Technical Report OSU-CISRC-4/05-TR22
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftp site: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2005**
File: **TR22.pdf**
Web site: **http://www.cse.ohio-state.edu/research/tech-report.html**

Pitch-Based Segregation of Reverberant Speech

Nicoleta Roman
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
niki@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive Science
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

Correspondence should be directed to D. Wang: Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210. Phone: (614)-292-6827, URL: www.cis.ohio-state.edu/~dwang.

ABSTRACT

In everyday listening, both background noise and reverberation degrade the speech signal. Psychoacoustic evidence suggests that human speech perception under reverberant conditions relies primarily 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 a degradation in performance due to the loss in periodicity for 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 with 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 locations 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.

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 eliminated by reverberation (Plomp, 1976; Culling et al., 2003) which emphasizes 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 (AR) 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) was able 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 enhancement 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; Nakatani and Okuno, 1998; Shamsoddini and Denbigh, 1999; 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 when they are presented 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, 2004). 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 ideal binary masks has been shown to be highly intelligible even when extracted from multi-source mixtures and also to produce substantial improvements in robust speech recognition (Cooke et al., 2001; Roman et al., 2003; Brungart et al., 2005).

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) exploits a differential strategy 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 system shows good performance when tested with a variety of noise intrusions under anechoic conditions. However, when the pitch varies with time in a

reverberant environment, reflected waves with different F0s arrive simultaneously with the direct sound at the ear. This multipath situation causes smearing of the signal in the sense that harmonic structure is less clear in the signal (Darwin and Hukin, 2000). Due to the loss of 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, for one-microphone recordings, perfect reconstruction exists only if the original room impulse response is a minimum-phase filter (Oppenheim and Schafer, 1989). This requirement is almost never satisfied in practical conditions. On the other hand, exact inverse filtering can be obtained using multiple microphones by assuming no common zeros among the different room impulse responses (Miyoshi and Kaneda, 1988). Inverse filtering techniques which partially dereverberate the reverberant signal have also been studied (Gillespie and Atlas, 2002). However, these algorithms assume *a priori* knowledge of the room impulse responses, which is often impractical. Several strategies 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 restoration of LP residuals results in both a reduction of perceived reverberation as well as an improvement of spectral fidelity in terms of harmonicity. In this paper, we employ a one-microphone adaptation of this strategy proposed by Wu (2003; Wu and Wang, 2005).

The dereverberation algorithms described above are designed to enhance a single reverberant source. Here, we investigate the effect of inverse filtering as pre-processing for a pitch-based speech segregation system in order to improve its robustness in a reverberant environment. The key idea is to estimate the filter that inverts the room impulse response corresponding to the target source. The effect of applying this inverse filter on the reverberant mixture is two-fold: it improves the harmonic structure of 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 the proposed pitch-based system. To our knowledge, the proposed system 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 employed in this paper. Section IV gives a detailed description of the proposed 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.

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

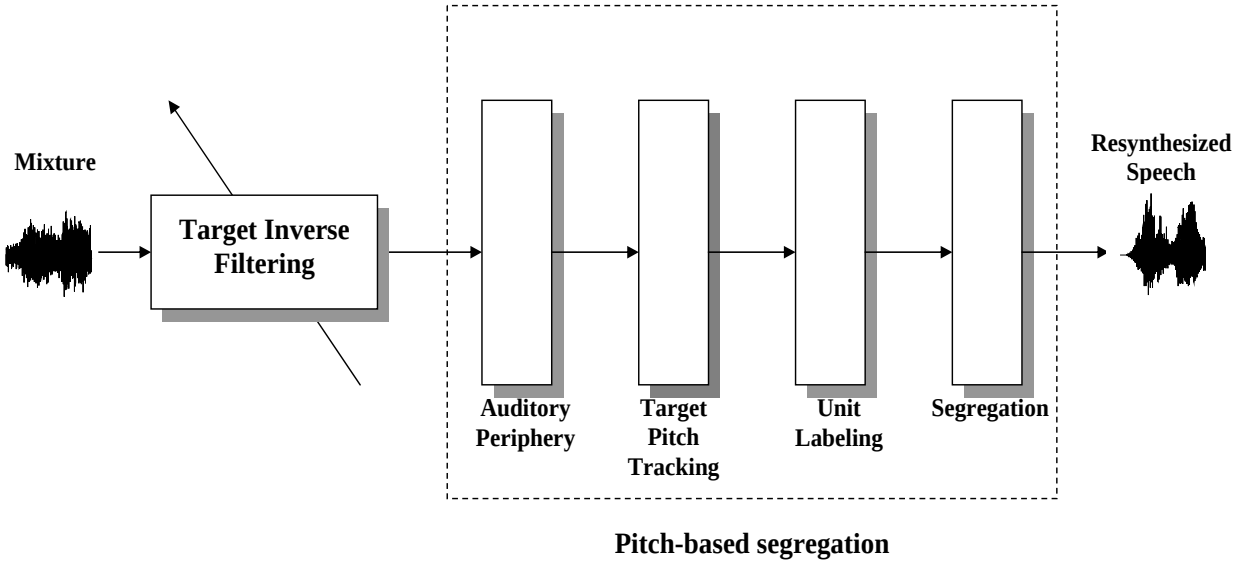

Figure 1. Schematic diagram of the proposed two-stage model.

