A SUPERVISED LEARNING APPROACH TO MONAURAL SEGREGATION OF REVERBERANT SPEECH

Zhaozhang Jin and *DeLiang Wang*

Department of Computer Science and Engineering
& Center for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
{jinzh, dwang}@cse.ohio-state.edu

ABSTRACT

Room reverberation degrades speech signals and poses a major challenge to current monaural speech segregation systems. Previous research relies on inverse filtering as a front-end for partially restoring the harmonicity of the reverberant signal. We show that the inverse filtering approach is sensitive to different room configurations, hence undesirable in general reverberation conditions. We propose a supervised learning approach to map a set of harmonic features into a pitch based grouping cue for each time-frequency (T-F) unit. We use a speech segregation method to estimate an ideal binary T-F mask which retains the reverberant mixture in a local T-F unit if and only if the energy of target is stronger than interference energy. Results show that our approach improves the segregation performance considerably.

Index Terms— Speech segregation, room reverberation, supervised learning, computational auditory scene analysis.

1. INTRODUCTION

In real-world acoustic environments, the speech signals reaching our ears are often corrupted by noise and room reverberation. Many techniques have been proposed to achieve monaural speech segregation or enhancement because one-microphone solutions are highly desirable in various applications (see [1]). Existing studies, however, are largely limited to anechoic conditions and little research has been developed to tackle the monaural segregation problem in reverberant environments. In this paper, we study monaural segregation of reverberant speech.

Inspired by the human auditory perception [2], computational auditory scene analysis (CASA) aims to achieve good performance in speech segregation based on perceptual principles [3]. From the perspective of speech segregation, the notion of an ideal binary time-frequency (T-F) mask has been proposed as the computational goal of 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 interference and 0 otherwise. Previous research has shown that speech reconstructed from the ideal binary mask is highly intelligible and produces large improvements in robust speech recognition [4].

Segmentation and grouping are the two main stages of CASA systems [5]. In segmentation, the input is decomposed into T-F segments, each of which is deemed to originate from a single source; in grouping, those segments that come from the same source are grouped into a stream by using harmonicity and other grouping cues. However, under reverberant conditions, the effectiveness of grouping by harmonicity is degraded because reverberation causes reflections of each harmonic to combine additively with the direct sound. Due to weakened harmonicity in the speech signal, the performance of most monaural CASA systems is expected to suffer under reverberant conditions.

To restore speech harmonicity, one method is to estimate and apply an inverse filter of the room impulse response corresponding to the target source [6]. Although inverse filtering partially counteracts the smearing effect of reverberation on speech spectrum, it assumes that a room configuration, e.g., room dimensions, wall reflection, source and microphone locations, etc., be stationary. Even if the source moves within a few centimeters range, the inverse filter needs to be re-estimated [7]. To quantify such adverse effects, we systematically evaluate the sensitivity of inverse filtering to a number of room configurations with different source and microphone locations and different reverberation times (T_{60}).

Here, we propose a supervised learning approach to achieve robustness against reverberation effects. Specifically, we estimate within each T-F unit a pitch based cue from a set of harmonic-related features extracted from a reverberant signal for grouping. A multilayer perceptron (MLP) is trained for each channel of a gammatone filterbank. The estimated grouping cues are utilized in the grouping stage to accomplish segregation of reverberant speech.

The rest of the paper is organized as follows. The next section evaluates the sensitivity of the inverse filtering approach. Section 3 presents the proposed supervised learning approach in detail. Our system is evaluated and results are given in Section 4. Conclusions are drawn in Section 5.

2. SENSITIVITY OF INVERSE FILTERING

The room impulse response characterizes reverberation of a specific room configuration (e.g., room size, reflection coefficient, and target/microphone location). A slight variation in the configuration could cause a big difference in the room impulse response and its inverse filter accordingly. In other words, if an inverse filter is estimated from the same room impulse response as what it applies to (i.e., matched inverse filtering), it enhances speech harmonicity; otherwise it further smears the harmonic structure. Fig. 1 demonstrates the effects of applying the same inverse filter to the matched room impulse response and a different room impulse response resulting from a moved source location. Both of the room reverberation times are $T_{60} = 0.3$ s. As can be seen in Fig. 1(b), the equalized response is much impulse-like, indicating the success of reverberation atten-

uation, while the mismatched room impulse response gets further smeared in Fig. 1(d) when convolving with the same inverse filter.

