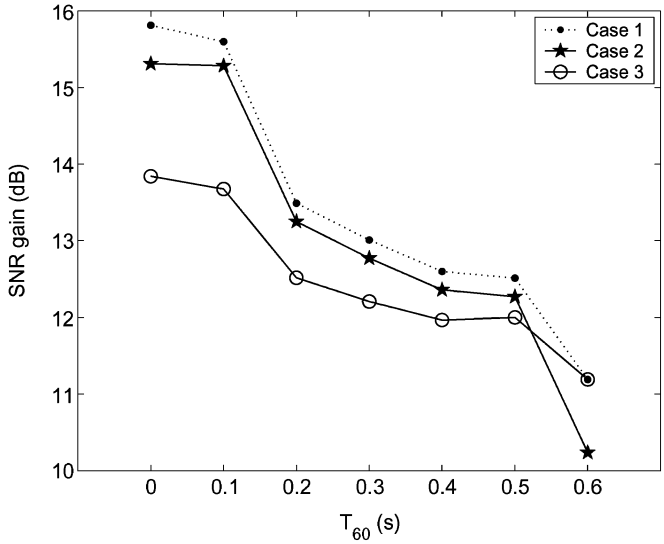


Fig. 6. Voiced speech segregation performance. SNR gain is measured under room conditions with T_{60} ranging from 0 to 0.6 s. Case 1: trained on each T_{60}. Case 2: trained on all T_{60}'s. Case 3: trained on $T_{60} = 0.6$ s only.

Case 1: Reverberation time is known. In this case, our evaluation is conducted within the same room but assesses the system's ability to generalize to different source/receiver locations. For example, we train on one set of 100 reverberant mixtures in Room 3 with $T_{60} = 0.3$ s and test the resulting system on the other two sets of mixtures in the same room in terms of SNR gain (the improvement over the initial SNR before segregation). The dotted line in Fig. 6 represents this case. The performance curve depicts the SNR gain of seven separate systems, each trained at a different T_{60}. This curve should represent the performance upper bound of our system in unknown reverberant conditions. The observed performance drop with increasing reverberation likely reflects the nature of the ascending difficulty of segregation. In other words, segregation in highly reverberant conditions is probably a harder task than in low reverberant conditions. Subjective tests reveal that human listeners' ability to separate competing voices degrades with increasing levels of reverberation [17].

Case 2: With unknown T_{60}, train on all different T_{60}'s. Specifically, we form a training corpus with a total of 700 reverberant mixtures by using the first set of mixtures in each room together with anechoic mixtures. The pentagram line in Fig. 6 shows the system performance in this case. This way of training gives a single system regardless of reverberant conditions and the performance is only about 0.5 dB worse on average compared to the known room case. A downside is that training now becomes computationally more expensive since the training set is seven times as large as training in a single room. On a 2.8-GHz PC with 1-GB memory, Case 2 needs 39.48 h for training—roughly seven times 5.51 h needed in Case 1.

Case 3: With unknown T_{60}, train on a single T_{60}. If we assume reverberation time is more likely above 0.3 s which is typical of rooms encountered in daily life [33], we can train at $T_{60} = 0.6$ s, the most reverberant condition because generalization to less reverberation may be better than the other way

around. The rationale here is to obtain the best possible classifier under the least favorable condition, often referred to as a MINIMAX solution [18], [54]. The SNR gain of this case is the circle line in Fig. 6. Some degradation is observed, but the system yields relatively good performance at high T_{60}'s.

Although the proposed system is designed to segregate reverberant speech, the above results suggest that our system also works well in the anechoic condition ($T_{60} = 0$ s). Using multiple features for estimating the grouping cue, our system shows a 13.8-dB SNR gain under the anechoic condition even when it is trained at $T_{60} = 0.6$ s. As a comparison, the segregation system by Hu and Wang, which is designed for and tested on voiced speech mixtures in anechoic conditions [25], produces an SNR gain of 12.9 dB on the same corpus. This indicates that our system performs a little better than the Hu–Wang system in the anechoic condition. If training is matched with the anechoic condition, our system achieves an SNR gain of 15.8 dB (see Fig. 6) which is significantly higher than that of their system.

To separate the contributions of the features and training strategies, we conduct experiments to compare classifiers trained on different features and using different objective functions. Fig. 7 gives the classification (unit labeling) performance in terms of SNR gain in different reverberant conditions. All classifiers are trained with one set of 100 reverberant mixtures in Room 6 with $T_{60} = 0.6$ s (as in Case 3). The proposed classifier—using six features and generalized MSE objective function—yields the top performance. Note that its SNR gain is significantly lower than the one presented in Fig. 6 because we measure the performance directly after the unit labeling stage, i.e., without performing segmentation and grouping. The classifier using just the two primary features $\{A(c,m,\tau_m), A_E(c,m,\tau_m)\}$ performs about 1.6 dB worse. This margin reflects the contribution of the other four features in encoding harmonicity under reverberant conditions. The classifier trained using the conventional MSE objective function performs 1 dB below the proposed one, showing the benefit of the generalized MSE objective function in MLP training.

