A Pitch-based Model for Separation of Reverberant Speech

Nicoleta Roman and DeLiang Wang

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

10.21437/Interspeech.2005-687

Abstract

In everyday listening, both background noise and reverberation degrade the speech signal. While monaural speech separation based on periodicity has achieved considerable progress in handling additive noise, little research has been devoted to reverberant scenarios. Reverberation smears the harmonic structure of speech signals, and our evaluations using a pitch-based separation algorithm show that an increase in the room reverberation time causes degradation in performance due to the loss in periodicity for the target signal. We propose a two-stage monaural speech 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 processing 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 separation. A systematic evaluation shows that the proposed system results in considerable signal-to-noise ratio gains across different conditions.

1. Introduction

In a natural environment, a desired speech signal often occurs simultaneously with other interfering sounds such as echoes and background noise. While the auditory system excels at separating speech from such complex mixtures, simulating this perceptual ability computationally remains a great challenge. Our monaural study is motivated by the following two considerations. First, a one-microphone solution to sound separation is highly desirable in many applications including automatic speech recognition and hearing aids. Second, although binaural listening improves speech intelligibility in anechoic conditions, this binaural advantage is largely eliminated by reverberation [1] which emphasizes the dominant role of monaural hearing in realistic conditions.

According to Bregman, the auditory system employs various cues including fundamental frequency (F0), onset time and location in a process known as auditory scene analysis (ASA) [2]. This theory has inspired a series of computational ASA (CASA) systems (see [3] for a review). At the core of these systems is a time-frequency (T-F) mask which selectively weights the acoustic mixture in order to enhance the desired signal. An ideal binary mask has been proposed as the computational goal for CASA [4]. 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 [5] [6].

Monaural separation of voiced speech has been studied previously by exploiting primarily pitch information (see e. g., [7] [8]). In this paper, we propose a pitch-based speech segregation method that follows the same principles as the system in [8] while simplifying the calculations required for extracting periodicities. The system shows good performance when tested with a variety of noise intrusions in anechoic conditions. However, when pitch varies with time in a reverberant environment, reflected waves with different F0s arrive simultaneously at the ear. This multipath situation causes smearing of the signal in the sense that harmonic structure is less clear in the signal [1]. Due to this 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 room impulse response is a minimum-phase filter. Several strategies have been proposed to estimate the inverse filter in unknown acoustical conditions [9] [10] [11]. In particular, the system developed by Gillespie et al. estimates the inverse filter from an array of microphones using an adaptive gradient-descent algorithm that maximizes the kurtosis of linear prediction (LP) residuals [10]. 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 in [12].

In this paper, 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. We show that the inverse filtering stage improves the separation performance of the proposed pitch-based system using signal-to-noise ratio (SNR) evaluations. To our knowledge, the proposed system is the first study that addresses monaural speech segregation with room reverberation.

The paper is organized as follows. Section 2 gives a detailed presentation of the model. Section 3 gives systematic results on pitch-based segregation both in the reverberant and the inverse-filtered condition. Section 4 concludes the paper.

2. Model Architecture

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

$$y(t) = h(t) * s(t) + n(t), \quad (1)$$

where '$*$' indicates convolution. $s(t)$ is the clean target signal, $h(t)$ is the room impulse response from the target location to the ear, and $n(t)$ is background noise. We propose a two-stage model for speech segregation: 1) inverse filtering with respect to target location in order to enhance the periodicity of target signal; 2) pitch-based speech segregation. The details are presented in the following two subsections.

2.1. Target inverse filtering

As described in the introduction, inverse filtering is a standard method for dereverberating the target. We employ the inverse filtering algorithm implemented in [12] which attempts to blindly estimate the inverse filter from one-microphone 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 $\tilde{z}(t)$:

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

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 $\mathbf{g}$ is an inverse filter of length L. The inverse filter is derived by maximizing the kurtosis of $\tilde{z}(t)$.

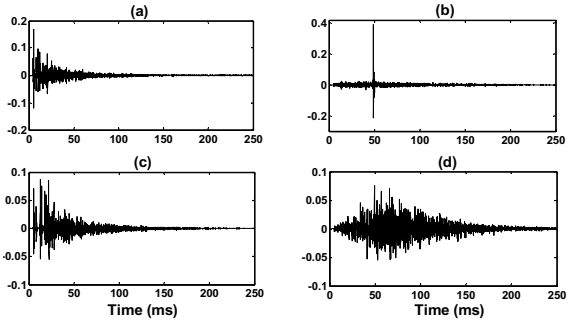

Figure 1. *(a) A room impulse response for a target source simulated in the median plane. (b) The effect of convolving the impulse response in (a) with the derived inverse filter. (c) A room impulse response for one interfering source at 45° azimuth. (d) The effect of convolving the impulse response in (c) with the derived inverse filter.*

The system is trained in the absence of interference on reverberant speech from the target source sampled at 16 kHz. 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 1 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 T_{60} room reverberation time is 0.35 s (T_{60} is the time required for the sound level to drop by 60 dB following the sound offset). As can be seen in Fig. 1(b), the equalized response for the target source is far more impulse-like compared to the room impulse response in Fig. 1(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. 1(d).

