Reverberant Speech Segregation Based on Multipitch Tracking and Classification

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

***Abstract*—Room reverberation creates a major challenge to speech segregation. We propose a computational auditory scene analysis approach to monaural segregation of reverberant voiced speech, which performs multipitch tracking of reverberant mixtures and supervised classification. Speech and nonspeech models are separately trained, and each learns to map from a set of pitch-based features to a grouping cue which encodes the posterior probability of a time–frequency (T-F) unit being dominated by the source with the given pitch estimate. Because interference may be either speech or nonspeech, a likelihood ratio test selects the correct model for labeling corresponding T-F units. Experimental results show that the proposed system performs robustly in different types of interference and various reverberant conditions, and has a significant advantage over existing systems.**

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

I. Introduction

SPEECH segregation in reverberant environments is a very challenging problem. A monaural (one-microphone) solution is highly desirable in many important applications, e.g., as a frontend for automatic speech recognition and hearing aid design in noisy backgrounds. Numerous methods have been developed for monaural speech enhancement [21]. These methods assume stationary or quasi-stationary interference and thus have intrinsic limitations in dealing with a general acoustic background. Model-based approaches have been proposed to perform monaural segregation. For example, Roweis [29] trained a factorial hidden Markov model for computing a time–frequency (T-F) mask for segregation. Bach and Jordan [3] proposed a spectral learning approach based on parameterized affinity matrices and segmentation in a T-F plane. Radfar and Dansereau [26] used a composite source model followed by a soft mask filter based on minimum mean square error for separating the underlying sources. However, none of these methods have been tested in reverberant conditions.

Computational auditory scene analysis (CASA) [35] aims to segregate a mixture signal into different streams based on perceptual principles of auditory scene analysis [6]. CASA systems tend to estimate the ideal binary mask (IBM) [33], where a value of 1 in the mask indicates that the target energy is stronger than the interference energy and 0 otherwise. IBM-segregated speech results in dramatic improvement of intelligibility in noise for both normal-hearing and hearing-impaired listeners [2], [7], [20], [36].

Few studies have addressed the monaural speech segregation problem in room reverberation. Roman and Wang [28] applied inverse filtering to partially counteract the smearing effect of reverberation on harmonic structure before segregation. However, the inverse filter is very sensitive to room configuration [17], [27]. Our previous work [17] developed a supervised learning approach to classify harmonic cues in order to achieve robust segregation performance against reverberant effects. However, this segregation system assumes the ground truth pitch of target speech.

This paper proposes a segregation system for reverberant speech by extending the above supervised classification approach in conjunction with detected pitch in reverberant mixtures [18]. We find that the classification approach continues to yield good performance when pitch-based features are extracted from estimated pitch, indicating good generalization. We further propose a novel unit labeling strategy for the time frames in which an interference pitch is also detected. Specifically, we train a multilayer perceptron (MLP) for target speech and a second MLP to model a variety of periodic interference. Because the MLP output estimates the posterior probability of the modeled source, a labeling criterion that compares the probabilities of the two underlying sources is expected to perform more reliably than one based on the posterior probability of only the target source. Here, we devise a likelihood ratio test to select the correct MLP model for the interference.

The paper is organized as follows. The Section II presents an overview of the proposed system. Sections III–V describe the segregation system stage by stage. Experimental results and comparisons are provided in Section VI. We discuss related issues and conclude the paper in Section VII.

II. System Overview

The proposed system has four computational stages. The first stage applies a recent multipitch tracking algorithm [18] to detect pitch contours for both target and interfering sources in reverberation. The algorithm uses an auditory filterbank to de-

Manuscript received June 13, 2010; revised December 15, 2010; accepted March 20, 2011. Date of publication April 05, 2011; date of current version August 19, 2011. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-08-1-0155 and in part by the National Science Foundation (NSF) under Grant IIS-0534707. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hui Jiang.

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. 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).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2134086

compose the input signal into the T-F domain and selects reliable channels to derive pitch scores for different pitch states. A hidden Markov model (HMM) then integrates these pitch scores and searches for the optimal pitch state sequence. This algorithm produces reliable pitch contours in noisy and reverberant conditions, laying a foundation for our pitch-based segregation approach.

In the second stage, we extract in each T-F unit a set of pitch-based features with respect to each detected pitch. These features are based on auditory filter responses and response envelopes to account for both resolved and unresolved harmonics. Section III gives the detail of multipitch tracking and feature extraction.

The next stage labels T-F units using trained MLP. As suggested in [17], the MLP is trained as a classifier in a cost sensitive way in order to maximize the signal-to-noise ratio (SNR) performance for the segregation task. The output of the MLP can be interpreted as the posterior probability of a T-F unit dominated by the source with the corresponding pitch period. Two MLPs are trained, one for speech and the other for nonspeech signals. Two scenarios are considered in the labeling procedure: 1) when the current time frame contains only one pitch, a simple labeling criterion is utilized; 2) when two pitch periods are detected in one frame, a comparative method is used to decide which harmonic source more likely dominates in the T-F unit. The second scenario further considers two hypotheses: speech interference and nonspeech interference. The different hypotheses call for different combinations of MLP models which lead to different labeling results. Finally, a likelihood ratio test is performed to choose the hypothesis that fits the data better. The proposed labeling procedure is described in Section IV.

The segmentation and grouping stage, described in Section V, refines the labeling results and generates the target and background stream. To obtain reliable segments in reverberant conditions, cross-channel correlation cues [34] are used in low-frequency channels and onset–offset cues [13] in high-frequency channels. These segments are subsequently grouped into the two streams resulting in an estimated IBM, which can then be used to resynthesize the segregated target speech.

III. Multipitch Tracking and Feature Extraction

The mixture signal $x(t)$ is first passed through a fourth-order gammatone filterbank [25] with 128 channels for time–frequency analysis. The center frequencies are quasi-logarithmically spaced from 50 to 8000 Hz. The response $x(c,t)$ of a filter channel c is further transduced by the Meddis hair cell model [22] to produce firing patterns in the simulated auditory nerve, denoted by $h(c,t)$. The output $h(c,t)$ is then divided into 20-ms time frames with 10-ms overlapping between consecutive frames. The resulting T-F representation is known as a cochleagram and further implementation details can be found in [35, Ch. 1]. In the following discussion, we use u_{cm} to denote a T-F unit in the cochleagram at time frame m and frequency channel c.