We also compare alternative segmentation methods: cross-channel correlation alone, onset/offset analysis alone, and the proposed method which combines these two methods. All methods use the same training with one set of 100 reverberant mixtures in Room 6 with $T_{60} = 0.6$ s from Cooke's corpus. Fig. 8 shows the system performance in terms of SNR gain in different reverberant conditions. The proposed segmentation using cross-channel correlation in low frequency and onset/offset analysis in high frequency yields the best performance. The method using just cross-channel correlation—$C(c,m)$ in low frequency and $C_E(c,m)$ in high frequency—performs about 0.4 dB worse (0.6 dB worse if measured just in the high-frequency range). This difference indicates the utility of onset/offset analysis in the high-frequency range. The method using onset/offset analysis across all frequencies gives the worst performance. This method forms segments by matching onset and offset fronts and segment boundaries tend to be block-like, missing detailed segment shapes. Although it performs a little better than the cross-channel correlation method in the high-frequency range, it underperforms the latter in the low-frequency range. Note that the good performance reported in [27] is based on a region-level measure not an SNR measure.

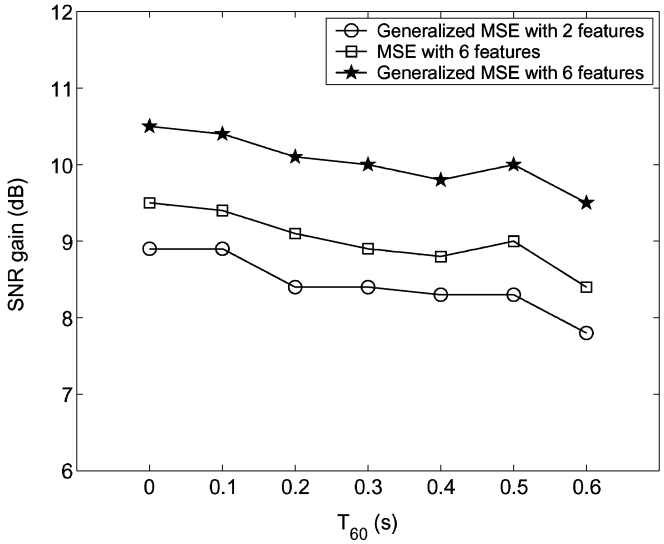


Fig. 7. Comparison of SNR gains among classifiers using different feature sets and different objective functions. Segmentation and grouping are not performed in this comparison.

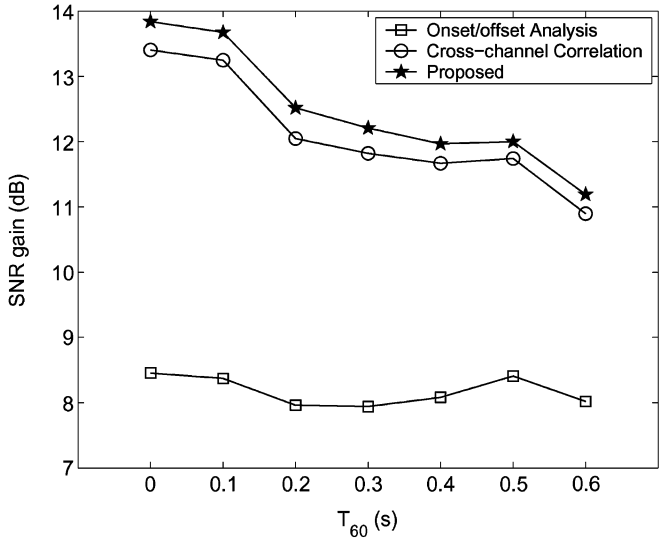


Fig. 8. Comparison of SNR gain between different segmentation methods.