2.2. Monaural speech segregation

Monaural CASA systems generally perform speech segregation using periodicity information following two stages: 1) segmentation and 2) grouping. In particular, a recent system proposed by Hu and Wang shows good performance across a variety of additive noise conditions [8]. In the low-frequency range, the system generates segments based on temporal continuity and cross-channel correlation, and groups them according to their periodicity. In the high frequency range, the signal envelope fluctuates with the pitch rate and the system makes use of amplitude modulation (AM). The segregation mechanism produces a binary mask which selects time-frequency (T-F) units where the target signal dominates the interference.

In this paper, we propose a pitch-based segregation method that follows the same principles as the Hu and Wang model while simplifying the calculations required for extracting periodicities. The proposed system incorporates a correlogram-based labeling strategy which while being simpler produces a more robust feature in the high frequency range compared to the AM rate computation in [8]. For pitch extraction, we employ a multi-pitch tracking algorithm which generates up to two pitch contours [13]. The system needs assignment of overlapping pitch contours in the case of a harmonic interference. For this, we use as ground truth the 'ideal' pitch contour extracted by Praat from the target signal.

The signal provided by the inverse filtering stage is filtered through a bank of 128 fourth-order gammatone filters. Envelopes are extracted in the high frequency channels as follows. 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 [13]. Finally, correlograms are computed using the normalized autocorrelation function in each frequency channel at 10 ms intervals using time windows of 20 ms.

The labeling of a T-F unit is carried out by comparing the estimated pitch lag p with the periodicity of the correlogram, $A(n)$. In the low-frequency range, the system selects the time lag l that corresponds to the closest peak in $A(n)$ from the pitch lag. For a particular channel, the distribution of 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 $|p - l|$ between the two lags exceeds a threshold θ_L. We have found empirically that a value of $\theta_L = 0.15 * (F_s / F)$ results in good performance, where F_s is the sampling frequency and F is the center frequency of the channel. Finally, the unit is labeled 1 if $A(l)$ is close to the maximum of $A(n)$ in the plausible pitch range:

$$\frac{A(l)}{\max_{n \in [32,200]} A(n)} > \theta_P, \quad (3)$$

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

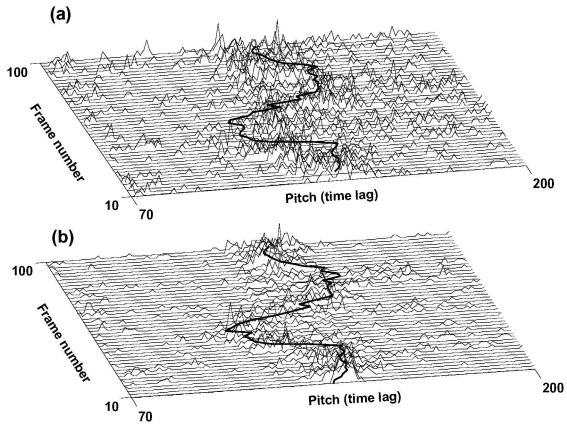

Figure 2. *Histograms of selected peaks in the high-frequency range (>800 Hz) for a male utterance. (a) Results for the reverberant target signal. (b) Results for the inverse filtered target signal.*

In the high-frequency range, we adapt the peak selection mechanism developed in [13]. First, the envelope-based correlogam $A_E(n)$ of periodic signals exhibits a peak both at the pitch lag and at the double of the pitch lag. Thus, the system selects first all the lags such that a peak exists at that lag and another peak exists within a 5 percent tolerance from the double of that lag. If no peaks are selected, the T-F unit is labeled 0. Second, a harmonic introduces peaks at lags around the multiple of its pitch lag. Therefore, our system selects the first peak that is higher than half of the maximum peak in $A_E(n)$ for $n \in [32, 200]$. 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 the 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 the test set.

The distortions on harmonic structure due to room reverberation are generally more salient in the high-frequency range. Figure 2 illustrates the effect of reverberation as well as inverse filtering in frequency channels above 800 Hz for a single male utterance. 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. Moreover, the channel selection mechanism retains 79 percent of the total signal energy by applying inverse-filtering as compared to 58 percent without inverse filtering.