The normalized correlogram $A(c,m,\tau)$ is then calculated for u_{cm} with a time delay τ by the following autocorrelation function:

$$A(c,m,\tau)=\frac{\sum\limits_{n=-N/2}^{N/2} h\left(\frac{c,mN}{2+n}\right) h\left(\frac{c,mN}{2+n+\tau}\right)}{\sqrt{\sum\limits_{n=-N/2}^{N/2} h^2\left(\frac{c,mN}{2+n}\right)}\sqrt{\sum\limits_{n=-N/2}^{N/2} h^2\left(\frac{c,mN}{2+n+\tau}\right)}} \tag{1}$$

where N is the frame size. The range for τ should include the plausible pitch range [32,200] in samples. For the sampling frequency of 16 kHz, N is equal to 320 samples and the above pitch range translates to [80,500] in Hz. The denominator in (1) normalizes the autocorrelation value to $[0,1]$ (due to the nonnegative output of the Meddis model). To effectively capture unresolved harmonics, we also calculate the correlogram $A_E(c,m,\tau)$ from the envelope of the hair cell response $h_E(c,t)$ using (1).

Multipitch tracking [18] is performed here to detect pitch contours for both target and interfering sources in the reverberant mixture. Based on the correlogram $A(c,m,\tau)$, pitch scores in each time frame are derived under the three pitch state spaces, namely zero-pitch, one-pitch, and two-pitch hypotheses. Under zero-pitch hypothesis, we detect and assign relatively low scores for silence, unvoiced speech or noise. Under the one-pitch hypothesis, a weighted summary correlogram over the set of reliable channels is computed for all possible pitch periods. Under the two-pitch hypothesis, the notion of IBM is employed to divide the selected channels into two groups, each corresponding to one of the underlying pitch periods. The pitch strength from each group is calculated and later combined into a single score for all pairs of plausible pitch periods. In the tracking stage, an HMM treats those pitch scores in a probabilistic form and searches over all the three pitch hypotheses for the optimal sequence of pitch states. This algorithm yields up to two dominant pitch periods, adequate for segregating foreground and background streams. See [18] for details of multipitch tracking of reverberant mixtures.

With results of multipitch tracking, we extract pitch-based features for each u_{cm}. As discussed in [17], in order to achieve robust performance in reverberant signals, the feature set should be sensitive to both resolved and unresolved harmonics. In low-frequency channels, harmonics are resolved because a filter does not respond to more than one harmonic. When bandwidth increases in high-frequency channels, a filter responds to multiple harmonics and therefore harmonics become unresolved. Following [15] and [17], the first three features are derived from the inner hair cell response $h(c,t)$ to detect resolved harmonics. Given an estimated target pitch period τ_m at frame m, $A(c,m,\tau_m)$ is one direct measure of how the periodicity in u_{cm} is consistent with τ_m. This feature has been proven to be effective under both anechoic and reverberant conditions [17]. As an alternative way of measuring the harmonicity, the average instantaneous frequency $\bar{f}(c,m)$ is estimated from the zero-crossing rate of $A(c,m,\tau)$ with respective to τ. Thus, by multiplying $\bar{f}(c,m)$ with τ_m, we derive two supplementary features: the nearest integer [·] to the product and the distance | · | between the product and its nearest integer. The former indicates a harmonic number while the latter measures the amount of deviation from the harmonic number. The next three features are similarly extracted from the envelope of the hair cell output $h_E(c,t)$, which better reveals the amplitude modulation (AM) of unresolved harmonics. A bandpass filter with the passband from 50 to 550 Hz is utilized to extract AM. Hence, we obtain the six pitch-based features for u_{cm}

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

IV. Unit Labeling Using MLP

As a key stage of the segregation system, unit labeling distinguishes target-dominant T-F units from interference-dominant ones. An MLP classifier learns a mapping from the aforementioned feature set $\mathcal{O}_{c,m}(\tau_m)$ to a grouping cue, which encodes the posterior probability of a T-F unit being dominated by the source with τ_m. In this section, we first discuss the training of the MLP classifier. Later, labeling criteria are derived from the MLP output.

A. MLP Training

In the training stage, features are extracted using the ground truth of the target pitch, which is obtained by running Praat [4] on the premixed reverberant target signal followed by manual correction. The desired value of the grouping cue is defined to be 1 if u_{cm} is dominated by the source with given pitch τ_m and 0 otherwise. When the given pitch period arises from the target speech, the above definition is consistent with the IBM. On the contrary, the interference pitch corresponds to the complement IBM, with 0s and 1s swapped. In the following, we focus on the case when τ_m is the target pitch. The second case will be considered in Section IV-C.

The model consists of 128 MLPs, one trained for each channel. Training minimizes a cost-sensitive objective function (i.e., error function). Specifically, the objective function J_c is defined as [17]

$$J_c = \frac{\sum_m (d_c(m) - y_c(m))^2 \cdot E_c(m)}{\sum_m E_c(m)}. \quad (3)$$

Given desired output $d_c(m)$ and actual output $y_c(m)$, the objective function is expressed as a weighted mean squared error (MSE) with unit energy $E_c(m)$ as weights, which directly relates to the goal of maximizing SNR in the segregation problem [17]. In implementation, each MLP has the same network architecture with 6 input nodes, 20 hidden nodes, and 1 output node. The size of the hidden layer is decided by tenfold cross-validation. We use a hyperbolic tangent sigmoid function as the transfer function of the hidden and output layers. A generalized Levenberg–Marquardt backpropagation algorithm [11] is adapted to learn unknown network parameters in conjunction with J_c as the learning goal.