The input to our model is a monaural mixture of two or more sound sources in a small reverberant room ($6\text{m} \times 4\text{m} \times 3\text{m}$). The receiver - the left ear of a KEMAR dummy head (Burkhard and Sachs, 1975) - is fixed at ($2.5\text{m} \times 2.5\text{m} \times 2\text{m}$) while the acoustic sources are located at a distance of 1.5 m from the receiver. The impulse response modeling the acoustic transfer function from one 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 of the KEMAR dummy head (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 in their structure. The reverberant signals are then obtained by convolving the original clean signals with the corresponding room impulse responses. Finally, signals are added together and sampled at 16 kHz.

In the first stage, a finite impulse response filter is estimated that inverts the target room impulse response $h(t)$. Adaptive filtering strategies for estimating this filter are sensitive to background noise (Haykin, 2002). For simplicity, here we perform this estimation during an

initial training stage in the absence of noise. We employ the inverse filtering strategy proposed by Gillespie et al. (2001), which is a practical system using a relatively small amount of training data. This method exploits the fact that the signal to be recovered is speech by employing an LP-based metric and produces improved harmonicity for the target source. The inverse filter is applied to the entire mixture and the result is fed to the next stage.

In the second stage, a pitch-based segregation system is employed to separate the inverse-filtered target signal from other interfering sounds. The signal is analyzed using a gammatone auditory filterbank (Patterson et al., 1988) in consecutive time frames to produce a time-frequency decomposition. A standard mechanism for periodicity extraction employs a correlogram which is a collection of autocorrelation functions computed at individual filters (Licklider, 1951; Slaney and Lyon, 1993). For a particular T-F unit in the low-frequency range, the autocorrelation faithfully encodes its periodicity information. In the high-frequency range, the filters have a wide bandwidth and multiple harmonics activate the same filter, thus creating beats at a rate corresponding to the fundamental period (Helmholtz, 1863). Such amplitude modulation can be detected using the envelope-based autocorrelation. Our system employs a peak selection mechanism to reveal likely periodicities in the autocorrelation functions of individual T-F units. Further, the system decides whether the underlying target is stronger than the combined interference by comparing these periodicities with a given target pitch.

However, labeling at the T-F unit level is a very local decision and 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). By definition, a segment is composed of spatially contiguous units dominated by a single sound source. Hence, grouping at the segment level improves the system robustness compared to the simple T-F labeling. Here, we combine the unit labeling described above with the segmentation framework proposed by Hu and Wang (2004). First, segments in the low-frequency range are generated using cross-channel correlation and temporal continuity. These segments are grouped into a target stream and a background stream according to the labeling of their T-F components. Similarly, segments are added to the target stream in the high-frequency range using envelope-based cross-channel correlation. The result of this process is a binary mask that assigns 1 to all the T-F units in the target stream and 0 otherwise.

Finally, a speech waveform 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 used to retain the acoustic energy from the mixture that corresponds to 1's in the mask and nullifies the others. This method achieves high-quality reconstruction.

II. TARGET INVERSE FILTERING

As described in the introduction, inverse filtering is a standard method used for deriving the original target signal. We employ the method proposed by Gillespie et al. (2001) which attempts to blindly estimate the inverse filter from reverberant speech data. Based on the observation that peaks in the LP residual of speech are smeared under reverberation, an online adaptive algorithm

estimates the inverse filter by maximizing the kurtosis of the inverse-filtered LP residual of reverberant speech $\tilde{z}(t)$:

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

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 $\boldsymbol{q}$ 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 $\boldsymbol{q}$ can be approximated as follows (Gillespie et al., 2001):

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

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

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

where $\hat{\boldsymbol{q}}(t)$ 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 equation (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 implementation for one microphone signals described by Wu and Wang (2005). 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 using the following update equations:

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

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

where $\mathbf{F}(m)$ and $\mathbf{Y}_r(m)$ represent the FFT of $f(t)$ and $\mathbf{y}_r(t)$ for the mth block, and $\boldsymbol{Q}(n)$ represents the estimate for the FFT of inverse filter $\boldsymbol{q}$ at iteration n. M represents the number of blocks and the superscript * is 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.