D. Comparison With Roman–Wang Model

In this subsection, we use Cooke's corpus to compare the performance of our proposed system to that of the inverse filtering based approach by Roman and Wang [48]. In their system, an inverse filter is first estimated by maximizing the kurtosis of the inverse-filtered linear prediction residual of the reverberant speech from the target location in the absence of interference [22], [61]. Then, the obtained inverse filter is applied to the reverberant mixture consisting of both the reverberant target and the reverberant interference. In order to make a fair comparison between the Roman–Wang and the proposed system, we use the same subset of reverberant mixtures for learning. Specifically, in the Roman–Wang system, we use the same inverse filter (as used in the above evaluation) that is estimated from the RIR of

TABLE III
COMPARISONS OF SNR GAIN (IN dB) BETWEEN THE PROPOSED SYSTEM AND THE ROMAN–WANG SYSTEM. MATCHED TRAINING CONDITIONS ARE SHOWN AS UNDERLINED BOLD

	ROMAN-WANG						PROPOSED					
$T_{60}(s)$	0.1	0.2	0.3	0.4	0.5	0.6	0.1	0.2	0.3	0.4	0.5	0.6
Set 1	10.24	9.58	**10.07**	8.36	8.92	6.69	14.44	13.52	**13.48**	11.71	12.59	8.92
Set 2	10.86	9.36	9.74	9.40	7.54	7.88	15.37	13.26	13.37	12.08	10.09	9.59
Set 3	10.59	8.04	7.99	9.15	9.51	7.16	14.84	11.53	12.18	11.65	12.16	8.93
Average	10.56	8.99	9.27	8.97	8.66	7.24	14.88	12.77	13.01	11.81	11.61	9.15

"1_T" in Room 3. In the proposed system, MLP learning also takes place on the reverberant mixtures generated by the first set of $\{\mathbf{r}_T, \mathbf{r}_I, \mathbf{r}_M\}$ in Room 3. Note that, like in our system, *a priori* pitch contours are used in their system in order to generate results free from pitch detection errors. Table III summarizes the SNR gain evaluation. Each number in the table presents the average SNR gain on reverberant mixtures generated by a particular set of $\{\mathbf{r}_T, \mathbf{r}_I, \mathbf{r}_M\}$ in one of the rooms. The two underlined numbers correspond to matched training conditions as stated above. Their difference shows the advantage of the proposed system when training and testing on the same set of mixtures. The number at the bottom of each column is the average of the three sets, which provides the average SNR gain in the same room. The SNR gains under the anechoic condition (not shown in the table) are 10.37 and 14.87 dB for the Roman–Wang and the proposed system, respectively. As can be observed in Table III, the proposed system achieves significantly higher SNR gains across all different reverberation times than the Roman–Wang system. The overall 3.4-dB improvement is mainly brought about by accurate unit labeling in high frequency: their system cannot reliably handle unresolved harmonics in those regions for reverberant signals. Through MLP training, the T–F units in both low and high frequencies can be labeled in our system. Note that this improvement partly results from segmentation using multiscale onset/offset analysis in high frequency (see Fig. 8).

In [48], Roman and Wang reported a 1.3-dB difference between the systems with and without using inverse filtering when $T_{60} = 0.35$ s. Without inverse filtering as a preprocessing stage, their system is quite similar to the Hu–Wang system [25], which was discussed earlier. This implies that our system performs much better than the Hu–Wang system when it is applied to reverberant mixtures. In other words, although there are reasons to expect that pitch-based grouping may not be very sensitive to reverberation, such separation algorithms unlikely perform well without dealing with the issue of reverberation.

E. Evaluation on TIMIT

Here we evaluate how the proposed system generalizes to new speakers and utterances. As described in Section V-A, four speakers, two males and two females, are randomly selected from the TIMIT database. We label the four speakers as: TrM, TrF, TeM, and TeF, where "M" stands for male, "F" female, and "Tr" and "Te" label speakers included in training and testing, respectively. In order to compensate for the discrepancies between male and female speakers, we train on the first set of mixtures at Room 6 from both genders. The training corpus contains 200 mixtures, one half from TrM and the other half from TrF. The resulting system may be called speaker-independent (SI). For each speaker, we also train a system on the first set of mixtures at Room 6 from that speaker only. These systems are called speaker-dependent, or SD. The motivation of training a system for each speaker is to evaluate performance in the matched training scenario, which offers a reference for performance analysis. The SNR difference between SI and SD indicates degradation due to unmatched training. Since utterances from the TIMIT database contain unvoiced speech while our system deals with only voiced speech, we calculate the SNR at voiced speech frames only. Fig. 9 shows the SNR comparison between SI and SD for each of the four speakers. Note that, the SNR performance in Fig. 9 is lower than that of the previous experiment due to the use of TIMIT sentences whose spectra significantly overlap with those of interferences. The SNR gain at each T_{60} condition is averaged over all three sets of mixtures in that condition, the same as the last row in Table III. TeM and TeF are the test cases using entirely different speakers, within which 70% of the sentences are new. The SI curve is not much lower than the SD curve for these two speakers, demonstrating that our system generalizes well to both unforseen speakers and utterances. Note that degradation also exists in TrM and TrF, albeit smaller than those in TeM and TeF. This degradation arises because the training corpus contains two speakers rather than a single matching one. Table IV gives numeric results of average SNR degradation across all T_{60}'s for different speakers. The maximum degradation is 0.67 dB for TeF, which is fairly small compared to SNR performance variations for different room conditions.

