Pitch-based monaural segregation of reverberant speech

Nicoleta Roman[a)]
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210

DeLiang Wang[b)]
Department of Computer Science and Engineering & Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

(Received 13 April 2005; revised 20 January 2006; accepted 23 March 2006)

In everyday listening, both background noise and reverberation degrade the speech signal. Psychoacoustic evidence suggests that human speech perception under reverberant conditions relies mostly on monaural processing. While speech segregation based on periodicity has achieved considerable progress in handling additive noise, little research in monaural segregation has been devoted to reverberant scenarios. Reverberation smears the harmonic structure of speech signals, and our evaluations using a pitch-based segregation algorithm show that an increase in the room reverberation time causes degraded performance due to weakened periodicity in the target signal. We propose a two-stage monaural separation system that combines the inverse filtering of the room impulse response corresponding to target location and a pitch-based speech segregation method. As a result of the first stage, the harmonicity of a signal arriving from target direction is partially restored while signals arriving from other directions are further smeared, and this leads to improved segregation. A systematic evaluation of the system shows that the proposed system results in considerable signal-to-noise ratio gains across different conditions. Potential applications of this system include robust automatic speech recognition and hearing aid design.
© *2006 Acoustical Society of America.* [DOI: 10.1121/1.2204590]

PACS number(s): 43.72.Dv [DOS] Pages: 458–469

I. INTRODUCTION

In a natural environment, a desired speech signal often occurs simultaneously with other interfering sounds such as echoes and background noise. While the human auditory system excels at speech segregation from such complex mixtures, simulating this perceptual ability computationally remains a great challenge. In this paper, we study the monaural separation of reverberant speech. Our monaural study is motivated by the following two considerations. First, an effective one-microphone solution to sound separation is highly desirable in many applications including automatic speech recognition and speaker recognition in real environments, audio information retrieval, and hearing prosthesis. Second, although binaural listening improves the intelligibility of target speech under anechoic conditions (Bronkhorst, 2000), this binaural advantage is largely diminished by reverberation (Plomp, 1976; Culling *et al.*, 2003); this underscores the dominant role of monaural hearing in realistic conditions.

Various techniques have been proposed for monaural speech enhancement including spectral subtraction (e.g., Martin, 2001), Kalman filtering (e.g., Ma *et al.*, 2004), subspace analysis (e.g., Ephraim and Trees, 1995), and autoregressive modeling (e.g., Balan *et al.*, 1999). However, these methods make strong assumptions about the interference and thus have difficulty in dealing with a general acoustic background. Another line of research is the blind separation of signals using independent component analysis (ICA). While standard ICA techniques perform well when the number of microphones is greater than or equal to the number of observed signals such techniques do not function in monaural conditions. Some recent sparse representations attempt to relax this assumption (e.g., Zibulevsky *et al.*, 2001). For example, by exploiting *a priori* sets of time-domain basis functions learned using ICA, Jang *et al.* (2003) attempted to separate two source signals from a single channel but the performance is limited.

Inspired by the human listening ability, research has been devoted to build speech separation systems that incorporate known principles of auditory perception. According to Bregman (1990), the auditory system performs sound separation by employing various cues including pitch, onset time, spectral continuity, and location in a process known as auditory scene analysis (ASA). This ASA account has inspired a series of computational ASA (CASA) systems that have significantly advanced the state-of-the-art performance in monaural separation (e.g., Weintraub, 1985; Cooke, 1993; Brown and Cooke, 1994; Wang and Brown, 1999; Hu and Wang, 2004) as well as in binaural separation (e.g., Roman *et al.*, 2003; Palomaki *et al.*, 2004). Generally, CASA systems follow two stages: segmentation (analysis) and grouping (synthesis). In segmentation, the acoustic input is decomposed into sensory segments, each of which originates from a single source. In grouping, the segments that likely come from the same source are put together. A recent overview of both monaural and binaural CASA approaches can be found in Brown and Wang (2005). Compared with speech enhance-

[a)]Electronic mail: roman.45@osu.edu
[b)]Electronic mail: dwang@cse.ohio-state.edu

 0001-4966/2006/120(1)/458/12/$22.50 © 2006 Acoustical Society of America

ment techniques described above, CASA systems make few assumptions about the acoustic properties of the interference and the environment.

CASA research, however, has been largely limited to anechoic conditions, and few systems have been designed to operate on reverberant input. A notable exception is the binaural system proposed by Palomaki *et al.* (2004) which includes an inhibition mechanism that emphasizes the onset portions of the signal and groups them according to common location. Evaluations in reverberant conditions have also been reported for a series of two-microphone algorithms that combine pitch information with binaural cues or signal-processing techniques (Luo and Denbigh, 1994; Shamsoddini and Denbigh, 2001; Barros *et al.*, 2002).

At the core of many CASA systems is a time-frequency (T-F) mask. Specifically, the T-F units in the acoustic mixture are selectively weighted in order to enhance the desired signal. The weights can be binary or real (Srinivasan *et al.*, 2004). The binary T-F masks are motivated by the masking phenomenon in human audition, in which a weaker signal is masked by a stronger one in the same critical band (Moore, 2003). Additionally, from the speech segregation perspective, the notion of an *ideal binary mask* has been proposed as the computational goal of CASA (Wang, 2005). Such a mask can be constructed from *a priori* knowledge about target and interference; specifically a value of 1 in the mask indicates that the target is stronger than the interference and 0 indicates otherwise. Speech reconstructed from the ideal binary mask has been shown to be highly intelligible even when extracted from multisource mixtures and also to produce large improvements in robust speech recognition and human speech intelligibility (Cooke *et al.*, 2001; Roman *et al.*, 2003; Brungart *et al.*, 2006).

Perceptually, one of the most effective cues for speech segregation is the fundamental frequency (F0) (Darwin and Carlyon, 1995). Accordingly, much work has been devoted to build computational systems that exploit the F0 of a desired source to segregate its harmonics from the interference (for a review see Brown and Wang, 2005). In particular, the system proposed by Hu and Wang (2004) employs differential strategies to segregate resolved and unresolved harmonics. More specifically, periodicities detected in the response of a cochlear filterbank are used at low frequencies to segregate resolved harmonics. In the high-frequency range, however, the cochlear filters have wider bandwidths and a number of harmonics interact within the same filter, causing amplitude modulation (AM). In this case, their system exploits periodicities in the response envelope to group unresolved harmonics. In this paper, we propose a pitch-based speech segregation method that follows the same principles while simplifying the calculations required for extracting periodicities. The method shows good performance when tested with a variety of noise intrusions under anechoic conditions. However, when F0 varies with time in a reverberant environment, reflected waves with different F0s arrive simultaneously with the direct sound. This multipath situation causes smearing of harmonic structure (Darwin and Hukin, 2000). Due to weakened harmonicity, the performance of pitch-based segregation degrades in reverberant conditions.

One method for removing the reverberation effect is to pass the reverberant signal through a filter that inverts the reverberation process and hence reconstructs the original signal. However, because a typical room impulse response is not minimum phase, perfect one-microphone reconstruction requires a noncausal infinite impulse response filter with a large delay (Neely and Allen, 1979). In addition, one needs to have *a priori* knowledge of the room impulse response, which is often impractical. Several methods have been proposed to estimate the inverse filter in unknown acoustical conditions (Furuya and Kaneda, 1997; Gillespie *et al.*, 2001; Nakatani and Miyoshi, 2003). In particular, the system developed by Gillespie *et al.* (2001) estimates the inverse filter from an array of microphones using an adaptive gradient-descent algorithm that maximizes the kurtosis of linear prediction (LP) residuals. The inverse filter results in reduction of perceived reverberation as well as enhanced harmonicity. In this paper, we employ a one-microphone adaptation of this method proposed by Wu (2003; Wu and Wang, 2006).

The dereverberation algorithms described above are designed to enhance a single reverberant source. Here, we investigate the effect of inverse filtering as preprocessing for a pitch-based speech segregation system in order to improve its robustness in reverberant environments. The key idea is to estimate a filter that inverts the room impulse response corresponding to the target source. The effect of applying this inverse filter on the reverberant mixture is twofold: It improves the harmonic structure of the target signal while smearing those signals originating at other locations. Using a signal-to-noise ratio (SNR) evaluation, we show that the inverse filtering stage improves the separation performance of our pitch-based system. To our knowledge, this is the first study that addresses monaural speech segregation with room reverberation.