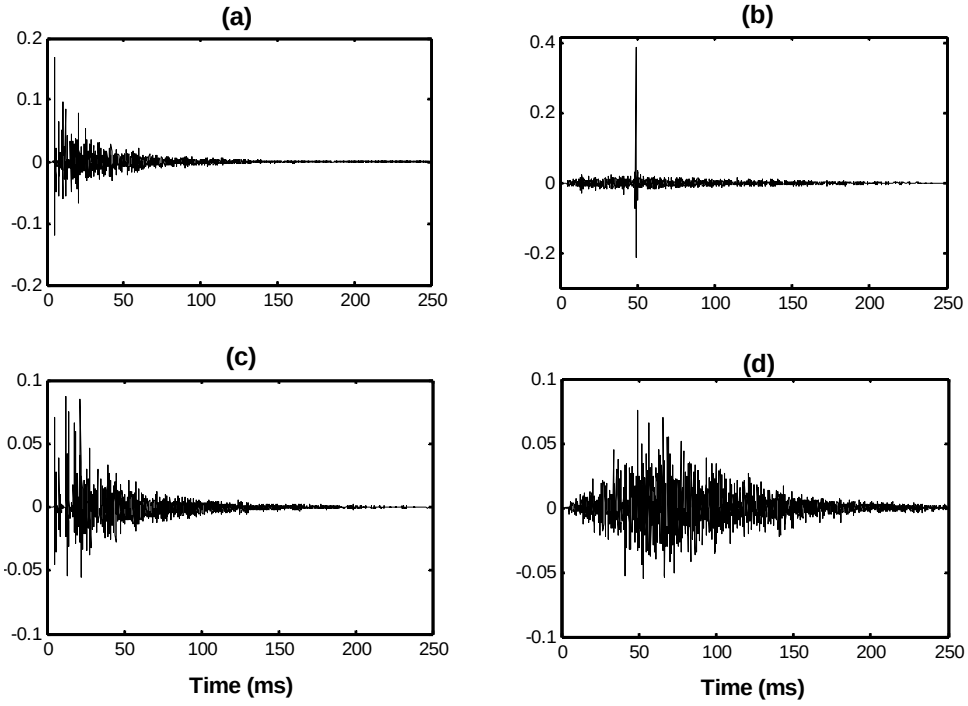

Figure 2. Effects of inverse filtering on room impulse responses. **(a)** A room impulse response for a target source presented in the median plane. **(b)** The effect of convolving the impulse response in (a) with an estimated inverse filter. **(c)** A room impulse response for one interfering source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

Figure 2 shows the outcome of convolving an estimated inverse filter with both the target room impulse response as well as the room 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 locations are 0° (target) and 45°. As can be seen in Fig. 2(b), the equalized response for the target source is far more impulse-like 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). Fig. 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 a clean signal to generate the signals in Fig. 3. For a constant pitch contour, reverberation produces elongated tails but

largely preserves the harmonicity. However, once the pitch changes 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. 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 Fig. 2(c) and Fig. 2(d) and the results are given in Fig. 3(d) and Fig. 3(e), respectively.

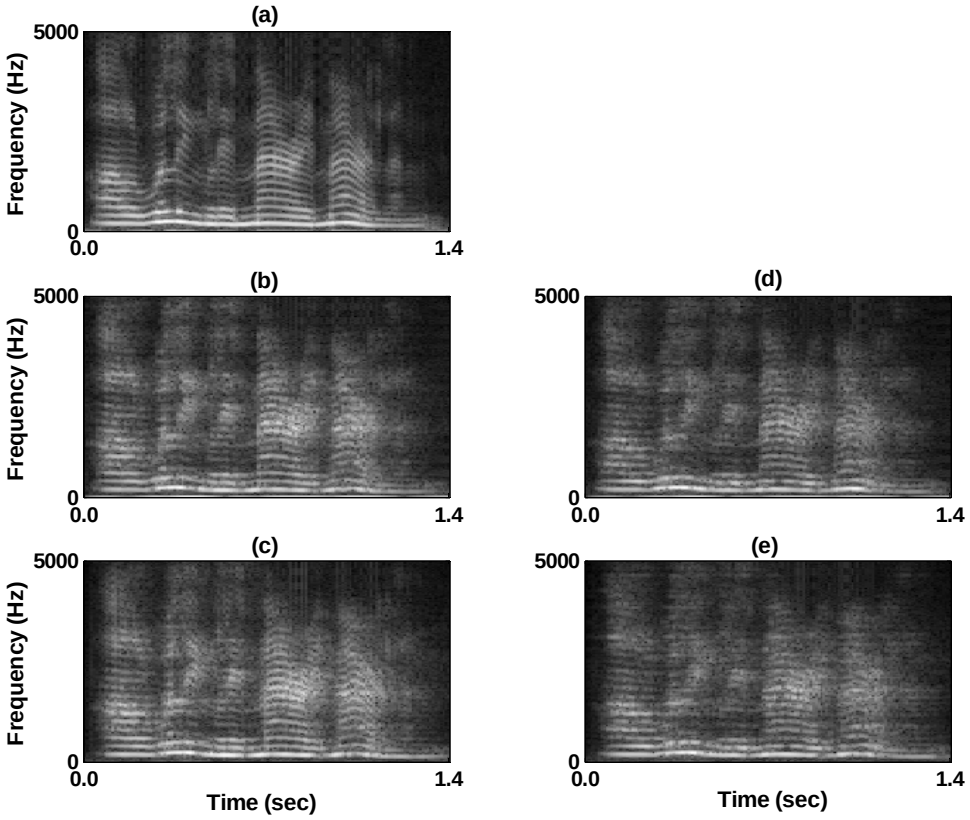

Figure 3. Effects of reverberation and target inverse filtering on the harmonic structure of a voiced utterance. **(a)** Spectrogram of the anechoic signal. **(b)** Spectrogram of the reverberant signal corresponding to the impulse response in Fig. 2(a). (**c**) Spectrogram of the inverse-filtered signal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

Finally, the target inverse filter is applied on the reverberant mixture composed of both target speech and interference and the resulting signal feeds to the second stage of our model.