VI. ROBUSTNESS ANALYSIS

The feature-based learning in our proposed system shows good generalization ability to various reverberant conditions as shown in Section V. For example, the system trained at Room 6 with $T_{60} = 0.6$ s performs well in other room conditions. Furthermore, the changes of source and microphone locations within a room little affect our system performance. To understand the surprisingly robust performance, this section provides an analysis at the feature level.

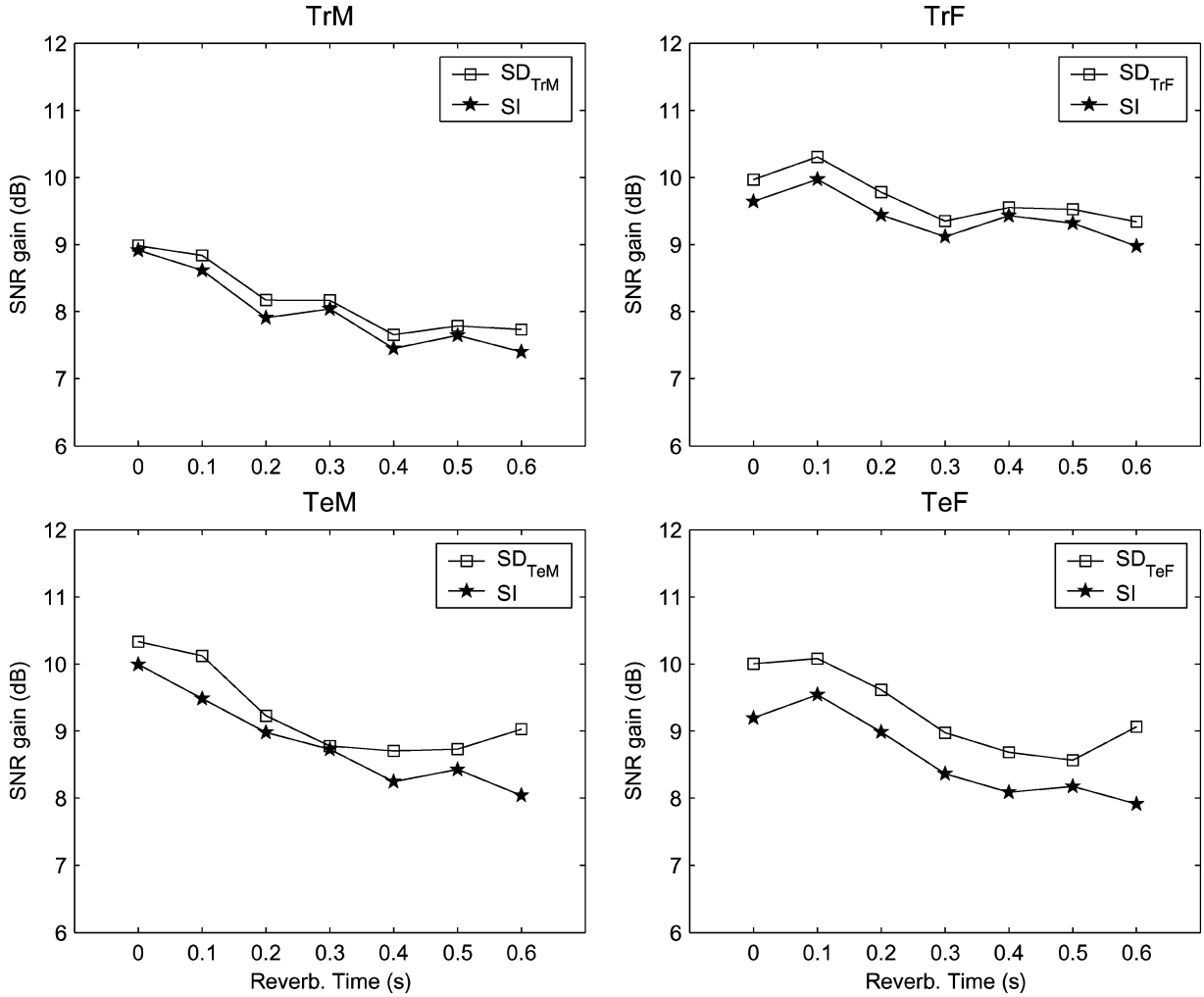


Fig. 9. Comparison of SNR gain between speaker-dependent (SD) and speaker-independent (SI) cases under room conditions with T_{60} ranging from 0 to 0.6 s. TrM and TrF are speakers included in training and TeM and TeF in testing.