The rest of the paper is organized as follows. The next section defines the problem domain and presents a model overview. Section III gives a detailed description of the dereverberation stage. Section IV gives a detailed description of the pitch-based segregation stage. Section V presents systematic results on pitch-based segregation both in reverberant and inverse-filtered conditions. We also make a comparison with the spectral subtraction method. Section VI concludes the paper.

II. MODEL OVERVIEW

The speech received at one ear in a reverberant enclosure undergoes both convolutive and additive distortions:

$$y(t) = h(t) * s(t) + n(t), \tag{1}$$

where "$*$" indicates convolution. $s(t)$ is the clean (anechoic) target speech signal to be recovered, $h(t)$ models the acoustic transfer function from target speaker location to the ear, and $n(t)$ is the reverberant background noise which usually contains interfering sources at other locations. As explained in the Introduction, the problem of monaural speech segregation has been studied extensively in the additive condition by employing the periodicity of target speech. However, room reverberation poses an additional challenge by smearing the

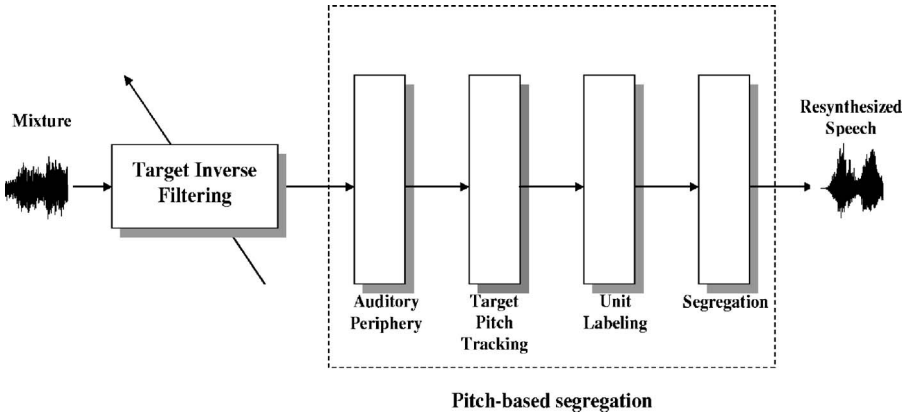

FIG. 1. Schematic diagram of the proposed two-stage model.

spectrum and weakening the harmonic structure. Consequently, we propose a two-stage speech segregation model: (1) inverse filtering with respect to the target location in order to enhance the periodicity of the target signal; (2) pitch-based speech segregation. Figure 1 illustrates the architecture of the proposed model.

The input to our model is a monaural mixture of two or more sound sources in a small reverberant room (6 m $\times$4 m$\times$3 m). The receiver—the left ear of a Knowles Electronic Manikin for Auditory Research (KEMAR) (Burkhard and Sachs, 1975)—is fixed at (2.5 m, 2.5 m, and 2 m) while the acoustic sources are located at a distance of 1.5 m from the receiver. The impulse response corresponding to the acoustic transfer function from a source to the receiver is simulated using a room acoustic model. Specifically, the simulated reflections from the walls are given by the image reverberation model (Allen and Berkley, 1979) and are convolved with the measured head related impulse responses (HRIR) of the KEMAR (Gardner and Martin, 1994). This represents a realistic input signal at the ear. Specific room reverberation times are obtained by varying the absorption characteristics of room boundaries (Palomaki *et al.*, 2004). Note that two different positions in the room produce impulse responses that differ greatly. The original clean signals are upsampled at the HRIR sampling frequency of 44 kHz and then convolved with the corresponding room impulse responses. Finally, the resulting reverberant signals are added together and resampled at 16 kHz.

In the first stage, a finite impulse response filter is estimated that inverts the target room impulse response. Adaptive filtering strategies for estimating this filter are sensitive to background noise (Haykin, 2002). For simplicity, we perform this estimation during an initial training stage using reverberant speech from the target location in the absence of background noise. We employ the inverse-filtering method by Gillespie *et al.* (2001), which uses a relatively small amount of training data. During testing, the inverse filter is applied to a mixture signal consisting of a reverberant target signal and interfering signals. The result is then fed to the next stage. We emphasize that this initial training is not utterance dependent; that is, the utterances used in training and testing can be totally different.

In the second stage, we employ a pitch-based segregation system to separate the inverse-filtered target signal. The signal is analyzed using a gammatone filterbank (Patterson *et al.*, 1988) in consecutive time frames to produce a T-F decomposition, where a basic T-F unit refers to the response of a particular filter channel in a particular time frame. Our system computes a correlogram which is a standard technique for periodicity extraction (Licklider, 1951; Slaney and Lyon, 1993). Specifically, autocorrelation is computed at the output of a particular channel and the set of the autocorrelations for all channels forms the correlogram. In the high-frequency range, we use response envelopes and extract AM rates. The system then groups those T-F units where the underlying target is stronger than the combined interference by comparing the extracted periodicities with an estimated target pitch. Labeling at the T-F unit level is a local decision and therefore prone to noise. Following Bregman's conceptual model, previous CASA systems employ an initial segmentation stage followed by a grouping stage in which segments likely to originate from the same source are grouped together (see, e.g., Wang and Brown, 1999). To enhance the robustness, we further perform segmentation. The result of this process is a binary T-F mask corresponding to the target stream.

Finally, a speech wave form is resynthesized from the resulting binary mask using a method described by Weintraub (1985; see also Brown and Cooke, 1994). The signal is reconstructed from the output of the gammatone filterbank. To remove across-channel differences, the output of the filter is time reversed, passed through the gammatone filter, and reversed again. The mask is employed to retain the acoustic energy from the mixture that corresponds to one's in the mask and nullifies the others.

III. TARGET INVERSE FILTERING

As described in the Introduction, inverse filtering is a standard strategy used for deriving the anechoic signal. We employ the method proposed by Gillespie *et al.* (2001) which attempts to blindly estimate the inverse filter from single-source reverberant speech. Their method was originally proposed for multi-microphone situations and has subsequently been extended to monaural recordings by Wu and Wang (2006). Based on the observation that peaks in the LP residual of speech are weakened by reverberation, an adaptive algorithm estimates the inverse filter by maximizing the kurtosis of the inverse-filtered LP residual of reverberant speech $\tilde{z}(t)$