III. PITCH-BASED SPEECH SEGREGATION

The proposed pitch-based speech segregation system uses a given target pitch contour to group harmonically related components from the target source. Our system follows the principles of segmentation and grouping from the system of Hu and Wang (2004). However, we simplify their algorithm by extracting periodicities directly from the correlogram. Also, compared to the sinusoidal modeling approach used for computing AM rates in Hu and Wang (2004), our simplified implementation is more robust to intrusions in the high frequency range resulting in more efficient T-F unit labeling. A detailed description of the 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 aligned using the equivalent rectangular bandwidth (ERB) between 80 and 5000 Hz (Patterson et al., 1988). In addition, envelopes are extracted for channels with center frequencies higher than 800 Hz as used by Rouat et al. (1997). A Teager energy operator is applied to the signal. This is defined as $E(t) = x^2(t) - x(t+1)x(t-1)$ for a signal $x(t)$. 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):

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

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 Hz 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 fundamental frequency for one periodic source. Hence, an additional envelope-based correlogram $A_E(c, j, \tau)$ is computed for channels in the high-frequency range (>800 Hz) by replacing the filter output g in equation (8) with its extracted envelope. This correlogram representation of the acoustic signal has been successfully used in Wu et al. (2003) for multi-pitch 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 as an effective feature in previous segmentation studies (see e. g. Wang and Brown, 1999). Similarly, envelope-based cross-channel correlation $C_E(c, j)$ is computed for channels in the high-frequency range (>800 Hz) to capture the amplitude modulation rate.

B. Unit Labeling

A pitch-based segregation system requires a robust pitch detection algorithm. We employ here the multi-pitch tracking algorithm proposed by Wu et al. (2003) that produces up to two pitch contours and has shown good performance for a variety of intrusions. The system combines correlogram-based pitch and channel selection mechanisms within a statistical framework in order to form multiple tracks that correspond to the active sources in the acoustic scene. However, the assignment of the overlapping pitch contours is needed when the interference also has harmonic structure. For this, the 'ideal' pitch contour is extracted using Praat (Boersma and Weenink, 2002) from the target signals and used as the ground truth for the sole purpose of deciding which of two overlapping pitch contours belongs to the target utterance. The resulting estimated pitch track is used for identifying individual T-F units that belong to target as described below.

The labeling of an individual T-F unit is carried out by comparing the target pitch lag p with the periodicity of the normalized correlogram. In the low-frequency range, the system selects the time lag l that corresponds to the closest peak in $A(c, j, \tau)$ from the pitch lag. For a particular channel, the distribution of the selected time lags is sharply centered around the pitch lag and its variance decreases as the channel center frequency increases. Here, 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 unit is labeled 1 if $A(c, j, l)$ is close to the maximum of $A(c, j, \tau)$ in the plausible pitch range:

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

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

In the high-frequency range, we adapt the peak selection mechanism developed by Wu et al. (2003). First, the envelope correlogam $A_E(c, j, \tau)$ 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, a harmonic interference introduces peaks at lags around the multiples of its pitch lag. Therefore, our system selects the first peak that is higher than half of the maximum peak in $A_E(c, j, \tau)$ for $\tau \in [32, 200]$. The T-F unit is labeled then 1 if the distance between the time lag of the selected

peak and the target pitch lag does not exceed a threshold $\Delta = 15$, 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.