TABLE IV
AVERAGE SNR DEGRADATION (IN dB) IN THE SPEAKER-INDEPENDENT SYSTEM RELATIVE TO SPEAKER-DEPENDENT SYSTEMS

TrM	TrF	TeM	TeF
0.19	0.28	0.42	0.67

Basically, the proposed system learns to distinguish between target dominant (class 1) and interference dominant (class 0) T–F units. Therefore, we reformulate the segregation problem into two-class classification. Intuitively, if features are robust, we expect that feature distributions in different reverberant conditions are close to each other. Hence, the distance between different feature sets can be a quantitative measure of feature robustness. From another perspective, the distance between class 0 and class 1 subsets within one feature set describes the classification complexity (or data separability) [52]. Therefore, the role of a distance measure is twofold: it models the feature variations in different reverberant conditions and it compares data separability in those conditions at the same time. An example is the following: Let $D(\cdot,\cdot)$ be a distance measure. Consider two feature sets $\{F_1, F_2\}$, each having two subsets corresponding to class 0 and class 1. A set of these subsets is constructed as

$$\mathcal{F}_2 = \{F_{1,0}, F_{1,1}, F_{2,0}, F_{2,1}\} \tag{14}$$

where $F_{i,j}$ indicates the subset of class j in F_i. On one hand, $D(F_{1,0}, F_{2,0})$ and $D(F_{1,1}, F_{2,1})$ measure the similarity between F_1 and F_2 in the two classes. When both values are small, good generalization from one set to the other is expected. On the other hand, $D(F_{1,0}, F_{1,1})$ and $D(F_{2,0}, F_{2,1})$ measure the separability of the feature sets, which relates to their performance upper bound discussed in Section V-C.

We use the Constrained Minimum (CM) distance [53] as the distance measure in this study. The CM distance is not only a metric, but also capable of measuring classification complexity because it is computed by comparing summary statistics of the data sets. According to [53], the CM distance is derived using the geometrical interpretation of the distribution and is of Mahalanobis type as

$$d_{CM}(D_1, D_2|S)^2 = (\theta_1 - \theta_2)^T cov^{-1}[S](\theta_1 - \theta_2) \tag{15}$$

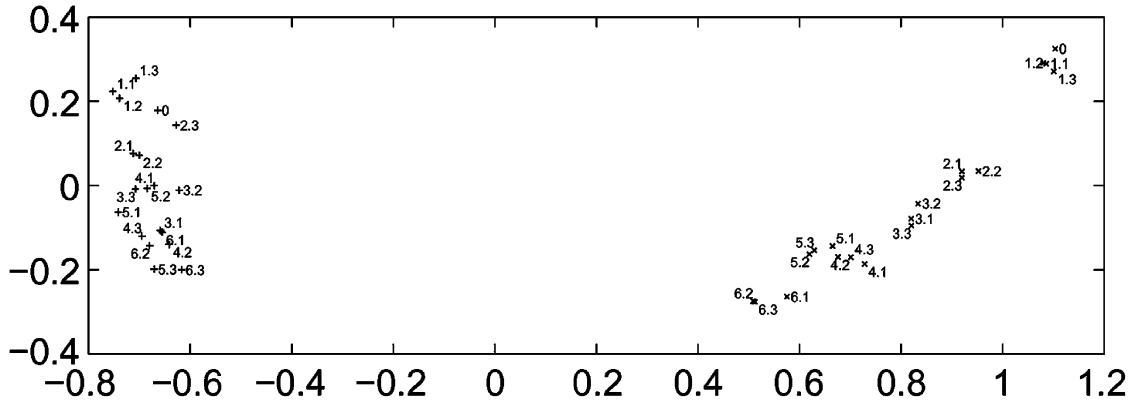


Fig. 10. Visualization of the constrained minimum distance. Plus marks stand for class 0 subsets and cross marks for class 1 subsets. Each mark is labeled in the form of "x.y" where x is the room index and y is the set index. "0" labels the anechoic situation.

where D_1 and D_2 are two data sets and their means are θ_1 and θ_2. S represents the true underlying feature distribution function. Since it is an unknown term, we estimate it from D_1 and D_2 and calculate its covariance matrix $cov[S]$ thereafter. In the case of N available feature sets in different reverberation conditions, we have

$$\mathcal{F}_N = \{F_{1,0}, F_{1,1}, F_{2,0}, F_{2,1}, \ldots, F_{N,0}, F_{N,1}\}. \tag{16}$$

The CM distance between every pair of the subsets in $\mathcal{F}_N$ is then calculated. To visualize the relationship between these subsets, we reconstruct 2-D spatial locations from their CM distances using the Metric Multidimensional Scaling (MDS) technique [16], which transforms a distance matrix into a set of coordinates such that the Euclidean distances derived from these coordinates approximate as well as possible original distances.