B. Single-Pitch Labeling

We apply the trained MLP to label u_{cm}. It should be noted that each frequency channel has a separately trained MLP, as mentioned in Section IV-A. Because the trained MLP directly estimates the posterior probability [24], we denote the MLP output by $P(D|\mathcal{O}_{c,m}(\tau_m))$, where D is the event of u_{cm} being dominated by the source with τ_m. Given that τ_m is the target pitch, a T-F unit u_{cm} is labeled as the target source if $P(D|\mathcal{O}_{c,m}(\tau_m))$ is greater than $P(\overline{D}|\mathcal{O}_{c,m}(\tau_m))$, where $\overline{D}$ is the complement of the event D. Because these two posterior probabilities sum to one, the above criterion is simplified to

$$P(D|\mathcal{O}_{c,m}(\tau_m)) > \frac{1}{2}. \quad (4)$$

Note that this criterion was derived in [17] with a different meaning of τ_m. The ground truth pitch of the target speech was previously used to concentrate on supervised learning without the influence of pitch estimation errors. In this paper, we avoid such prior knowledge and rely on estimated pitch. We instead assume the “ideal” assignment of detected pitch contours. Specifically, we assign a detected pitch contour to the target if it is closer to the ground truth target pitch, and to the interference otherwise.

However, inaccurate pitch estimation will lead to unit labeling errors, and hence degrade the segregation performance. Quantitative analysis will be given in Section VI-B. Section IV-C proposes a novel strategy to alleviate performance degradation.

C. Double-Pitch Labeling

When the interfering signal has periodic components or is another speech, the second pitch can be utilized to improve the labeling accuracy. In this situation, instead of evaluating how likely u_{cm} is dominated by the target speech, we ask whether the target or the interference more likely dominates u_{cm}. This criterion can be written as

$$P(D|\mathcal{O}_{c,m}(\tau_m)) > P(D|\mathcal{O}_{c,m}(\tau'_m)) \quad (5)$$

where τ_m and τ'_m are target and interference pitch periods, respectively. Such a comparison is made under the same conditions except for different pitch values, and as a result we expect it to be more accurate than a comparison with a fixed value as in (4) (see also [15]).

Due to the fundamental difference between speech and periodic nonspeech noise, we train two MLP models, denoted by M_1 and M_2, for speech and nonspeech sources, respectively. Each model learns a distinct mapping from the pitch-based features to the grouping cues for each type of harmonic sources. Two hypotheses are considered consequently:

H_1: Speech mixed with speech;
H_2: Speech mixed with nonspeech signal.

Since H_1 assumes that both target and interference signals are speech, the criterion in (5) is replaced by

$$P_{M_1}(D|\mathcal{O}_{c,m}(\tau_m)) > P_{M_1}(D|\mathcal{O}_{c,m}(\tau'_m)). \quad (6)$$

Similarly, the labeling criterion under H_2 is

$$P_{M_1}(D|\mathcal{O}_{c,m}(\tau_m)) > P_{M_2}(D|\mathcal{O}_{c,m}(\tau'_m)) \quad (7)$$

where P_{M_1} is the MLP output from M_1 and P_{M_2} from M_2.

A likelihood ratio test is then applied to select the correct hypothesis at the sentence level. Let L_1 be the unit label under H_1:

$$L_1(c,m) = \begin{cases} 1, & \text{if } P_{M_1}(D|\mathcal{O}_{c,m}(\tau_m)) > P_{M_1}(D|\mathcal{O}_{c,m}(\tau'_m)) \\ 0, & \text{else.} \end{cases} \quad (8)$$

and L_2 under H_2:

$$L_2(c,m) = \begin{cases} 1, & \text{if } P_{M_1}(D|\mathcal{O}_{c,m}(\tau_m)) > P_{M_2}(D|\mathcal{O}_{c,m}(\tau'_m)) \\ 0, & \text{else.} \end{cases} \quad (9)$$

Both are computed in two-pitched frames. We further define $\mathcal{F}$ to be the set of the overlapping interference-dominant units (0s) in L_1 and L_2

$$\mathcal{F} = \{(c,m) : L_1(c,m) = 0\} \cap \{(c,m) : L_2(c,m) = 0\}. \quad (10)$$

A likelihood ratio test thus can be conducted:

$$\frac{\prod_{(c,m)\in\mathcal{F}} P_{M_1}(D|\mathcal{O}_{c,m}(\tau'_m))}{\prod_{(c,m)\in\mathcal{F}} P_{M_2}(D|\mathcal{O}_{c,m}(\tau'_m))} \gtrless 1. \quad (11)$$

Essentially, the test compares the likelihoods of M_1 and M_2 and selects the model that better fits the interference data in $\mathcal{F}$.

The overall labeling strategy is illustrated in Fig. 1. After the detection of target and interference pitch contours $\{\tau_m\}$ and $\{\tau'_m\}$, we first label units on single-pitch frames based on the criterion in (4). For two-pitch frames, we branch the labeling process using the following two hypotheses: H_1 (speech plus speech, left branch), and H_2 (speech plus nonspeech, right branch). The speech model M_1 is applied to both of the sources under the first hypothesis and the labeling result is obtained according to (6). The second hypothesis calls for M_1 and M_2, respectively, and the result is obtained using (7). Given the two-talker mixture in this example, the likelihood ratio test in (11) chooses the labeling result from the left branch.

V. Segmentation and Grouping

In CASA systems, it is preferable to segment the auditory scene into contiguous T-F regions, each of which is deemed to mainly originate from a single source. A segment contains more global information of the underlying source that is missing from individual units, and thus is expected to provide more robust components for grouping and improve segregation performance [12], [35].

However, room reverberation makes it difficult to obtain accurate segments. Following [17], we perform segmentation on reverberant mixtures using two different methods in different frequency ranges. Due to the fact that the filter responses in low-frequency channels (center frequency below 800 Hz) are less corrupted by reverberation than those in high-frequency channels [30], cross-channel correlation [34] remains effective as the segmentation cue at low frequencies. On the other hand, signal onsets are relatively unaffected by reverberation because the direct sound arrives earlier than its echoes. Hence, the high-frequency regions are segmented using onset/offset cues [13]. The complete segmentation is formed by combining the segments generated from the above two methods.

Based on the unit labeling results, we then group each segment into the target stream if the total energy of the target-labeled T-F units is greater than that labeled as the non-target units. In the final step, each target segment is expanded by iteratively recruiting its neighboring units that are labeled as target and do not belong to any segments. The resulting binary mask is thus formed and the target speech can be resynthesized from this mask [34].