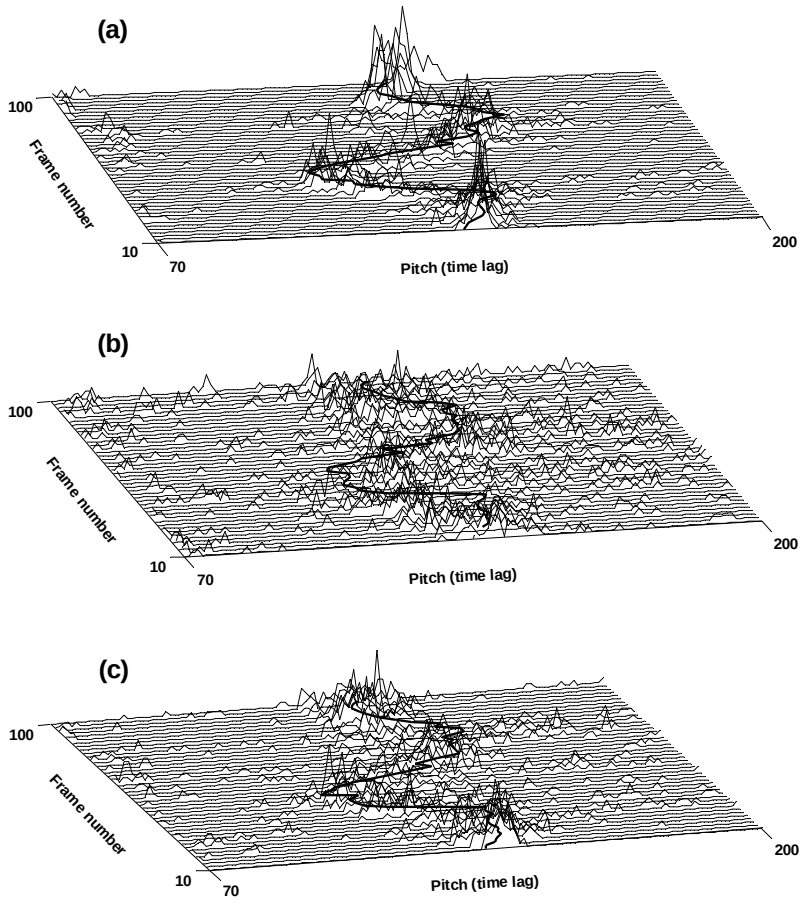

Figure 4. Histograms of selected peaks in the high-frequency range (>800 Hz) for a male utterance. **(a)** Results for the clean signal. **(b)** Results for the reverberant signal. **(c)** Results for the inverse filtered signal. The solid lines are the corresponding pitch contours.

The distortions on harmonic structure due to room reverberation are generally more salient in the high-frequency range**.** Figure 4 illustrates the effect of reverberation as well as inverse filtering in frequency channels above 800 Hz for a single male utterance. The filters in Fig. 2(a) and Fig. 2(b) are used to simulate 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 pitch contours are extracted using Praat (Boersma and Weenink, 2002) for each separate condition. From a different measure, the channel selection mechanism retains 79 percent of the total signal

energy by applying inverse-filtering as compared to 58 percent without inverse filtering. As a reference, the system retains 94 percent signal energy in the anechoic condition.

C. Segregation

The final segregation of the acoustic mixture into a target and a background stream is based on combined segmentation and grouping. The main objective is to improve on the pitch-based T-F unit labeling described above using segment-level features. The following steps follow the general segregation strategy from the Hu and Wang model (2004).

In the first step, segments are formed using temporal continuity and the gammatone-based 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 θ_C=0.985 (Hu and Wang, 2004). The segments formed at this stage are primarily located in the low-frequency range. A segment agrees with the target pitch at a given time frame if more 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 is grouped in the background stream.

The second step primarily deals with potential 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 (Hu and Wang, 2004). All these segments are grouped to the target stream.

This 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 present. 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 waveform can then be resynthesized from the resulting binary T-F mask for systematic performance evaluation, to be discussed in the next section.

IV. RESULTS

Two types of ASA cues that can potentially help a listener to segregate one talker in noisy conditions are: localization 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 also observe that pitch tracking is robust to moderate levels of reverberation. To illustrate this, Figure 5 compares the results of a pitch tracking algorithm (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 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 in reverberation degrades this 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. 4, inverse filtering with respect to target location improves 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 Section IV is applied separately on the reverberant mixture and the inverse-filtered mixture. As described in Section II, we have evaluated the system on the left-ear response of a KEMAR dummy head, using a room acoustic model implemented by Palomaki et al. (2004). In addition, the inverse filter of the target room impulse response is obtained from training data as explained in Section III and applied on the whole reverberant mixture to obtain the inverse filtered mixture.