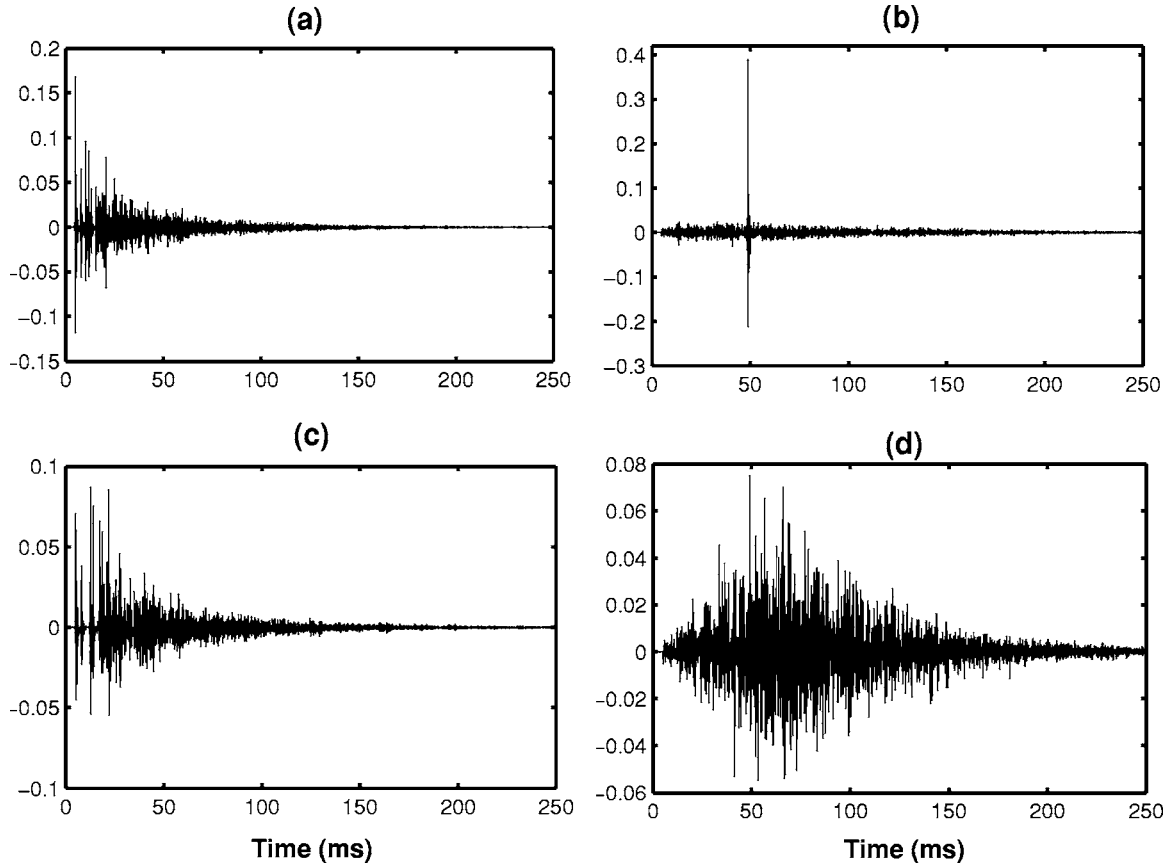

FIG. 2. Effects of inverse filtering on room impulse responses. (a) A room impulse response for a target source presented in the median plane. (b) The effect of convolving the impulse response in (a) with an estimated inverse filter. (c) A room impulse response for one interfering source at 45° azimuth. (d) The effect of convolving the impulse response in (c) with the estimated inverse filter.

$$\tilde{z}(t) = \mathbf{q}\mathbf{y}_r^T(t), \tag{2}$$

where $\mathbf{y}_r(t) = [y_r(t-L+1), \ldots, y_r(t-1), y_r(t)]$ and $y_r(t)$ is the LP residual of the reverberant speech from the target source, and $\mathbf{q}$ is an inverse filter of length L. The inverse filter is derived by maximizing the kurtosis of $\tilde{z}(t)$, which is defined as

$$J = \frac{E[\tilde{z}^4(t)]}{E^2[\tilde{z}^2(t)]} - 3. \tag{3}$$

The gradient of the kurtosis with respect to the inverse filter $\mathbf{q}$ can be approximated as follows (Gillespie *et al.*, 2001):

$$\frac{\partial J}{\partial \mathbf{q}} \approx \left\{ \frac{4(E[\tilde{z}^2(t)]\tilde{z}^3(t) - E[\tilde{z}^4(t)]\tilde{z}(t))}{E^3[\tilde{z}^2(t)]} \right\} \mathbf{y}_r(t). \tag{4}$$

Consequently, the optimization process in the time domain is given by the following update equation:

$$\hat{\mathbf{q}}(t+1) = \hat{\mathbf{q}}(t) + \mu f(t) \hat{\mathbf{y}}_r(t), \tag{5}$$

where $\hat{\mathbf{q}}(t)$ is the estimate of the inverse filter at time t, μ denotes the update rate, and $f(t)$ denotes the term inside the braces of Eq. (4).

However, a direct time-domain implementation of the above update equation is not desirable since it results in very slow convergence or no convergence at all under noisy conditions (Haykin, 2002). In this paper, we use the fast-block LMS (least mean square) implementation for one microphone signals described by Wu and Wang (2006). This method shows good convergence when applied to one-microphone reverberant signals for a range of reverberation times. The signal is processed block by block using a size L for both filter length and block length with the following update equations:

$$\mathbf{Q}'(n+1) = \mathbf{Q}(n) + \frac{\mu}{M} \sum_{m=1}^{M} \mathbf{F}(m) \mathbf{Y}_r^*(m), \tag{6}$$

$$\mathbf{Q}(n+1) = \frac{\mathbf{Q}'(n+1)}{|\mathbf{Q}'(n+1)|}, \tag{7}$$

where $\mathbf{F}(m)$ and $\mathbf{Y}_r(m)$ represent the fast Fourier transform (FFT) of $f(t)$ and $\mathbf{y}_r(t)$ for the m*th* block, and $\mathbf{Q}(n)$ represents the estimate for the FFT of inverse filter $\mathbf{q}$ at iteration n. M represents the number of blocks and the superscript * indicates the complex conjugation. Equation (7) ensures that the estimate of the inverse filter is normalized.

The system is trained on reverberant speech from the target source sampled at 16 kHz and presented alone. We employ a training corpus consisting of ten speech signals from the TIMIT database: five female utterances and five male utterances. An inverse filter of length $L = 1024$ is adapted for 500 iterations on the training data.

Figure 2 shows the outcome of convolving an estimated inverse filter with both the target impulse response as well as

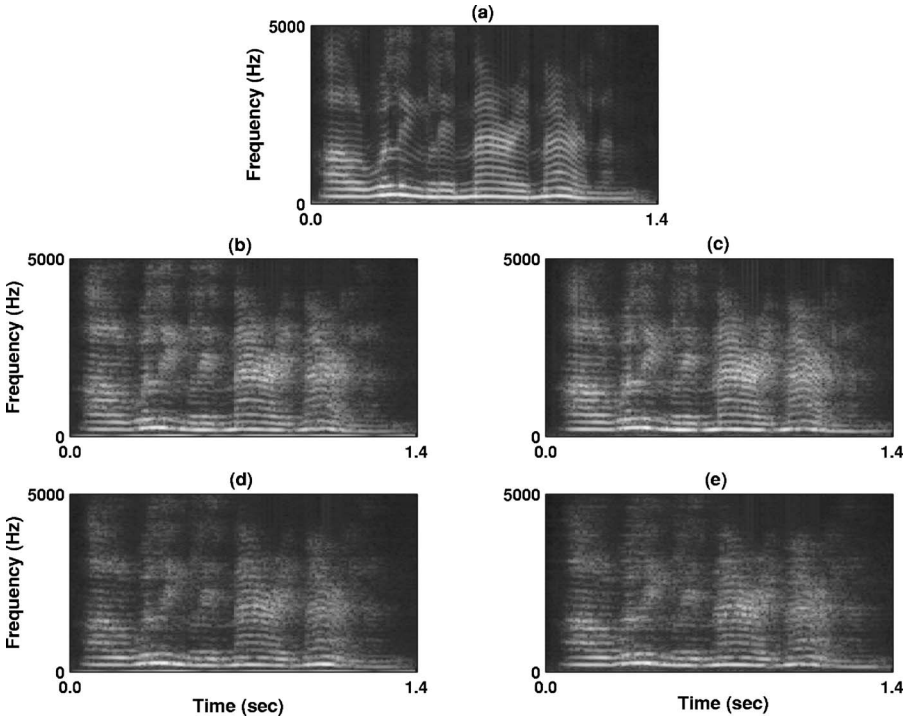

FIG. 3. Effects of reverberation and target inverse filtering on the harmonic structure of a voiced utterance. (a) Spectrogram of the anechoic signal. (b) Spectrogram of the reverberant signal corresponding to the impulse response in Fig. 2(a). (c) Spectrogram of the inverse-filtered signal corresponding to the equalized impulse response in Fig. 2(b). (d) Spectrogram of the reverberant signal corresponding to the room impulse response in Fig. 2(c). (e) Spectrogram of the inverse filtered signal corresponding to the impulse response in Fig. 2(d).

the impulse response at a different source location. The room reverberation time T_{60} is 0.35 s (T_{60} is the time required for the sound level to drop by 60 dB following the sound offset). The two source azimuths are 0° (target) and 45°. As can be seen in Fig. 2(b), the equalized response for the target source is far more impulselike compared to the room impulse response in Fig. 2(a). On the other hand, the impulse response corresponding to the interfering source is further smeared by the inverse filtering process, as seen in Fig. 2(d). Figure 3 illustrates the effect of reverberation as well as that of inverse filtering on the harmonic structure of a voiced utterance. The filters in Fig. 2 are convolved with an anechoic signal to generate the signals in Fig. 3. For a constant pitch contour, reverberation produces elongated tails but preserves the harmonicity. However, once the pitch varies reverberation smears the harmonic structure. For a given change in pitch frequency, higher harmonics vary their frequencies more rapidly compared to lower ones. Consequently, higher harmonics are more susceptible to reverberation as can be seen in Fig. 3(b). Figure 3(c) shows that an inverse filter is able to recover some of the harmonic components in the signal; for example, the harmonic series starting at about 1.0 s is more visible in Fig. 3(c) than in Fig. 3(b). To exemplify the smearing effect on the spectrum of an interfering source, we show the convolution of the same utterance with the filters corresponding to Figs. 2(c) and 2(d) and the results are given in Figs. 3(d) and 3(e), respectively.