Fig. 10 presents the 2-D visualization of the CM distance of feature subsets. The 38 plotted subsets are derived from 19 feature sets, which correspond to the 19 sets of 100 reverberant mixtures generated from the voiced corpus described in Section V-A. Different reverberant conditions are denoted as $x.y$ where $x \in \{1, \ldots, 6\}$ represents room index and $y \in \{1, 2, 3\}$ is set index. As shown in the figure, the features from each room tend to cluster together, suggesting a strong similarity between them. This indicates that our features are robust to source/microphone location changes within a room. Based on the observation that features in rooms with close T_{60}'s also have relatively short distances, we can conclude that the features are robust to different reverberant rooms when these rooms have close T_{60}'s. On the right side ("$\times$") of Fig. 10, there is a clear pattern of position change with the change of T_{60}. Such a trend is not as prominent on the left side ("$+$") because our features are pitch-based and background T–F units may not be sensitive to such features. However, it will not affect the comparison of classification complexity as the changes on the left side are smaller than those on the right side and can therefore be ignored. Classification complexity can be compared by measuring the distance between the $+$ and the $\times$ mark of the same label, which indicates that the two subsets come from the same feature set. Fig. 10 suggests that the classification in low T_{60}'s is easier because of its relatively large distance while the classification in the most reverberant situation (i.e., $T_{60} = 0.6$ s) is the most difficult. This is consistent with the results in Fig. 6.

VII. Discussion

A key problem in reverberant speech separation is smeared harmonicity, which has negative impact on harmonic cues and results in significant performance degradation in previous CASA systems (e.g., [25]). The approach of inverting reverberant effects is sensitive to specific configurations, although it achieves good performance in matched configuration. In this paper, a set of six pitch-based features is extracted and these features incorporate information of both filter responses and their envelopes. Therefore, unlike [25] and [26], unit labeling can be handled together. The harmonic index and the deviation from the nearest harmonic, first proposed in [24], are demonstrated in this study to be effective supplementary features in modeling harmonicity under reverberant conditions—when excluding those features in the feature set, the overall performance has a significant drop in our experiment. MLP provides a way to combine these features into a unified grouping cue. Significantly different from other CASA systems, our supervised learning approach produces substantially better performance and generalizes well to different reverberant conditions. It is worth emphasizing that the proposed system also generalizes well to unseen speakers and utterances.

In the Bayesian framework, MLP may be viewed as an optimal classifier that discriminates target-dominant units from those belonging to the interference. One common problem of designing a classifier is the uncertainty in *a priori* class probability. Although the current study controls the SNR of all mixtures at 0 dB in both training and test phases, the above problem is of concern when SNR varies. For example, when testing on 10-dB mixtures, the system tends to label fewer target units than it should, indicating a bias towards the interference class. A common practice to increase system robustness against uncertain priors is training the classifier over a data set with the least biased priors [5], [34]. To use a training corpus of 0-dB mixtures is consistent with the above idea. However, the use of a training set with equal priors represents a solution that is unbiased towards any priors, but it does not theoretically imply robustness against other priors [1]. Some research provides clues [38], [46],

[50] on how to work with different SNRs. An adaptive solution uses incoming information to reduce the uncertainty and improve the classifier. More specifically, the SNR of the incoming mixture can be estimated to infer real priors and then the classifier can be adapted according to estimated priors [28].

Determination of pitch is a fundamental problem in CASA and reliable pitch estimation is critical for applying harmonic grouping. Although pitch may be a relatively robust feature to reverberation [60], few pitch determination algorithms are developed in both noisy and reverberant conditions. In this paper, we use *a priori* pitch, calculated from reverberant target speech before mixing, and future study needs to address the pitch determination problem in room reverberation.

Segregation of unvoiced speech, not dealt with in this paper, is an important and little studied problem in CASA presumably because of the difficulty of the task. Unvoiced speech lacks harmonicity and is more susceptible to interference due to its relatively weak energy compared to voiced speech. Under anechoic conditions, acoustic-phonetic features have been recently used to segregate unvoiced speech from nonspeech noise [28]. When interference is relatively stationary, spectral subtraction can be used to remove noise within intervals of unvoiced speech with noise estimation from intervals of segregated voiced speech [29]. No study has been performed on unvoiced speech segregation in reverberant conditions, and this is a topic that requires future research.