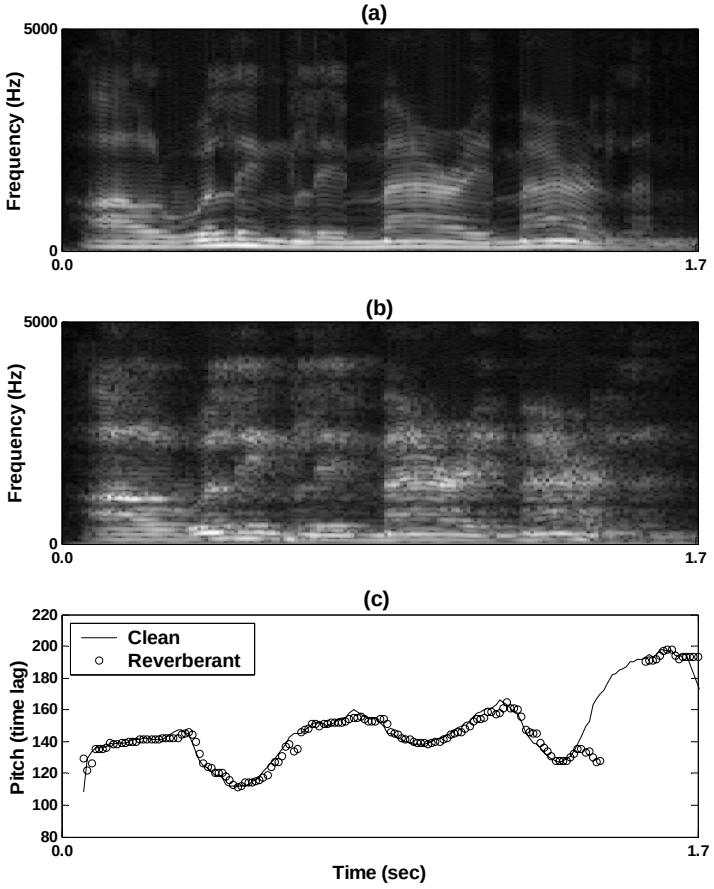

Figure 5. Comparison of pitch tracking in anechoic and reverberant conditions for a male voiced utterance. **(a)** Spectrogram of the anechoic signal. **(b)** Spectrogram of the reverberant signal corresponding to the impulse response in Fig. 2(a). **(c)** Pitch tracking results. The solid line indicates the anechoic pitch track. The 'o' track indicates the reverberant track.

Figure 6 shows the binary masks obtained for a mixture of target male speech presented at 0° and interference female speech at 45°. Reverberant signals as 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 SNR in both conditions. The ideal binary mask is constructed from the premixing target and intrusion as follows: a T-F unit in the mask is assigned 1 if the target energy in the unit is greater than the intrusion energy and 0 otherwise. This corresponds to a 0 dB local SNR criteria for ideal mask generation (see Brungart et al., 2005). The figure shows an improved segregation capacity in the high frequency range in the inverse-filtered case (Fig. 6 (c)) as compared to the reverberant case (Fig. 6 (a)).

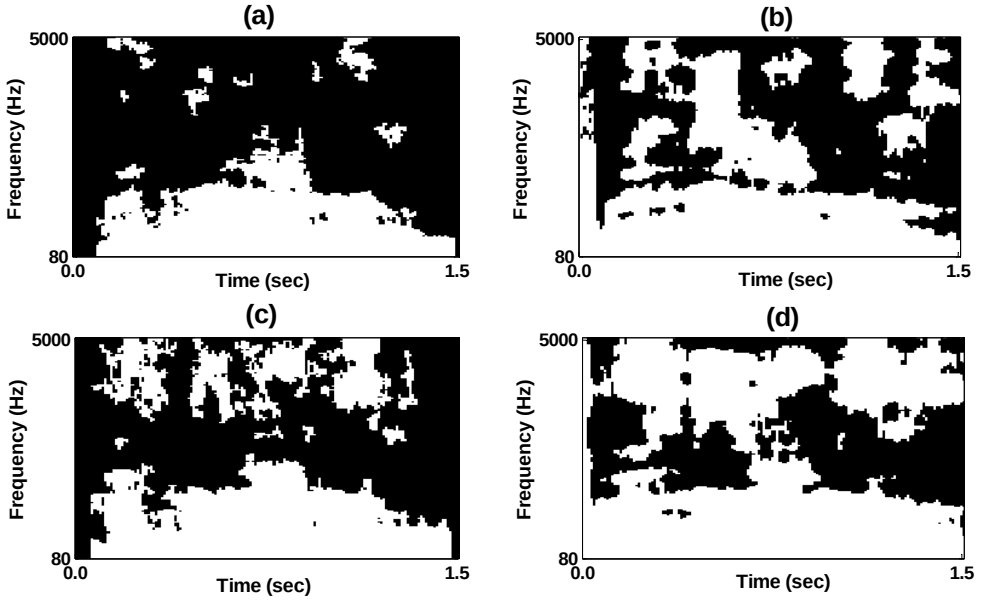

Figure 6. Binary mask estimation for a mixture of target male utterance and interference female speech in reverberant and inverse-filtered conditions. **(a)** The estimated binary mask on the reverberant mixture. **(b)** The ideal binary mask for the reverberant condition. **(c)** The estimated binary 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.

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

$$SNR = 10\log 10 \frac{\sum_t s_{IBM}^2(t)}{\sum_t (s_{IBM}(t) - s_E(t))^2}, \tag{11}$$

where $s_{IBM}(t)$ 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.

We perform the SNR evaluations using as target the set of 10 voiced male sentences collected by Cooke (1993) for the purpose of evaluating voiced speech segregation systems. The following 5 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, target is fixed at 0°. The babble noise is obtained by presenting natural speech utterances from the TIMIT database at the following 8 separated positions around the target source: $\pm 20^o, \pm 45^o, \pm 60^o, \pm 135^o$. 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 strategies the target signal in each case is scaled to yield a desired input SNR. Each value in the following tables represents the average output SNR of one particular intrusion mixed with the 10 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 s to 0.35 s. The mixtures are obtained using the female speech utterance as interference and three levels of input SNR: -5 dB, 0 dB, 5 dB. The ideal pitch contours are used here to generate the results. As expected, the system performance degrades gradually with increasing reverberation. The 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. 6. The decrease in performance 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 across a range of reverberation times, 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.