Finally, the target inverse filter is applied on the reverberant mixture and the resulting signal feeds to the second stage of our model described below.

IV. PITCH-BASED SPEECH SEGREGATION

The proposed pitch-based segregation system uses a given target pitch track to group harmonically related components from the target source. Our system follows the segmentation and grouping steps of Hu and Wang (2004). However, we simplify their algorithm by extracting periodicities directly from the correlogram. Also, compared to the sinusoidal modeling scheme for computing AM rates in Hu and Wang (2004), our simplified method is more robust to intrusions in the high frequency range. A detailed description of our model is given below.

A. Auditory periphery and feature extraction

The signal is filtered through a bank of 128 fourth-order gammatone filters with center frequencies between 80 and 5000 Hz (Patterson *et al.*, 1988). In addition, envelopes are extracted for channels with center frequencies higher than 800 Hz. A Teager energy operator is applied to the signal to extract its envelope (Rouat *et al.*, 1997). This is defined as $E(n)=x^2(n)-x(n+1)x(n-1)$ for a signal $x(n)$, where n denotes the sampling step. Then, the signals are low-pass filtered at 800 Hz using a third-order Butterworth filter and high-pass filtered at 64 Hz.

The correlogram $A(c,j,\tau)$ for channel c, time-frame j, and lag τ is computed by the following autocorrelation using a window of 20 ms ($K=320$):

$$A(c,j,\tau)=\frac{\sum_{k=0}^{K} g(c,j-k)g(c,j-k-\tau)}{\sqrt{\sum_{k=0}^{K} g^2(c,j-k)}\sqrt{\sum_{k=0}^{K} g^2(c,j-k-\tau)}}, \quad (8)$$

where g is the gammatone filter output and the correlogram is updated every 10 ms. The range for τ corresponding to the plausible pitch range of 80 to 500 Hz is from 32 to 200. At high frequencies, the autocorrelation based on response envelopes reveals the amplitude modulation rate that coincides with the F0 for one periodic source. Hence,

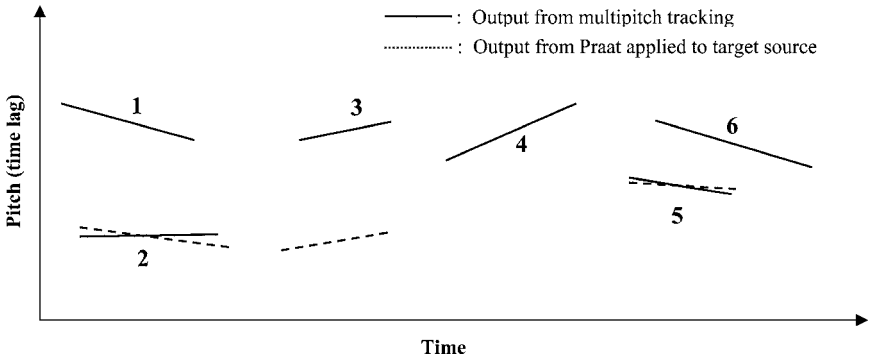

FIG. 4. Illustration of sequential organization. Solid lines illustrate a set of pitch contours from a multipitch tracking algorithm, each denoted by a number. Dashed lines show a set of pitch contours from Praat applied to the target signal before mixing. Note that these contours are drawn here for purposes of explanation, i.e., they are not actually produced from the algorithms. A comparison between the sets results in the selection of contours 2 and 5 as estimated target pitch contours.

an additional envelope correlogram $A_E(c,j,\tau)$ is computed for channels in the high-frequency range (>800 Hz) by replacing the filter output g in Eq. (8) with its extracted envelope. This correlogram representation of the acoustic signal has been successfully used in Wu *et al.* (2003) for multipitch analysis.

Finally, the cross-channel correlation between normalized autocorrelations in adjacent channels is computed in each T-F unit as

$$C(c,j) = \sum_{\tau=0}^{N-1} A(c,j,\tau)A(c+1,j,\tau), \tag{9}$$

where $N=200$ corresponds to the minimum pitch frequency of 80 Hz. Since adjacent channels activated by the same source tend to have similar autocorrelation responses, the cross-channel correlation has been used in previous segmentation studies (see, e.g., Wang and Brown, 1999). Similarly, envelope cross-channel correlation $C_E(c,j)$ is computed for channels in the high-frequency range (>800 Hz) to capture common amplitude modulation.

B. Unit labeling

A pitch-based segregation system requires a robust pitch detection algorithm. We employ the multipitch tracking (estimation) algorithm proposed by Wu *et al.* (2003) that gives good performance for a variety of intrusions. The system combines correlogram-based peak and channel selection within a statistical framework in order to form multiple tracks that correspond to different harmonic sources. When the interference is also a harmonic source, their system produces two pitch tracks each of which consists of a set of continuous pitch contours which do not overlap with each other, but the two sets may overlap in time; a pitch contour is a consecutive set of pitch points. The multipitch tracking system, however, does not address the issue of whether a particular pitch contour belongs to the target source or the interference. Assigning individual pitch contours to either the target or the interference is the issue of sequential organization (Bregman, 1990), and a challenging computational task which has been little addressed in previous CASA studies (Brown and Cooke, 1994; Hu and Wang, 2004). A recent study by Shao and Wang (2006) uses trained speaker models to address the sequential organization problem in the specific context of cochannel speech (two-speaker mixtures). In this paper, we do not attempt to address this problem and instead assume an "ideal" assignment for the two pitch tracks, i.e., an "ideal" binary decision for each of the contours in the contour union of the two tracks (as each track generally contains multiple contours). For this, an estimated pitch track from the target signal is extracted using Praat (Boersma and Weenink, 2002) and then used for the sole purpose of assigning whether an individual pitch contour corresponds to the target pitch track. This is explained in Fig. 4, which illustrates a set of pitch contours from the multipitch tracking algorithm of Wu *et al.* (2003) and the corresponding target pitch contours from Praat. The contours from the mixture data are marked as solid lines with numerical labels, while the target pitch contours from Praat are marked as dashed lines. In this situation, a comparison between the two sets results in the selection of contours 2 and 5 as estimated target pitch contours, which are used to group individual T-F units that belong to the target as described below. See Wu *et al.* (2003) for extensive treatment of multipitch tracking for noisy speech.