The final segregation of the acoustic mixture is based on combined segmentation and grouping. The motivation is to improve on the T-F unit labeling using segment-level features. Here, we combine the labeling described above with the segmentation framework proposed in [8]. 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, segregated target speech is resynthesized from the resulting T-F binary mask.

3. Results

We have evaluated the system on the left-ear response of a KEMAR dummy head, simulated using the room acoustic model described in [14] for a small rectangular room (6m×4m×3m). The position for the KEMAR is fixed in a corner at (2.5m×2.5m×2m) while the sources are located at 1.5 m from the receiver. In all cases, the target is fixed at 0° and the interference is at 45°, unless otherwise specified. The inverse filter of the target room impulse response is estimated from the training data as explained in Section 2.1 and applied on the whole reverberant mixture. We asses the performance of the pitch-based segregation system in two conditions: 1) reconstructing the reverberant target from the reverberant mixture and 2) reconstructing the inverse-filtered target from the filtered mixture.

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:

$$SNR = 10 \log 10 \left(\sum_t s_{IBM}^2(t) \Big/ \sum_t (s_{IBM}(t) - s_E(t))^2 \right), \qquad (4)$$

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 SNR evaluations using as target the set of 10 voiced male sentences collected by Cooke for the purpose of evaluating voiced speech segregation systems (see also [8]). The following 5 noise intrusions are used: white noise, babble noise, music, a female utterance and a male utterance. These intrusions represent typical acoustical interferences occurring in real environments. Each value in the following tables represents the average output SNR of one particular intrusion mixed with the 10 target sentences.

Table 1 shows the performance of our pitch-based speech segregation system applied directly on reverberant mixtures when the reverberation time increases from 0.05 s to 0.35 s. The mixtures are obtained by mixing the reverberant target speech with a female speech utterance at 3 SNR levels: -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 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.

Now we analyze how inverse-filtering pre-processing impacts the overall performance of our speech segregation system. The results in Table 2 are given for both the reverberant case (R) and inverse-filtered case (I-F) at three SNR levels: -5 dB, 0 dB and 5 dB. 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 (I-F) 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.

Table 1. *Output SNR results for 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

Table 2. *Output SNR results for target mixed with different noise types at T_{60} = 0.35 s. Target at 0° and noise at 45°.*

Input SNR	-5 dB		0 dB		5 dB	
	R	I-F	R	I-F	R	I-F
Female	3.21	4.90	4.61	6.42	5.70	7.71
Male	0.92	2.71	3.50	4.89	5.27	6.93
White noise	4.07	4.32	5.17	5.5	5.8	6.99
Babble noise	0.85	1.26	2.97	3.78	4.78	6.00
Rock Music	2.76	3.64	4.49	5.61	5.64	7.10
Average	**2.36**	**3.36**	**4.14**	**5.24**	**5.43**	**6.94**

Table 3. *Output SNR results for target mixed with different noise types at T_{60} = 0.35 s. Target and noise at 0°.*

Input SNR	-5 dB		0 dB		5 dB	
	R	I-F	R	I-F	R	I-F
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 3 shows results when the interference location is fixed at 0°, the same as the target location. As expected, in the white noise case, the SNR gains are similar to the ones presented in Table 2. However, for the female speech interference, the relative improvement obtained using inverse filtering 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 segregation performance.

Finally, we compare our system with spectral subtraction which is a standard speech enhancement technique. For this, the SNR is computed using the reverberant target as ground truth for the spectral subtraction and the inverse-filtered target as ground truth for our system. Spectral subtraction performs significantly worse than our system, especially at low levels of input SNR because of its well known deficiency in dealing with non-stationary interferences. For example, at -5 dB input SNR, the average output SNR is -1.81 dB compared to the 4.01 dB produced by our system. At high input SNR, however, spectral subtraction although does not remove all the noise it introduces little distortion to the target signal. By comparison, our system does not retain the inharmonic target components. Hence, spectral subtraction performs slightly better than our system in those cases.

4. Conclusion

We have investigated pitch-based monaural segregation in room reverberation and report the first systematic results on this challenging problem. Reverberation causes huge problems for pitch-based segregation systems due to the smearing of harmonic structure. 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. Extensive evaluations show that our system yields substantial SNR gains across a variety of noise conditions. According to ASA, auditory cues such as onsets, acoustic-phonetic properties of speech are also important for monaural separation. Future work will therefore attempt to utilize these cues to enhance the current performance.

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

5. References

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

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

[3] G. J. Brown and D. L. Wang, "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, 2005.

[4] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, ed., Norwell MA: Kluwer Academic, pp. 181-197, 2004.

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

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

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

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

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

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

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

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

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

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