In summary, we have proposed a system capable of segregating reverberant target speech. Two novel ideas are employed. First, a supervised learning approach establishes the mapping from a set of pitch-based features to a grouping cue and a new objective function is proposed to maximize SNR. Second, a multiscale onset/offset analysis is employed to form reliable segments in the high-frequency range.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their useful suggestions/criticisms.

REFERENCES

[1] R. Alaiz-Rodríguez, A. Guerrero-Curieses, and J. Cid-Sueiro, "Minimax classifiers based on neural networks," *Pattern Recognition*, vol. 38, pp. 29–39, 2005.

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[3] 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, pp. 480–492, 2006.

[4] F. Bach and M. Jordan, "Blind one-microphone speech separation: A spectral learning approach," in *Proc. NIPS*, 2004, pp. 65–72.

[5] R. Barandela, J. Sanchez, V. García, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition*, vol. 36, pp. 849–851, 2003.

[6] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[7] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (Version 4.3.14)." 2005 [Online]. Available: http://www.fon.hum.uva.nl/praat

[8] S. F. Boll, "Supression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[9] M. S. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proc. IEEE ICASSP*, 1998, pp. 3613–3616.

[10] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT, 1990.

[11] G. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, pp. 297–336, 1994.

[12] G. J. Brown and K. J. Palomäki, "Reverberation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley/IEEE Press, 2006, pp. 209–250.

[13] 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, pp. 4007–4018, 2006.

[14] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[15] M. P. Cooke, *Modeling Auditory Processing and Organization*. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[16] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*. New York: Chapman & Hall, 2001.

[17] J. F. Culling, K. I. Hodder, and C. Y. Toh, "Effects of reverberation on perceptual segregation of competing voices," *J. Acoust. Soc. Amer.*, vol. 114, pp. 2871–2876, 2003.

[18] R. O. Duda, P. E. Hart, and D. Stock, *Pattern Classification*. New York: Wiley, 2001.

[19] 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. 32, pp. 1109–1121, 1984.

[20] Y. Ephraim and H. L. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus, "CDROM 1993 [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[22] B. W. Gillespie, H. S. Malvar, and D. A. F. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE ICASSP*, 2001, pp. 3701–3704.

[23] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*. Boston, MA: PWS Publishing, 1996.

[24] G. Hu, "Monaural speech organization and segregation," Ph.D. dissertation, Biophys. Program, The Ohio State Univ., Columbus, 2006.

[25] 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.

[26] G. Hu and D. L. Wang, "An auditory scene analysis approach to monaural speech segregation," in *Topics in Acoustic Echo and Noise Control*, E. Hansler and G. Schmidt, Eds. New York: Springer, 2006, pp. 485–515.

[27] 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.

[28] G. Hu and D. L. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Amer.*, vol. 124, pp. 1306–1319, 2008.

[29] K. Hu, P. Divenyi, D. Ellis, Z. Jin, B. G. Shinn-Cunningham, and D. L. Wang, "Preliminary intelligibility tests of a monaural speech segregation system," in *Proc. SAPA*, 2008, pp. 11–16.

[30] J. J. Jetzt, "Critical distance measurement of rooms from the sound energy spectral response," *J. Acoust. Soc. Amer.*, vol. 65, pp. 1204–1211, 1979.

[31] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," in *Proc. IEEE ICASSP*, 2007, pp. 921–924.

[32] M. Kukar and I. Kononenko, "Cost-sensitive learning with neural networks," in *Proc. Eur. Conf. Artif. Intell.*, 1998, pp. 445–449.

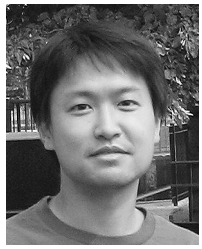
[33] H. Kuttruff, *Room Acoustics*. New York: Taylor & Francis, 2000.

[34] S. Lawrence, I. Burns, A. Back, A. Tsoi, and C. L. Giles, "Neural network classification and unequal prior class probabilities," in *Tricks of the Trade, Lecture Notes in Computer Science State-of-the-Art Surveys*, G. Orr, K. R. Müller, and R. Caruana, Eds. Berlin, Germany: Springer, 1998, pp. 299–314.

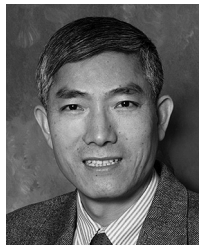
[35] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, pp. 1673–1682, 2008.

[36] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, pp. 197–210, 1978.