The labeling of an individual T-F unit is carried out by comparing the estimated target pitch with the periodicity of the correlogram. The correlogram has the well-known property that it exhibits a peak at the signal period as well as the multiples of the period. Note that an autocorrelation response is quasiperiodic due to the bandpass nature of a filter channel and the number of peaks in the correlogram increases with increasing center frequency of the channel. For a particular T-F unit, we should select the peak that best captures the periodicity of the underlying signal. In the low-frequency range, the system selects the peak for which the corresponding time lag l is the closest to the estimated target pitch lag p in $A(c,j,\tau)$. Statistics collected in individual channels show that the distribution of selected time lags is sharply centered around the target pitch lag and its variance decreases with increased center frequency. Hence, a T-F unit is discarded if the distance between the two lags $|p-l|$ exceeds a threshold θ_L. We have found empirically that a value of $\theta_L = 0.15(F_s/F_c)$ results in a good performance, where F_s is the sampling frequency and F_c is the center frequency of channel c. Finally, the peak height indicates the strength of the target signal in the mixture. The unit is thus labeled 1 if $A(c,j,l)$ is close to the maximum of $A(c,j,\tau)$ in the plausible pitch range

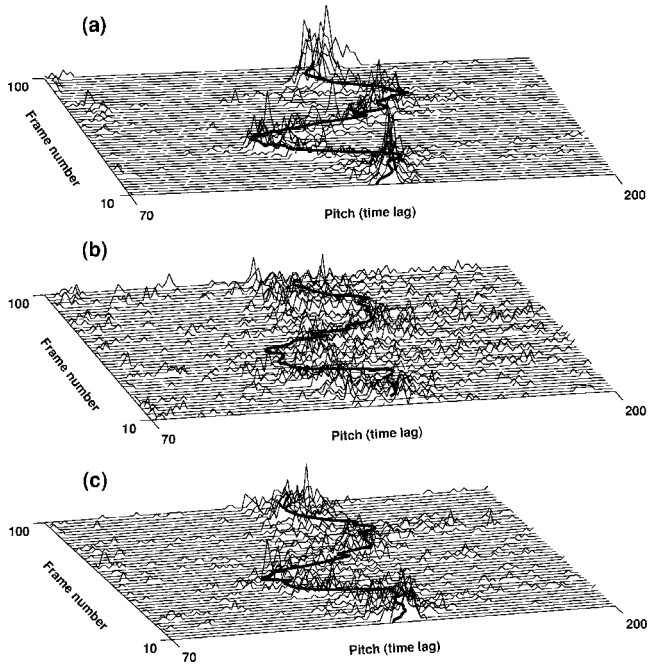

FIG. 5. Histograms of selected peaks in the high-frequency range (>800 Hz) for a male utterance. (a) Results for the anechoic signal. (b) Results for the reverberant signal. (c) Results for the inverse-filtered signal. The solid lines are the corresponding pitch tracks.

$$\frac{A(c,j,l)}{\max\limits_{\tau\in[32,200]} A(c,j,\tau)} > \theta_P, \tag{10}$$

where θ_P is fixed to 0.85. The unit is labeled 0 otherwise.

In the high-frequency range, we adapt the peak selection method of Wu *et al.* (2003). First, the envelope correlogram $A_E(c,j,\tau)$ of a periodic signal exhibits a peak both at the pitch lag and at the double of the pitch lag. Thus, the system selects all the peaks that satisfy the following condition: A peak with time lag l must have a corresponding peak that falls within the 5% interval around the double of l. If no peaks are selected, the T-F unit is labeled 0. Second, to deal with the situation where the pitch lag corresponding to the interference is half that of the target pitch, our system selects the first peak that is higher than half of the maximum peak in $A_E(c,j,\tau)$ for $\tau\in[32,200]$. Finally, the T-F unit is labeled 1 if the distance between the time lag corresponding to the selected peak and the estimated target pitch lag does not exceed a threshold of $\Delta=15$. The unit is labeled 0 otherwise. All the above parameters were optimized by using a small training set and found to generalize well over a test set.

The distortions on harmonic structure due to room reverberation are generally more severe in the high-frequency range. Figure 5 illustrates the effect of reverberation as well as inverse filtering in frequency channels above 800 Hz for a single male utterance. The filters in Figs. 2(a) and 2(b) are used to generate the reverberant signal and the inverse-filtered signal, respectively. At each time frame, we display the histogram of time lags corresponding to selected peaks. As can be seen from the figure, inverse filtering results in sharper peak distributions and improved harmonicity in comparison with the reverberant condition. The corresponding

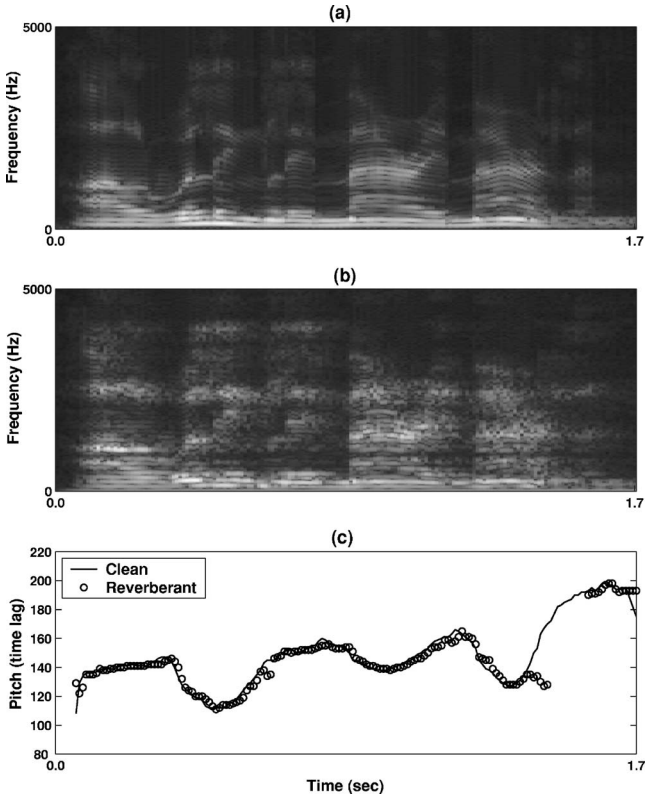

FIG. 6. Comparison of pitch tracking in anechoic and reverberant conditions for a male voiced utterance. (a) Spectrogram of the anechoic signal. (b) Spectrogram of the reverberant signal corresponding to the impulse response in Fig. 2(a). (c) Pitch tracking results. The solid line indicates the anechoic pitch track. The 'o' track indicates the reverberant track.

pitch tracks are extracted using Praat for each separate condition. To illustrate the effect of inverse filtering on the harmonic structure of the signals originating at the target location, we apply the T-F labeling described above to both the reverberant as well as the inverse-filtered male utterance. The signals are then reconstructed from the resulting T-F masks using the resynthesis method described in Sec. II. The reconstructed signal retains 79% of the target energy in the inverse-filtered condition compared to only 58% in the reverberant condition. As a reference, the corresponding labeling in the anechoic condition retains 94% of the target energy.

C. Segregation

The final segregation of the acoustic mixture into a target and a background stream is based on combined segmentation and grouping. A segment is a contiguous region of T-F units, each of which should be dominated by the same sound source. The main objective of the final segregation is to improve on the T-F unit labeling described above using segment-level features. The following steps follow the general segregation strategy in the Hu and Wang model (2004).

In the first step, segments are formed using temporal continuity and cross-channel correlation. Specifically, neighboring T-F units are iteratively merged into segments if their corresponding cross-channel correlation $C(c,j)$ exceeds a threshold $\theta_C=0.985$. The segments formed in this step are primarily located in the low-frequency range. A segment agrees with the target pitch at a given time frame if more

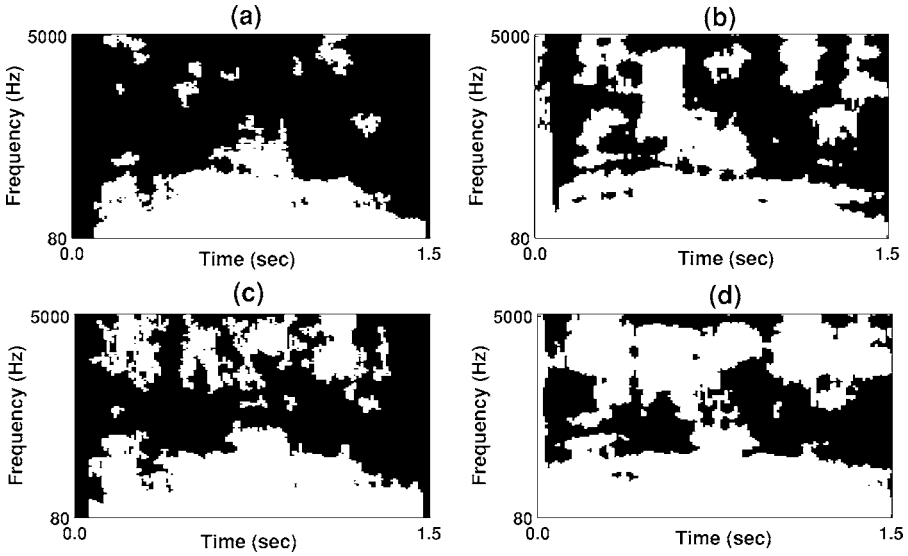

FIG. 7. Binary mask estimation for a mixture of target male utterance and interference female speech in reverberant and inverse-filtered conditions. (a) The estimated binary mask on the reverberant mixture. (b) The ideal binary mask for the reverberant condition. (c) The estimated binary mask on the filtered mixture. (d) The ideal binary mask for the inverse-filtered condition. The white regions indicate T-F units that equal 1 and the black regions indicate T-F units that equal 0.

than half of its T-F units are labeled 1. A segment that agrees with the target pitch for more than half of its length is grouped into the target stream; otherwise it goes to the background stream.