VI. Experimental Results

A. Data Preparation

To simulate room acoustics, we use the image model [1], [19] which produces the room impulse response (RIR) given the input of room dimensions, reflection coefficients, and the physical locations corresponding to sound sources and the microphone. The basic idea is to represent the RIR as an infinite number of image sources that are created by reflecting the actual sound source in six walls. Therefore, a source location together with a microphone location decide RIR in a fixed room. We first construct two acoustic rooms by specifying their room dimensions and reflection coefficients. Details are given in Table I. To simulate both convolutive and additive distortions in our mixture signals, we specify in each configuration a random location for the microphone and then choose two other locations for two sources (target and interference) randomly but control source-microphone distances to ensure that close-talking scenarios are avoided and signal-to-reverberant energy ratios are roughly constant in each simulated room. Note that, even in the same room, the RIRs from different source locations to the microphone may differ significantly [23], [27]. Consequently, a reverberant mixture is created by convolving each source with its corresponding RIR and mixing the two reverberant sources at 0-dB SNR. It is worth emphasizing here that the goal of our system is to segregate the reverberant target source from the mixture signal. A more detailed description of the simulation procedure can be found in [17].

The evaluation corpus is constructed by mixing ten randomly selected TIMIT utterances [10] with 15 different types of interference. In Table II, the interferences are classified into three categories: 1) those with no pitch; 2) those with some pitch qualities; and 3) other speech utterances, such that the segregation performance can be evaluated differently in each category. The interfering signals are from NOISEX-2 [32], Cooke's corpus [8], and TIMIT. To evaluate different reverberant conditions, we choose within each room three random configurations and use each of these configurations to construct a reverberant mixture. Consequently, we have a total of 1050 mixtures, with the

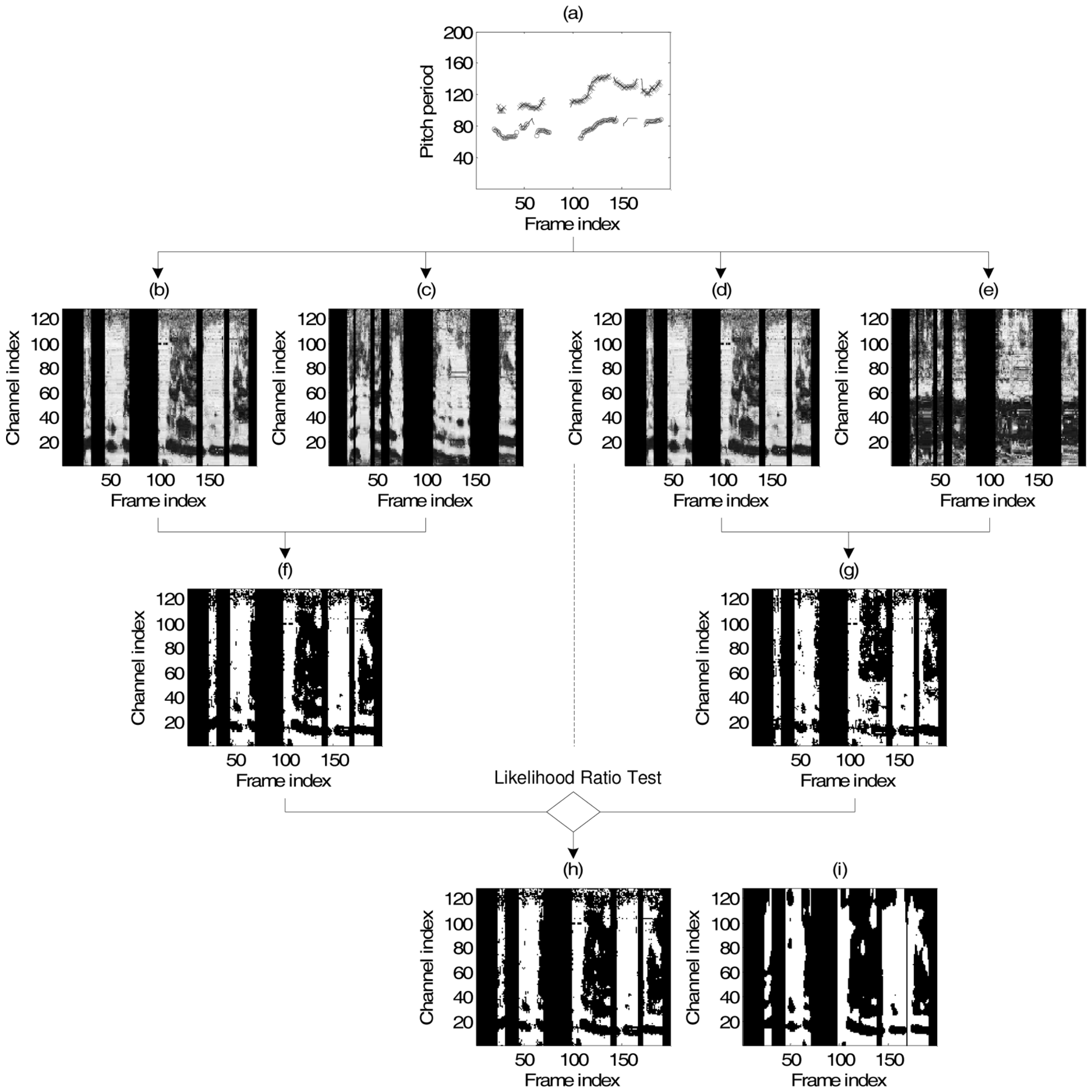

Fig. 1. Illustration of the unit labeling procedure. (a) Pitch tracking results for a mixture of one male (target) and one female (interference) utterance. The solid lines indicate the ground truth pitch contours. The crosses represent the estimated male pitch contours whereas the circles the female pitch contours. (b) The M_1 output corresponding to the target pitch contours. Brighter color indicates higher posterior probability. (c) The M_1 output corresponding to the interference pitch contour. (d) Same as (b). (e) The M_2 output corresponding to the interference pitch contours. (f) The binary mask under the H_1 hypothesis using (4) on single-pitch frames and (6) on two-pitch frames. The target-dominant units are labeled as white and the interference-dominant ones black. (g) The binary mask for the H_2 hypothesis. (h) The binary mask chosen by a likelihood ratio test. (i) The ideal binary mask for the purpose of comparison.