[37] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Amer.*, vol. 106, pp. 1633–1654, 1999.
[38] V. Matousek, P. Mautner, J. Ocelíková, and P. Sojka, Eds., in *Proc. Text, Speech, Dialogue—Second Int. Workshop, TSD'99, Plzen, Czech Republic, September 1999*, 1999, vol. 1692, Lecture Notes in Computer Science, Springer.
[39] R. Meddis, "Simulation of auditory-neural transduction: Further studies," *J. Acoust. Soc. Amer.*, vol. 83, pp. 1056–1063, 1988.
[40] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *J. Sound Vibr.*, vol. 102, pp. 217–228, 1985.
[41] H. Ney, "On the probabilistic interpretation of neural network classifiers and discriminative training criteria," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 2, pp. 107–119, Feb. 1995.
[42] K. J. Palomäki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Commun.*, vol. 43, pp. 361–378, 2004.
[43] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An Efficient Auditory Filterbank Based on the Gammatone Function," Appl. Psychol. Unit, Cambridge, U.K., APU Rep. 2341, 1988.
[44] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft masking filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.
[45] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: Robustness results," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 311–319, May 2000.
[46] M. D. Richard and R. P. Lippmann, "Neural network classifiers estimate bayesian a posteriori probabilities," *Neural Comput.*, vol. 3, pp. 461–483, 1991.
[47] N. Roman, S. Srinivasan, and D. L. Wang, "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4040–4051, 2006.
[48] N. Roman and D. L. Wang, "Pitch-based monaural segregation of reverberant speech," *J. Acoust. Soc. Amer.*, vol. 120, pp. 458–469, 2006.
[49] S. T. Roweis, "One microphone source separation," in *Proc. NIPS*, 2000, pp. 793–799.
[50] M. Saerens, P. Latinne, and C. Decaestecker, "Adjusting a classifier for new *a priori* probabilities: A simple procedure," *Neural Comput.*, vol. 14, pp. 21–41, 2002.
[51] M. Sayles, B. Schouten, N. J. Ingham, and I. M. Winter, "The effect of reverberation on the temporal representation of the f0 of frequency swept harmonic complexes in the ventral cochlear nucleus," in *Hearing: From Sensory Processing to Perception*, B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey, Eds. Berlin, Germany: Springer, 2007, pp. 35–42.
[52] S. Singh, "Multisolution estimates of classification complexity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1534–1539, Dec. 2003.
[53] N. Tatti, "Distances between data sets based on summary statistics," *J. Mach. Learn. Res.*, vol. 8, pp. 131–154, 2007.
[54] H. L. V. Trees, *Detection, Estimation and Modulation Theory*. New York: Wiley, 1968.
[55] 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: Kluwer, 2005, pp. 181–197.
[56] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 684–697, May 1999.
[57] , D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.
[58] M. Weinraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Dept. of Elect. Eng., Stanford Univ., Stanford, CA, 1985.
[59] R. Weiss and D. Ellis, "Monaural speech separation using source-adapted models," in *Proc. IEEE WASPAA*, 2007, pp. 114–117.
[60] M. Wu and D. L. Wang, "A pitch-based method for the estimation of short reverberation time," *Acta Acustica United With Acustica*, vol. 92, pp. 337–339, 2006.
[61] M. Wu and D. L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 774–784, May 2006.



Zhaozhang Jin (S'06) received the B.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, and the M.S. degree in computer science and engineering from The Ohio State University, Columbus, where he is currently pursuing the Ph.D. degree. His research interests include signal processing, machine learning, and computational auditory scene analysis.



DeLiang Wang (M'90–SM'01–F'04) received the B.S. and M.S. degrees in computer science from Peking (Beijing) University, Beijing, China, in 1983 and 1986, respectively, and the Ph.D. degree in computer science from the University of Southern California, Los Angeles, in 1991.

From July 1986 to December 1987, he was with the Institute of Computing Technology, Academia Sinica, Beijing. Since 1991, he has been with the Department of Computer Science and Engineering and the Center for Cognitive Science at The Ohio State University, Columbus, where he is currently a Professor. From October 1998 to September 1999, he was a Visiting Scholar in the Department of Psychology at Harvard University, Cambridge, MA. From October 2006 to June 2007, he was a Visiting Scholar at Oticon A/S, Denmark. His research interests include machine perception and neurodynamics.

Dr. Wang received the National Science Foundation Research Initiation Award in 1992, the Office of Naval Research Young Investigator Award in 1996, and the Helmholtz Award from the International Neural Network Society in 2008. He also received the 2005 Outstanding Paper Award from the IEEE TRANSACTIONS ON NEURAL NETWORKS.