The second step primarily deals with potentially missing segments in the high-frequency range. Segments are formed by iteratively merging T-F units that are labeled 1 but not selected in the first step for which the envelope cross-channel correlation $C_E(c,j)$ exceeds the threshold θ_C. Segments shorter than 50 ms are removed. All these segments are grouped to the target stream.

The final step performs an adjustment of the target stream so that all T-F units in a segment bear the same label and no segments shorter than 50 ms are grouped. Furthermore, the target stream is iteratively expanded to include neighboring units that do not belong to either stream but are labeled 1.

With the T-F units belonging to the target stream labeled 1 and the other units labeled 0, the segregated target speech wave form is then resynthesized from the resulting binary T-F mask for systematic performance evaluation, to be discussed in the next section.

V. RESULTS

Two types of ASA cues that can potentially help a listener to segregate one talker in noisy conditions are location and pitch. Darwin and Hukin (2000) compared the effects of reverberation on spatial, prosodic, and vocal-tract size cues for a sequential organization task where the listener's ability to track a particular voice over time is examined. They found that while location cues are seriously impaired by reverberation, the F0 contour and vocal-tract length are more resistant cues. In our experiments, we have also observed that pitch tracking is robust to moderate levels of reverberation. To illustrate this, Fig. 6 compares the results of the pitch tracking algorithm of Wu *et al.* (2003) on a single male utterance in anechoic and reverberant conditions where T_{60}=0.35 s. The only distortions observed in the reverberant pitch track compared to the anechoic one are elongated tails and some deletions in the time frames where pitch changes rapidly.

Culling *et al.* (2003) have shown that while listeners are able to exploit the information conveyed by the F0 contour to separate a desired talker, the smearing of individual harmonics caused by reverberation degrades their separation capability. However, compared to location cues, the pitch cue degrades gradually with increasing reverberation and remains effective for speech separation (Culling *et al.*, 2003). In addition, as illustrated in Fig. 5, inverse filtering with respect to target location enhances signal harmonicity. We therefore assess the performance of two viable pitch-based strategies: (1) segregating the reverberant target from the reverberant mixture and (2) segregating the inverse-filtered target from the inverse-filtered mixture. Consequently, the speech segregation system described in Sec. IV is applied separately on the reverberant mixture and the inverse-filtered mixture.

To conduct a systematic SNR evaluation, a segregated signal is reconstructed from a binary mask following the method described in Sec. II. Given our computational objective of identifying T-F regions where the target is stronger than the interference, we use the signal reconstructed from the ideal binary mask as the ground truth to compute the output SNR (see Hu and Wang, 2004)

$$\mathrm{SNR}_{\mathrm{OUT}} = 10 \log_{10} \frac{\sum_t s_{\mathrm{IBM}}^2(t)}{\sum_t [s_{\mathrm{IBM}}(t) - s_E(t)]^2}, \tag{11}$$

where $s_{\mathrm{IBM}}(t)$ represents the target signal reconstructed using the ideal binary mask and $s_E(t)$ the estimated target reconstructed from the binary mask produced by our model. The input SNR is computed in the standard way as the ratio of target signal energy to noise signal energy expressed in decibels. Note that the target signal refers to the reverberant target signal in the reverberant condition and to the inverse-filtered signal in the inverse-filtered condition.

Figure 7 shows the binary masks produced by our system for a mixture of target male speech presented at 0° and interference female speech at 45°. Reverberant signals as

TABLE I. Output SNR results for target speech mixed with a female interference at three input SNR levels and different reverberation times.

Reverberation time (s)	−5 dB	0 dB	5 dB
Anechoic	8.78	11.61	13.93
T_{60}=0.05	7.25	8.54	10.65
T_{60}=0.10	7.35	8.16	9.46
T_{60}=0.15	6.37	7.09	8.24
T_{60}=0.20	5.59	6.52	7.39
T_{60}=0.25	4.74	6.06	6.79
T_{60}=0.30	4.47	5.57	6.22
T_{60}=0.35	4.55	5.36	6.13

well as inverse-filtered signals for both target and interference are produced by convolving the original anechoic utterances with the filters from Fig. 2. The signals are mixed to give an overall 0 dB input SNR in both conditions. The figure also displays the ideal binary masks. The results show an improved segregation capacity in the high frequency range in the inverse-filtered case [Fig. 7(c)] as compared to the reverberant case [Fig. 7(a)].

We perform the SNR evaluations using as target a set of ten voiced male sentences collected by Cooke (1993) for the purpose of evaluating voiced speech segregation systems. The following five noise intrusions are used: white noise; babble noise; a male utterance; music; and a female utterance. These intrusions represent typical acoustical interferences occurring in real environments. In all cases, the target is fixed at 0°. The babble noise is obtained by presenting natural speech utterances from the TIMIT database at the following eight separate directions around the target source: ±20°; ±45°; ±60°; and ±135°. For the other intrusions, the interfering source is located at 45°, unless otherwise specified. Also, the reverberation time for the experiments described below equals 0.35 s, unless otherwise specified. This reverberation time falls in the typical range for living rooms and office environments. When comparing the results between the two segregation strategies the target signal in each case is scaled to yield the desired input SNR. Each value in the following tables resents the average output SNR of one particular intrusion mixed with the ten target sentences.

We first analyze how pitch-based speech segregation is affected by reverberation. Table I shows the performance of our pitch-based segregation system applied directly on reverberant mixtures when T_{60} increases from 0.05 to 0.35 s. The mixtures are obtained using the female speech utterance as interference and three levels of input SNR: −5; 0; and 5 dB. The ideal pitch contours, not estimated ones, are used here for testing purposes. As expected, the system performance degrades gradually with increasing reverberation. Individual harmonics are increasingly smeared and this results in a gradual loss in energy, especially in the high-frequency range as illustrated also in Fig. 7. The decrease in output SNR for T_{60}=0.35 s compared to the anechoic condition ranges from 4.23 dB at −5 dB input SNR to 7.80 dB at 5 dB input SNR. Overall, however, the segregation algorithm provides consistent gains, showing the robustness of the pitch cue. Observe that a sizeable gain of 9.55 dB is obtained for the −5 dB input SNR even when T_{60}=0.35 s.

Now we analyze how the inverse-filtering stage impacts the overall performance. The results in Table II are given for both the reverberant case (Reverb) and inverse-filtered case (Inverse) at three input SNR levels: −5; 0; and 5 dB. The results are obtained using estimated pitch tracks as explained in Sec. IV B. The performance depends on input SNR and type of interference. A maximum improvement of 12.46 dB is obtained for the female interference at −5 dB input SNR. The proposed system (Inverse) has an average gain of 10.11 dB at −5 dB, 6.45 dB at 0 dB, and 2.55 dB at 5 dB. When compared to the reverberant condition a 2–3 dB improvement is observed for the male and female intrusions at all input SNR conditions. Almost no improvement is observed for white noise or babble noise. Moreover, inverse filtering decreases the system performance in the case of white noise at low SNRs because of the over-grouping of T-F units in the high-frequency range. For comparison, results using the ideal pitch tracks are presented in Table III. The improvement obtained by using ideal pitch tracks is small and shows that the pitch estimation method is accurate. We note that the variation in the output SNR values across different target sentences is relatively small—the standard deviation ranges from 1 to 2 dB—in both reverberant and inverse-filtered conditions.