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

ROOM NO.	L × W × H (m)	Reflection Coeff.	T_{60} (s)
1	6 × 4 × 3	0.73	0.3
2	9 × 5 × 3	0.87	0.6

TABLE II
CATEGORY OF INTERFERING SIGNALS

Category 1	White noise, pink noise, car noise, F16 cockpit noise, speech shape noise
Category 2	1 kHz tone, "cocktail party" noise, rock music, siren, trill telephone
Category 3	3 female utterances, 2 male utterances

original 150 mixtures in anechoic and $2 \times 3 \times 150$ mixtures in reverberant conditions.

B. SNR Evaluation

Given the computational goal of estimating the IBM, we use the SNR measure in [12] to assess the segregation performance using the resynthesized speech from the IBM as a ground truth

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_t s_I^2(t)}{\sum_t \left(s_I(t) - s_E(t)\right)^2}. \tag{12}$$

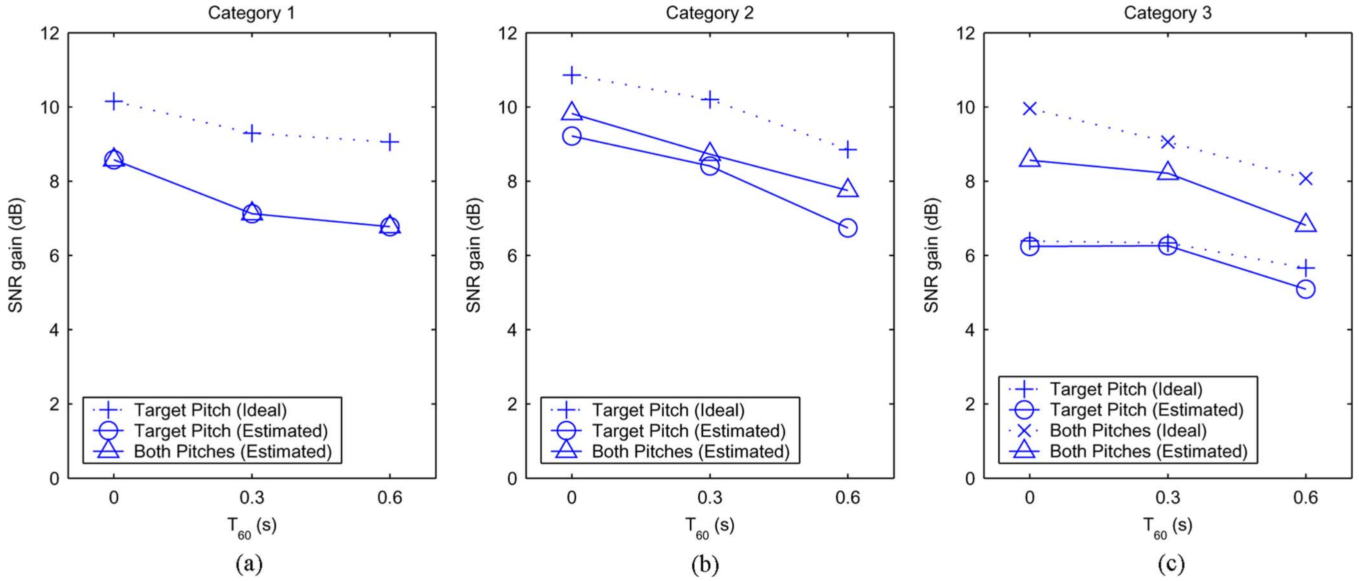

Fig. 2. SNR gains with different labeling strategies. The three panels (from left to right) indicate three different categories of interference.

where $s_I(t)$ and $s_E(t)$ are signals resynthesized from the IBM and an estimated mask, respectively. In the following, we evaluate the segregation performance in terms of the SNR gain, which is the improvement over the SNR before segregation. The latter is calculated by passing the mixture to the all-one mask as $s_E(t)$ in (12).

When the acoustic environment changes, the pitch-based features are likely to vary accordingly and thus generalization could become an issue. Different training strategies were compared in [17], results indicate that generalization to less reverberant conditions is better than the other way around. Therefore, we trained the speech model M_1 under the more reverberant condition (Room 2, with reverberation time $T_{60} = 0.6$ s) using 150 mixtures from one of the three configurations. Because nonspeech periodic signals only exist in Category 2, the nonspeech model M_2 is trained using 50 mixtures from that category under the same reverberant condition.

We first assess the performance degradation brought about by pitch estimation errors. The dotted lines in Fig. 2 show the segregation performance in different types of interference using the ground-truth target pitch, which is the system described in [17]. Note that this performance is not obtainable due to the use of prior pitch knowledge. The SNR performance drops with increasing T_{60}, reflecting the nature of the rising difficulty of segregation due to room reverberation. The circle lines show the SNR performance with the estimated pitch. Different amounts of degradation are observed in each category of interference. The first two categories experience an average of 2.1-dB SNR loss from the anechoic to the most reverberant conditions. In the third category, the discrepancy is less than 0.4 dB on average. These results are consistent with the pitch estimation errors reported in [18] and indicate that the MLP model is able to generalize when the input pitch is not accurate.

The triangle lines in Fig. 2 demonstrate the advantage of the proposed double-pitch labeling strategy. In Category 1, there is no SNR improvement due to the lack of the second pitch from interfering signals. In the second and the third categories, the interference pitch is often available and enables the labeling method in Section IV-C to be utilized. As can be seen, significant SNR improvement is realized. In Category 2, the SNR gain improves about 1 dB in all reverberant conditions, bringing the performance curve closer to the performance with the ground truth pitch. We note that the nonspeech model is trained on a closed set of interferences and generalization to new noises remains to be studied. We also note that, when using a single model for both speech and nonspeech interference, the double-pitch labeling strategy performs even slightly worse than single-pitch labeling (not shown in the figure) due to the model's inability to capture both speech and nonspeech characteristics. In Category 3, the double-pitch labeling strategy shows a considerable advantage over single-pitch labeling. The two-pitch labeling strategy using estimated pitches even outperforms the ground-truth single-pitch performance by about 2 dB in the anechoic condition and 1 dB when T_{60} is 0.6 s. With ideal pitches, the performance is further improved by more than 1 dB as shown in the figure. There are three reasons contributing to this performance. First, we can detect accurate pitch contours when the target and the interference are both speech signals. Second, the voiced parts of the two competing voices often overlap significantly, resulting a high percentage of two-pitch frames conducive to high SNR improvement. Third, the likelihood ratio test works perfectly in the two-talker category. To illustrate the improved performance, Figs. 3 and 4 show the estimated IBMs from the single- and double-pitch labeling strategies in Category 2 and 3 interferences, respectively. The IBMs are also displayed for the purpose of comparison.