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

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

Now we analyze how inverse-filtering pre-processing impacts the overall performance of our speech segregation system. The results in Table II are given for both the reverberant case (Reverb) and inverse-filtered case (Inverse) at three SNR levels: -5 dB, 0 dB and 5 dB. The results are obtained using the estimated pitch tracks provided by the multi-pitch tracking algorithm of Wu et al. (2003) as explained in Section 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 only 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 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 by attempting to over-group 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 chosen pitch estimation method is accurate.

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°.

Input SNR	**-5 dB**		**0 dB**		**5 dB**	
	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°.

Input SNR	**-5 dB**		**0 dB**		**5 dB**	
	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**

As seen in the results presented above, the major advantage of the inverse-filtering stage occurs for 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 shows SNR results for both white noise as well as 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 obtained using inverse filtering compared to the reverberant condition is largely attenuated 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.

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

Input SNR	**-5 dB**		**0 dB**		**5 dB**	
	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°.

Input SNR	**-5 dB**		**0 dB**		**5 dB**	
	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**

As mentioned in Section I, our system is the 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 non-stationary 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 with relatively steady intrusion, the spectral subtraction algorithm tends to subtract little intrusion but it introduces little distortion to the target signal. By comparison, our system is a target-centered algorithm that attempts to reconstruct the target signal on the basis of periodicity. Target components made inharmonic by reverberation are therefore removed by our algorithm, thus introducing more distortion to the target signal. It is worth noting that the ceiling performance of our algorithm without any interference is 8.89 dB.

V. DISCUSSION

In natural settings, reverberation alters many of the acoustical properties of a sound source reaching our ears, including smearing out 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 normal 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 a 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 for a two-talker situation with T_{60} = 0.35 s. Observe that this decrement is consistent with the 5 dB increase in speech reception threshold reported by Culling et al. (2003).

To reduce the smearing effects on the target speech, we have proposed a pre-processing stage which equalizes the room impulse response that corresponds to target location. This pre-processing results in both improved harmonicity for signals arriving from the target direction as well as smearing of competing sources at other locations, and thus provides a better input signal for the pitch-based segregation system. The extensive evaluations show that our system yields substantial SNR gains across a variety of noise conditions.

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 a practical 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 proposed by Gillespie and Atlas (2002). This type of pre-processing 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, walking) will have an adverse impact on the inverse-filtering stage. Second, the stage needs to identify signal intervals that contain no interference to allow for the filter adaptation**.** On the other hand, our pitch-based segregation stage can function without training and is robust to a variety of environmental changes. Hence, whenever the adaptation of the inverse filter is infeasible, one can use 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.

ACKNOWLEDGMENTS

This research was supported in part by an AFOSR grant (FA9550-04-1-0117), an NSF grant (IIS-0081058) and an AFRL grant (FA8750-04-1-0093).

References

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

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

R. Balan, A. Jourjine and J. Rosca (**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.

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

M. Brandstein and D. Ward, Eds. (**2001**). *Microphone Arrays: Signal Processing Techniques and Application*, Berlin: Springer.

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

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

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

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

D. Brungart, P. Chang, B. Simpson and D. L. Wang (**2005**). "Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask," submitted.

M. D. Burkhard and R. M. Sachs (**1975**). "Anthropometric manikin for acoustic research," J. Acoust. Soc. Am., vol. 58, pp. 214-222.

M. P. Cooke (**1993**). *Modeling Auditory Processing and Organization*, Cambridge University Press, Cambridge U. K.

M. P. Cooke, P. Green, L. Josifovski and A. Vizinho (**2001**). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Comm., vol. 34, pp. 267-285.

J. F. Culling, K. I. Hodder and C. Y. Toh (**2003**). "Effects of reverberation on perceptual segregation of competing voices," J. Acoust. Soc. Am., vol. 114, pp. 2871-2876.

C. J. Darwin and R. P. Carlyon (**1995**). "Auditory grouping," in *The handbook of perception and cognition*, vol. 6, B. C. J. Moore, Ed., London:Academic, pp. 387-424.

C. J. Darwin and R. W. Hukin (**2000**). "Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention," J. Acoust. Soc. Am., vol. 108, pp. 335-342.

Y. Ephraim and H. L. Trees (**1995**). "A signal subspace approach for speech enhancement," IEEE Trans. Speech Audio Proc., vol. 3, pp. 251-266.

K. Furuya and Y. Kaneda (**1997**). "Two-channel blind deconvolution for non-minimum phase impulse responses," Proc. ICASSP, pp. 1315-1318.

W. G. Gardner and K. D. Martin (**1994**). "HRTF measurements of a KEMAR dummy-head microphone," MIT Media Lab Perceptual Computing Technical Report #280.

B. W. Gillespie and L. E. Atlas (**1994**). "Acoustic diversity for improved speech recognition in reverberant environments," Proc. ICASSP, pp. 557-560.