As seen in the results presented above, the major advantage of the inverse-filtering stage occurs for a harmonic interference. In all the cases presented above the interfering source is located at 45°, and the inverse filtering stage further smears its harmonic structure. However, if the interfering source is located at a location near the target source the inverse filter will dereverberate the interference also. Table IV

TABLE II. Output SNR results using estimated pitch tracks for target speech mixed with different noise types at three input SNR levels and T_{60}=0.35 s. Target is at 0° and interference at 45°.

	−5 dB		0 dB		5 dB	
Input SNR	Reverb	Inverse	Reverb	Inverse	Reverb	Inverse
White noise	5.75	4.92	6.22	5.87	6.37	7.39
Babble noise	2.50	2.81	4.76	5.27	5.95	6.94
Male	0.67	4.54	3.96	6.68	5.76	7.76
Music	3.27	5.82	5.58	6.72	6.24	7.70
Female	4.87	7.46	5.51	7.70	6.13	7.95
Average	3.41	5.11	5.21	6.45	6.03	7.55

TABLE III. Output SNR results using ideal pitch tracks for target speech mixed with different noise types at three input SNR levels and T_{60}=0.35 s. Target is at 0° and interference at 45°.

	−5 dB		0 dB		5 dB	
Input SNR	Reverb	Inverse	Reverb	Inverse	Reverb	Inverse
White noise	5.94	5.38	6.19	6.10	6.37	7.56
Babble noise	3.25	4.23	5.14	5.71	5.95	7.40
Male	1.90	5.08	4.49	6.96	5.76	7.80
Music	3.89	6.25	5.73	6.93	6.24	7.80
Female	4.55	7.23	5.36	7.71	6.13	8.30
Average	3.90	5.63	5.38	6.68	6.09	7.77

shows SNR results for both white noise and female speech intrusions when the interference location is fixed at 0°, the same as the target location. As expected, in the white noise case, the results are similar to the ones presented in Table III. However, the relative improvement in output SNR obtained using inverse filtering is reduced to the range of 0.5–1 dB. This shows that smearing the harmonic structure of the interfering source plays an important role in boosting the segregation performance in the inverse-filtered condition.

As mentioned in Sec. I, this paper is the first study on monaural segregation of reverberant speech. As a result, it is difficult to quantitatively compare with existing systems. In an attempt to put our performance in perspective, we show a comparison with the spectral subtraction method, which is a standard speech enhancement technique (O'Shaughnessy, 2000). To apply spectral subtraction in practice requires robust estimation of interference spectrum. To put spectral subtraction in a favorable light, the average noise power spectrum is computed *a priori* within the silent periods of the target signal for each reverberant mixture. This average is used as the estimate of intrusion and is subtracted from the mixture. The SNR results are given in Table V, where the reverberant target signal is used as ground truth for the spectral subtraction algorithm and the inverse-filtered target signal is used as ground truth for our algorithm. As shown in the table, the spectral subtraction method performs significantly worse than our system, especially at low levels of input SNR. This is because of its well-known deficiency in dealing with nonstationary interferences. At 5 dB input SNR the spectral subtraction outperforms our system when the interference is white noise, babble noise, or music. In those cases of high-input SNR and relatively steady intrusion, the spectral subtraction algorithm tends to subtract little intrusion but it also introduces little distortion to the target signal. By comparison, our system focuses on target extraction that attempts to reconstruct the target signal on the basis of periodicity. Target components made inharmonic by reverberation are removed by our algorithm, thus introducing more target signal loss. It is worth noting that the ceiling performance of our algorithm without any interference is 8.89 dB output SNR.

VI. DISCUSSION

In natural settings, reverberation alters many of the acoustical properties of a sound source reaching our ears, including smearing its harmonic and temporal structures. Despite these alterations, moderate reverberant speech remains highly intelligible for normal-hearing listeners (Nabelek and Robinson, 1982). When multiple sound sources are active, however, reverberation adds another level of complexity to the acoustic scene. Not only does each interfering source constitute an additional masker for the desired source, but also does reverberation blur many of the cues that aid in source segregation. The recent results of Culling *et al.* (2003) suggest that reverberation degrades human ability to exploit differences in F0 between competing voices, producing a 5 dB increase in speech reception threshold for normally intonated sentences in monaural conditions.

We have investigated pitch-based monaural segregation in room reverberation and report the first systematic results on this challenging problem. We observe that pitch detection is relatively robust in moderate reverberation. However, the segregation capacity is reduced due to the smearing of the harmonic structure, resulting in gradual degradation in performance as the room reverberation time increases. As seen in Table I, compared to anechoic conditions there is an average decrement of 5.33 dB output SNR for a two-talker situation with T_{60}=0.35 s. This decrement is, however, consistent with the 5 dB increase in speech reception threshold reported by Culling *et al.* (2003).

TABLE IV. Output SNR results using ideal pitch tracks for target speech mixed with two types of noise at three input SNR levels and T_{60}=0.35 s. Target and interference are both located at 0°.

	−5 dB		0 dB		5 dB	
Input SNR	Reverb	Inverse	Reverb	Inverse	Reverb	Inverse
White noise	6.37	6.76	6.30	6.82	6.21	7.28
Female	4.82	5.51	5.74	6.65	6.28	7.57

TABLE V. Comparison between the proposed algorithm and spectral subtraction (SS). Results are obtained for target speech mixed with different noise types at three input SNR levels and T_{60}=0.35 s. Target is at 0° and interference at 45°.

	−5 dB		0 dB		5 dB	
Input SNR	SS	Proposed	SS	Proposed	SS	Proposed
White noise	2.40	3.36	6.54	4.93	10.47	6.48
Babble noise	−2.76	2.74	1.98	4.66	6.65	6.42
Male	−4.05	4.11	0.77	6.17	5.59	7.24
Music	−1.37	4.45	3.22	6.01	7.68	7.07
Female	−3.31	5.40	1.46	6.71	6.19	7.56
Average	−1.81	4.01	2.79	5.69	7.31	6.95

To reduce the smearing effects on the target speech, we have proposed a preprocessing stage which equalizes the room impulse response corresponding to target location. This preprocessing results in both improved harmonicity for signals arriving from the target direction and smearing of competing sources at other directions. We have found that this effect provides a better input signal for pitch-based segregation. The extensive evaluations show that our system yields substantial SNR gains across a variety of noise conditions. Our previous study shows a strong correlation between SNR gains measured against the ideal binary mask and improvements in automatic speech recognition and speech intelligibility scores (Roman *et al.*, 2003). Hence we expect similar improvements for the SNR gains achieved in the present study, although further evaluation is required to substantiate this projection.

The improvement in speech segregation obtained in the inverse-filtering case is limited by the accuracy of the estimated inverse filter. In our study, we have employed an algorithm that estimates the inverse filter directly from reverberant speech data. When the room impulse response is known, better inverse-filtering methods exist, e.g., the linear least square equalizer by Gillespie and Atlas (2002). This type of preprocessing leads to increased target signal fidelity and thus produces large improvements in speech segregation. In terms of applications to real-world scenarios our inverse-filtering faces several drawbacks. First, the adaptation of the inverse filter requires data on the order of a few seconds and thus any fast change in the environment (e.g., head movements and walking) will have an adverse impact on the inverse-filtering stage. Second, this stage needs to perform filter adaptation in the presence of no or weak interference. On the other hand, our pitch-based segregation stage can be applied without such limitations. Hence, whenever the adaptation of the inverse filter is infeasible, one can still apply our pitch-based segregation algorithm directly on the reverberant mixture.

Speech segregation in high input SNR conditions presents a challenge to our system. We employ a figure-ground segregation strategy that attempts to reconstruct the target signal by grouping harmonic components. Consequently, inharmonic target components are removed by our approach even in the absence of interference. While this problem is common in both anechoic and reverberant conditions, it worsens in reverberation due to the smearing of harmonicity. To address this issue probably requires examining the inharmonicity induced by reverberation and distinguishing such inharmonicity from that caused by additive noise. This is a topic of further investigation.