We have also evaluated the proposed system using the conventional SNR measure. In all three categories of interference and different reverberant conditions, the conventional SNR gains are about 1.5 dB lower but show similar trends to those in Fig. 2. This is consistent with earlier comparisons [12]. Since it directly relates to our computational objective, we use the IBM-based SNR measure in later experiments.

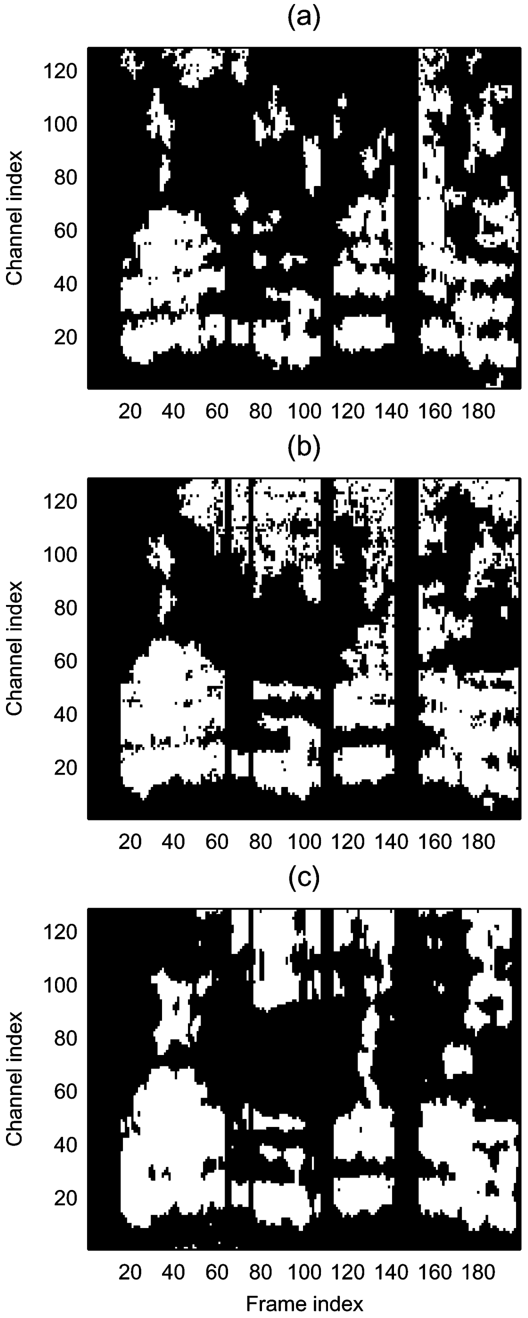

Fig. 3. IBM estimation for the reverberant mixture of one female utterance and rock music with $T_{60} = 0.6$ s. (a) Labeling by the target pitch. (b) Labeling by both the target and the interference pitches. (c) Ideal binary mask.

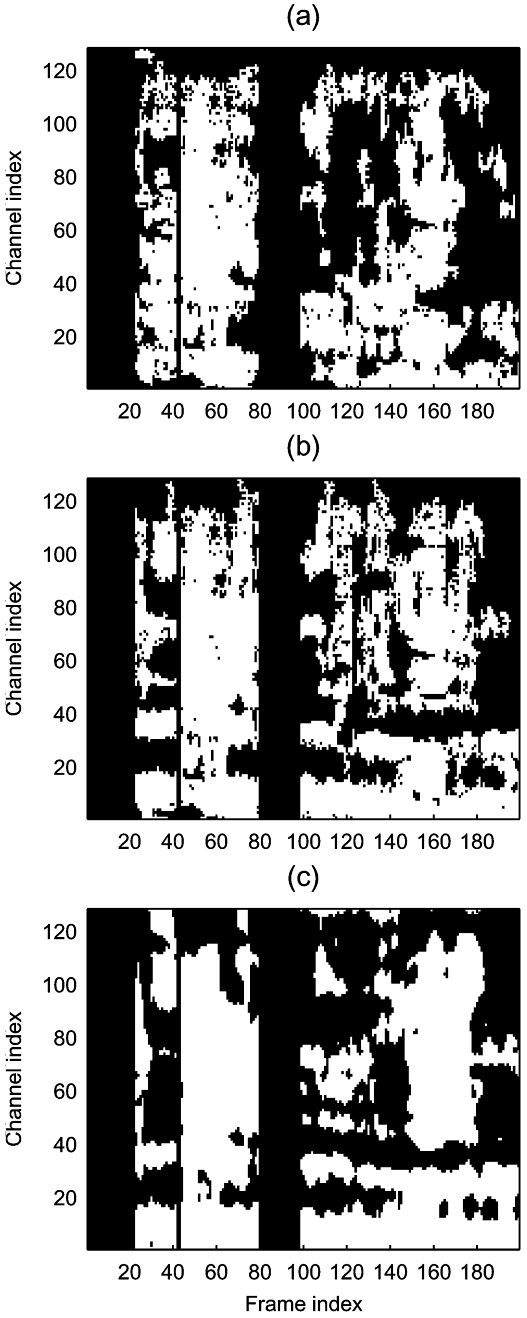

Fig. 4. IBM estimation for the reverberant mixture of one male (target) and one female (interference) utterance with $T_{60} = 0.6$ s. (a) Labeling by the target pitch. (b) Labeling by both the target and the interference pitches. (c) Ideal binary mask.