B. W. Gillespie, H. S. Malvar and D. A. F. Florencio (**2001**). "Speech dereverberation via maximum-kurtosis subband adaptive filtering," Proc. ICASSP, vol. 6, pp. 3701-3704.

S. Haykin (**2002**). *Adaptive Filter Theory*, 4th ed., Upper Saddle River, New Jersey: Prentice Hall.

Helmholtz, H. (**1863**). *On the Sensation of Tone* (A. J. Ellis, Trans.), 2nd English ed., New York: Dover Publishers.

G. Hu and D. L. Wang (**2003**). "Separation of stop consonants," Proc. ICASSP, vol. 2, pp. 749-752.

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

G. Hu and D. L. Wang (**2005**). "Separation of fricatives and affricates," Proc. ICASSP, vol. 1, pp. 1101-1104.

G.-J. Jang, T.-W. Lee, Y.-H. Oh (**2003**). "Single channel signal separation using time-domain basis functions" IEEE Signal Proc. Letters, vol. 10, nr. 6, pp. 168-171.

J. C. R. Licklider (**1951**). "A duplex theory of pitch perception," Experientia, vol. 7, pp. 128-134.

H. Y. Luo and P. N. Denbigh (**1994**). "A speech separation system that is robust to reverberation," Proc. ISSIPNN, pp. 339-342.

N. Ma, M. Bouchard and R. Goubran (**2004**). "Perceptual Kalman filtering for speech enhancement in colored noise", Proc. ICASSP, vol. 1, pp.717-720.

R. Martin (**2001**). "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Proc., vol. 9, pp. 504-512.

M. Miyoshi and Y. Kaneda (**1988**). "Inverse filtering of room impulse response," IEEE Trans. Acoust., Speech, Signal Proc., vol. 36, pp. 145-152.

B. C. J. Moore (**2003**). *An introduction to the Psychology of Hearing*, 5th ed., San Diego, CA: Academic.

A. K. Nabelek and P. K. Robinson (**1982**). "Monaural and binaural speech perception in reverberation for listeners of various ages," J. Acoust. Soc. Am., vol. 71, pp. 1242-1248.

T. Nakatani and M. Miyoshi (**2003**). "Blind dereverberation of single channel speech signal based on harmonic structure," Proc. ICASSP, pp. 92-95.

T. Nakatani and H. G. Okuno (**1999**). "Harmonic sound stream segregation using localization and its application to speech stream segregation," Speech Comm., vol. 27, pp. 209-222.

O' Shaughnessy, D. (**2000**). *Speech Communications: Human and Machine*, 2nd ed., Piscataway, NJ: IEEE Press.

A. V. Oppenheim and R. W. Schafer (**1989**). *Discrete-time signal processing*, Englewood Cliffs, NJ: Prentice-Hall.

K. J. Palomaki, G. J. Brown and D. L. Wang (**2004**). "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," Speech Comm., vol. 43, pp. 361-378.

R. D. Patterson, I. Nimmo-Smith, J. Holdsworth and P. Price (**1988**). "APU Report 2341: An efficient auditory filterbank based on the gammatone function," Applied Psychology Unit, Cambridge.

R. Plomp (**1976**). "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of a single competing sound source (speech or noise)," Acustica, vol. 34, pp. 200-211.

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

J. Rouat, Y. C. Liu and D. Morissette (**1997**). "A pitch determination and voice/unvoiced decision algorithm for noisy speech," Speech Comm., vol. 21, pp. 191-207.

A. Shamsoddini and P. N. Denbigh (**2001**). "A sound segregation algorithm for reverberant conditions," Speech Comm., vol. 33, pp. 179-196.

M. Slaney and R. F. Lyon (**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. New York: Wiley, pp. 95-116.

S. Srinivasan, N. Roman and D.L. Wang (**2004**). " On binary and ratio time-frequency masks for robust speech recognition," Proc. ICSLP, pp. 2541-2544.

D. L. Wang (**2004**). "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 Academic, pp. 181-197.

D. L. Wang and G. J. Brown (**1999**). "Separation of speech from interfering sounds based on oscillatory correlation," IEEE Trans. Neural Net., vol. 10, pp. 684-697.

M. Weintraub (**1985**). "A theory and computational model of auditory monaural sound separation," Ph.D. Dissertation, Stanford University Department of Electrical Engineering.

M. Wu (**2003**). "Pitch tracking and speech enhancement in noisy and reverberant environments," PhD thesis, The Ohio State University.

M. Wu and D. L. Wang (**2005**). "A two-stage algorithm for one-microphone reverberant speech enhancement," IEEE Trans. Speech, Audio Proc., in press.

M. Wu, D.L. Wang and G. J. Brown (**2003**). "A multipitch tracking algorithm for noisy speech," IEEE Trans. Speech, Audio Proc., vol. 11, pp. 229-241.

M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev (**2001**). "Blind source separation by sparse decomposition", in *Independent Component Analysis: Principles and Practice*, S. J. Roberts, and R.M. Everson, Eds., Cambridge.