In the segregation stage, our system utilizes only pitch cues and thus is limited to the segregation of voiced speech. Other ASA cues such as onsets, offsets, and acoustic-phonetic properties of speech are also important for monaural separation (Bregman, 1990). Recent research has shown that these cues can be used to separate unvoiced speech (Hu and Wang, 2003; 2005). Future work will need to address unvoiced separation in reverberant conditions. Another limitation, already mentioned in Sec. IV B, concerns sequential grouping. Like previous studies, our system avoids this issue by assuming an "ideal" assignment of estimated pitch contours. Although some progress has been made on sequential grouping of cochannel speech (e.g., Shao and Wang, 2006), the general problem of sequential organization remains a considerable challenge in CASA.

ACKNOWLEDGMENTS

This research was supported in part by an AFOSR grant (FA9550-04-1-0117) and an NSF grant (IIS-0081058).

Allen, J. B., and Berkley, D. A. (**1979**). "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**, 943–950.

Barros, A. K., Rutkowski, T., Itakura, F., and Ohnishi, N. (**2002**). "Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets," IEEE Trans. Neural Netw. **13**, 888–893.

Balan, R., Jourjine, A., and Rosca, J. (**1999**). "AR processes and sources can be reconstructed from degenerate mixtures," *Proc. 1st Int. Workshop on Independent Component Analysis and Signal Separation*, pp. 467–472.

Boersma, P., and Weenink, D. (**2002**). *Praat: doing Phonetics by Computer*, Version 4.0.26 (http://www.fon.hum.uva.nl/praat).

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

Bronkhorst, A. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acustica **86**, 117–128.

Brown, G. J., and Cooke, M. (**1994**). "Computational auditory scene analysis," Comput. Speech Lang. **8**, 297–336.

Brown, G. J., and Wang, D. L. (**2005**). "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, eds. (Springer, New York), pp. 371–402.

Brungart, D., Chang, P., Simpson, B., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask," (unpublished).

Burkhard, M. D., and Sachs, R. M. (**1975**). "Anthropometric manikin for

acoustic research," J. Acoust. Soc. Am. **58**, 214–222.
Cooke, M. P. (**1993**). *Modeling Auditory Processing and Organization* (Cambridge University Press, Cambridge, U.K).
Cooke, M. P., Green, P., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Commun. **34**, 267–285.
Culling, J. F., Hodder, K. I., and Toh, C. Y. (**2003**). "Effects of reverberation on perceptual segregation of competing voices," J. Acoust. Soc. Am. **114**, 2871–2876.
Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory grouping," in *The Handbook of Perception and Cognition*, vol. 6, B. C. J. Moore, ed. (Academic, London), pp. 387–424.
Darwin, C. J., and Hukin, R. W. (**2000**). "Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention," J. Acoust. Soc. Am. **108**, 335–342.
Ephraim, Y., and Trees, H. L. (**1995**). "A signal subspace approach for speech enhancement," IEEE Trans. Speech Audio Process. **3**, 251–266.
Furuya, K., and Kaneda, Y. (**1997**). "Two-channel blind deconvolution for non-minimum phase impulse responses," *Proc. ICASSP*, pp. 1315–1318.
Gardner, W. G., and Martin, K. D. (**1994**). "HRTF measurements of a KEMAR dummy-head microphone," MIT Media Lab Perceptual Computing Technical Report #280.
Gillespie, B. W., and Atlas, L. E. (**2002**). "Acoustic diversity for improved speech recognition in reverberant environments," *Proc. ICASSP*, pp. 557–560.
Gillespie, B. W., Malvar, H. S., and Florencio, D. A. F. (**2001**). "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Proc. ICASSP*, vol. 6, pp. 3701–3704.
Haykin, S. (**2002**). *Adaptive Filter Theory*, 4th ed. (Prentice-Hall, Upper Saddle River, NJ).
Hu, G., and Wang, D. L. (**2003**). "Separation of stop consonants," *Proc. ICASSP*, vol. 2, pp. 749–752.
Hu, G., and Wang, D. L. (**2004**). "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Netw. **15**, 1135–1150.
Hu, G., and Wang, D. L. (**2005**). "Separation of fricatives and affricates," *Proc. ICASSP* vol. 1, pp. 1101–1104.
Jang, G.-J., Lee, T.-W., and Oh, Y.-H. (**2003**). "Single channel signal separation using time-domain basis functions" IEEE Signal Process. Lett. **10**(6), 168–171.
Licklider, J. C. R. (**1951**). "A duplex theory of pitch perception," Experientia **7**, 128–134.
Luo, H. Y., and Denbigh, P. N. (**1994**). "A speech separation system that is robust to reverberation," *Proc. ISSIPNN*, pp. 339–342.
Ma, N., Bouchard, M., and Goubran, R. (**2004**). "Perceptual Kalman filtering for speech enhancement in colored noise," *Proc. ICASSP*, vol. 1, pp. 717–720.
Martin, R. (**2001**). "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Process. **9**, 504–512.
Moore, B. C. J. (**2003**). *An Introduction to the Psychology of Hearing*, 5th ed. (Academic, San Diego, CA).
Nabelek, A. K., and Robinson, P. K. (**1982**). "Monaural and binaural speech perception in reverberation for listeners of various ages," J. Acoust. Soc. Am. **71**, 1242–1248.
Nakatani, T., and Miyoshi, M. (**2003**). "Blind dereverberation of single channel speech signal based on harmonic structure," *Proc. ICASSP*, pp. 92–95.
Neely, S. T., and Allen, J. B. (**1979**). "Invertibility of a room impulse response," J. Acoust. Soc. Am. **66**, 165–169.
O' Shaughnessy, D. (**2000**). *Speech Communications: Human and Machine*, 2nd ed. (Piscataway, IEEE Press, NJ).
Palomaki, K. J., Brown, G. J., and Wang, D. L. (**2004**). "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," Speech Commun. **43**, 361–378.
Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Price, P. (**1988**). "APU Report 2341: An efficient auditory filterbank based on the gammatone function," Applied Psychology Unit, Cambridge.
Plomp, R. (**1976**). "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of a single competing sound source (speech or noise)," Acustica **34**, 200–211.
Roman, N., Wang, D. L., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.
Rouat, J., Liu, Y. C., and Morissette, D. (**1997**). "A pitch determination and voice/unvoiced decision algorithm for noisy speech," Speech Commun. **21**, 191–207.
Shao, Y., and Wang, D. L. (**2006**). "Model-based sequential organization in cochannel speech," IEEE Trans. Audio, Speech, Lang. Proc. **14**, 289–298.
Shamsoddini, A., and Denbigh, P. N. (**2001**). "A sound segregation algorithm for reverberant conditions," Speech Commun. **33**, 179–196.
Slaney, M., and Lyon, R. F. (**1993**). "On the importance of time—A temporal representation of sound," in *Visual Representations of Speech Signals*, M. P. Cooke, S. Beet, and M. Crawford, eds. (Wiley, New York), pp. 95–116.
Srinivasan, S., Roman, N., and Wang, D. L. (**2004**). "On binary and ratio time-frequency masks for robust speech recognition," *Proc. ICSLP*, pp. 2541–2544.
Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, ed. (Kluwer Academic, Norwell, MA), pp. 181–197.
Wang, D. L., and Brown, G. J. (**1999**). "Separation of speech from interfering sounds based on oscillatory correlation," IEEE Trans. Neural Netw. **10**, 684–697.
Weintraub, M. (**1985**). "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Stanford University Department of Electrical Engineering.
Wu, M. (**2003**). "Pitch tracking and speech enhancement in noisy and reverberant environments," PhD thesis, The Ohio State University, Department of Computer and Information Science.
Wu, M., and Wang, D. L. (**2006**). "A two-stage algorithm for one-microphone reverberant speech enhancement," IEEE Trans. Audio, Speech, Lang. Proc. **10**, 774–784.
Wu, M., Wang, D. L., and Brown, G. J. (**2003**) "A multipitch tracking algorithm for noisy speech," IEEE Trans. Speech Audio Process. **11**, 229–241.
Zibulevsky, M., Pearlmutter, B. A., Bofill, P., and Kisilev, P. (**2001**). "Blind source separation by sparse decomposition," in *Independent Component Analysis: Principles and Practice*, S. J. Roberts and R. M. Everson, eds. (Cambridge University Press, Cambridge).