C. Comparison

To put the performance of our approach in perspective, we compare with the tandem algorithm [15] and spectral subtraction [5]. The tandem algorithm is a recent CASA approach which performs pitch estimation and speech segregation jointly and iteratively and reports very good performance. Specifically, it starts with an initial estimate of pitch and uses this estimate to segregate target speech based on harmonicity and temporal continuity. The segregated speech is then used to re-estimate pitch and the improved pitch estimate leads to better segregation, and so on. In the segregation stage, a classification based approach is used for pitch-based grouping. There are two major differences between the tandem algorithm and our approach: 1) the classifier in the tandem algorithm is trained using a conventional MSE objective function, which is suboptimal for the goal of maximizing SNR [17]; and 2) the tandem algorithm uses the same training to deal with both speech and nonspeech harmonic sources. Like in our system, ideal sequential grouping is used to obtain the segregated target. Since the tandem model is developed and trained for anechoic signals, the performance is expected to degrade in reverberant environments. Fig. 5 shows in circle lines the SNR performance of the tandem algorithm. On average, it performs closely to the proposed system in the anechoic condition; its slight advantage reflects the fact that it is trained in the anechoic condition unlike our model which is

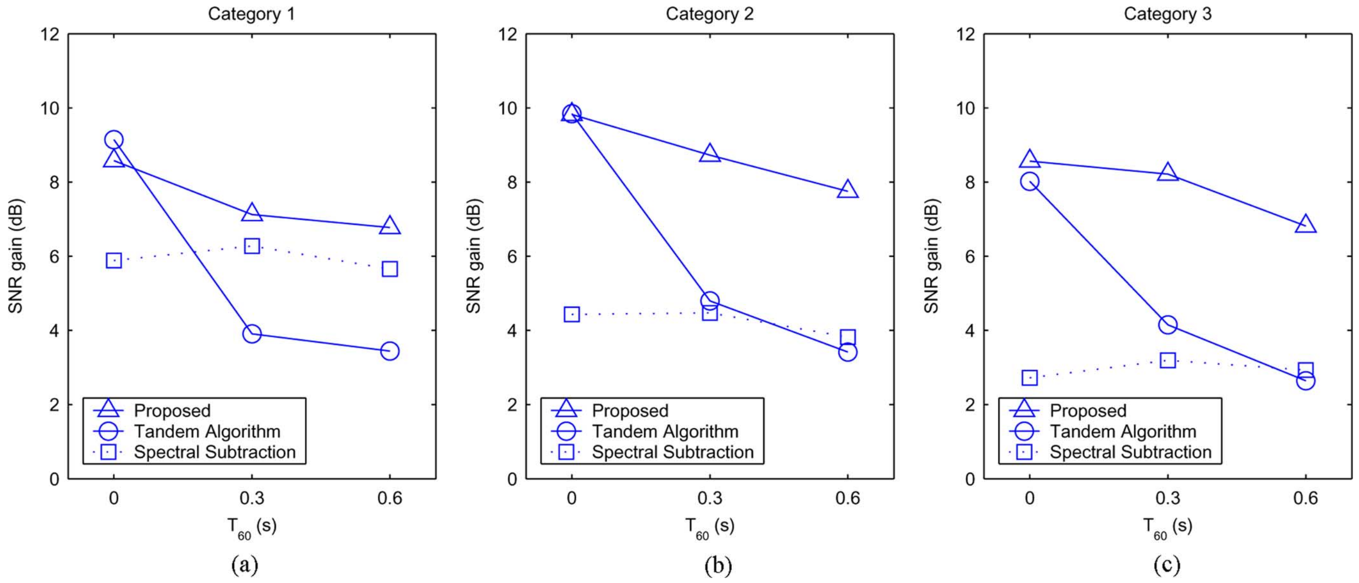

Fig. 5. Comparison of SNR gain among the proposed method, the tandem algorithm and spectral subtraction. The three panels (from left to right) indicate three different categories of interferences.

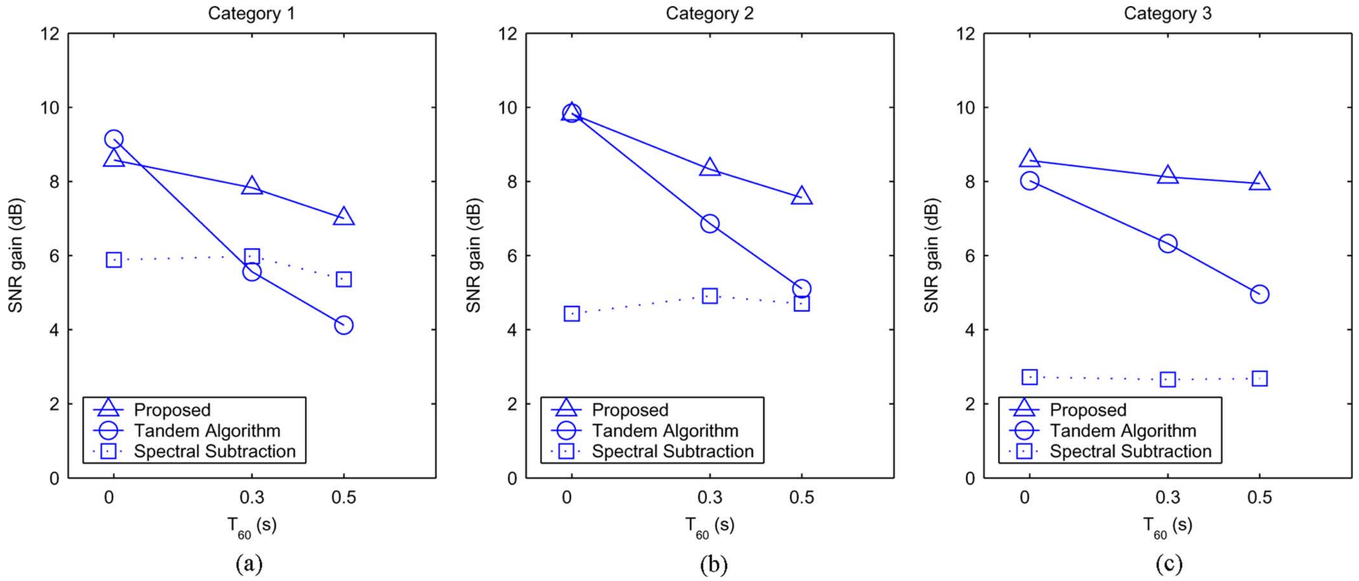

Fig. 6. Comparison of SNR gain among the proposed method, the tandem algorithm and spectral subtraction in real rooms. The three panels (from left to right) indicate three different categories of interferences.

trained at $T_{60} = 0.6$ s. As expected, the tandem model loses accuracy in both pitch estimation and pitch-based grouping when room reverberation occurs. The performance gap, compared to the proposed system, is about 3 dB in broadband noise and more than 4 dB in the next two categories of interference. Apart from the mismatched training, the pitch estimator in the tandem algorithm also has problems dealing with reverberant signals.

To make spectral subtraction perform on all types of interference, we provide the prior knowledge of the silent frames in the target speech for the required noise spectrum estimation. In each time frame, interference is attenuated by subtracting the most recently updated noise estimate from the spectrum of the mixture. Fig. 5 shows the performance of the spectral subtraction method in square lines. As can be seen, it yields reasonable performance in Category 1 because broadband noise largely conforms to the assumption of stationary noise. In Category 2, our system outperforms spectral subtraction by 4 dB. In the two-talker category, our system performs about 6 dB better in the anechoic condition and 4.5 dB better in reverberant conditions. We should point out that the performance of spectral subtraction is not sensitive to room reverberation because the quality of noise estimation described above does not change much with respect to the level of reverberation.

Finally, we construct another corpus using RIRs recorded in real rooms (see the Acknowledgement). We choose two acoustic rooms with $T_{60} = 0.3$ and 0.5 s, respectively. In each room, two RIRs corresponding to the first two omnidirectional microphones are selected for generating reverberant mixtures. The same target and interference signals as described in Section VI-A are used for generating the new corpus. We use the same speech and nonspeech models (M_1 and M_2) trained on the simulated reverberant conditions; in other words, no retraining is conducted. Fig. 6 presents and compares the SNR gains of the proposed system, the tandem algorithm and the spectral subtraction method in real rooms. These results are broadly consistent with those in Fig. 5. Similar to simulated RIRs, the proposed algorithm performs significantly better than the other two algorithms in all reverberant conditions. These results demonstrate that our trained classifiers generalize well from simulated reverberant conditions to real environments.

VII. Concluding Remarks

The paper proposes a CASA system for segregating reverberant speech by incorporating multipitch tracking and supervised classification to deal with corrupted harmonic features due to room reverberation. In [17], the trained classifier labels T-F units reliably and generalizes well to unseen reverberant conditions, speakers, and utterances. With the same learning scheme, our system exhibits similarly good generalization performance. Furthermore, the proposed model yields good performance when the input pitch is not accurate, which shows another aspect of generalizability. Our approach trains a separate model for periodic nonspeech signals that may occur in the background and use a likelihood ratio test for model selection. Evaluation and comparison show that our approach produces substantial SNR gains across different levels of room reverberation, and performs significantly better than related segregation methods.

Our study avoids the problem of sequential grouping by using "ideal" assignment of pitch contours. In monaural conditions, human listeners utilize multiple grouping principles to perform sequential organization [6]. Previous work has attempted to model source characteristics for sequential grouping (e.g., [9] and [31]). However, such effort has not addressed the issue of reverberation. How to perform sequential organization in reverberant conditions is an important topic for future research.

The problem of unvoiced speech segregation is not dealt with in this work. Due to its lack of harmonicity, unvoiced speech is intrinsically different from its voiced counterpart and cannot be captured by pitch related features in (2). Under anechoic conditions, Hu and Wang [14] developed an unvoiced speech segregation algorithm by analyzing signal onsets/offsets and classifying acoustic-phonetic features. A subsequent method [16] applies spectral subtraction to enhance unvoiced speech with noise estimated from pitched intervals. To our knowledge, however, no study has attempted to tackle the problem of unvoiced speech segregation under reverberant conditions and further research is required here.

Acknowledgment

The authors would like to thank the anonymous reviewers for their useful suggestions/criticisms. The authors would also like to thank J. West and G. Elko of Bell Labs, and C. Avendano of UC Davis for recording the real room impulse responses used in this study.

References

[1] 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.

[2] 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.

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

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

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

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

[7] 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.

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

[9] D. Ellis, "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley/IEEE Press, 2006, pp. 115–146.

[10] 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

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

[12] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[13] 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.

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

[15] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.

[16] K. Hu and D. L. Wang, "Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction," Dept. Comp. Sci. and Eng., The Ohio State Univ., Columbus, 2009, Tech. Rep. 51.

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

[18] Z. Jin and D. L. Wang, "A multipitch tracking algorithm for noisy and reverberant speech," in *Proc. IEEE ICASSP*, 2010, pp. 4218–4221.

[19] E. Lehmann, "Image-source method: Matlab code implementation," 2008 [Online]. Available: http://www.eric-lehmann.com/ism_code.html

[20] 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.

[21] P. C. Loizou, *Speech Enhancement: Theory and Practice*. New York: Taylor & Francis, 2007.

[22] R. Meddis, "Simulation of auditory-neural transduction: Further studies," *J. Acoust. Soc. Amer.*, vol. 83, pp. 1056–1063, 1988.

[23] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *J. Sound Vibr.*, vol. 102, pp. 217–228, 1985.

[24] 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.

[25] 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., 1988, APU Rep. 2341.

[26] 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.

[27] 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, pp. 311–319, Nov. 2000.

[28] N. Roman and D. L. Wang, "Pitch-based monaural segregation of reverberant speech," *J. Acoust. Soc. Amer.*, vol. 120, pp. 458–469, 2006.

[29] S. T. Roweis, "One microphone source separation," in *Proc. NIPS*, 2000, pp. 793–799.

[30] M. Sayles and I. M. Winter, "Reverberation challenges the temporal representation of the pitch of complex sounds," *Neuron*, vol. 58, pp. 789–801, 2008.

[31] Y. Shao and D. L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 289–298, 2006.

[32] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," in *Speech Commun.*, 1993, vol. 12, pp. 247–251.

[33] 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.

[34] 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.

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

[36] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time–frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, pp. 2336–2347, 2009.

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 from Peking (Beijing) University, Beijing, China, in 1983 and 1986, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, CA, in 1991, all in computer science.

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, 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, 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.