We also quantitatively evaluate the sensitivity of inverse filtering to different room configurations. Signal-to-reverberant energy ratio (SRR) is essentially a measure of intelligibility of reverberant speech [8] and hence a good indicator of the effectiveness of inverse filtering. SRR is defined as:

$$SRR = 10\log_{10}\left(\int_0^{t_1} p^2(t)dt \Big/ \int_{t_1}^{\infty} p^2(t)dt\right). \quad (1)$$

Here, $p(t)$ is the instantaneous sound pressure of the room impulse response measured at time t and t_1 is the arrival time of the first peak from the reflected impulses. A larger SRR value indicates higher intelligibility. Table 1 shows the SRR improvement after applying the inverse filter to room impulse responses for six different rooms. Each room has three sets of random (source1, source2, microphone) locations corresponding to six room impulse responses (1a, 1b, 2a, 2b, 3a, 3b), where "a" refers to the impulse response from source1 to the microphone and "b" from source2 (see Fig. 2 in Section 4). The inverse filter is estimated from the reverberant speech generated by the room impulse response (1a) in the room with $T_{60} = 0.3$ s using [9]. To examine the sensitivity of inverse filtering, this estimated inverse filter is used to convolve with all 36 room impulse responses and SRR's are calculated accordingly. It is evident, in Table 1, that significant SRR improvement only occurs under the matched inverse filtering condition. The SRR drops for almost all the other cases, implying a further smearing effect caused by mismatched inverse filtering. In conclusion, the inverse filtering approach is sensitive to different rooms and different source and microphone locations. Such a limitation hinders the applicability of this approach.

Table 1. Signal-to-reverberant ratio (SRR) improvement (in dB) by applying the estimated inverse filtering to each room impulse function. Dimensions (in meter) and reverberation time of each room are listed in the first column.

Rm Dim. ($T_{60}(s)$)	**1a**	**1b**	**2a**	**2b**	**3a**	**3b**
4×4×3 (0.1)	-7.8	-6.8	-7.4	-6.0	-7.5	-8.4
5×4×3 (0.2)	-3.7	-2.8	-3.8	-4.5	-4.4	-2.7
6×4×3 (0.3)	7.2	-2.1	-0.2	-2.9	-2.8	-2.3
7×5×3 (0.4)	-1.5	-1.1	-0.3	-2.7	-0.8	-0.5
8×5×3 (0.5)	-2.1	-0.4	-2.2	-1.4	-1.4	-1.4
9×5×3 (0.6)	0.0	1.2	-0.3	0.2	1.3	-0.1

3. SYSTEM DESCRIPTION

The signal received at a microphone, $y(t)$, in a reverberant enclosure undergoes both convolutive and additive distortions:

$$y(t) = h_T(t) * s(t) + h_I(t) * n(t), \quad (2)$$

where "$*$" indicates convolution. $s(t)$ is the clean (or anechoic) target speech and $h_T(t)$ models the room impulse response from the target speaker to the microphone, while $n(t)$ is the anechoic interference and $h_I(t)$ models the room impulse response from the interference to the microphone. Given a one-microphone recording with the above setup, the goal of our system is to segregate the reverberant target out of the mixture. This is accomplished by retaining time-frequency regions where a reverberant target speech is stronger than reverberant interference and discarding those regions where the reverberant interference is stronger. A monaural segregation system is thus proposed for this purpose. It estimates grouping cues from pitch-based acoustic features using trained MLP networks. Such cues are then applied in the grouping stage of a monaural CASA system to produce a binary mask that segregates the reverberant target.

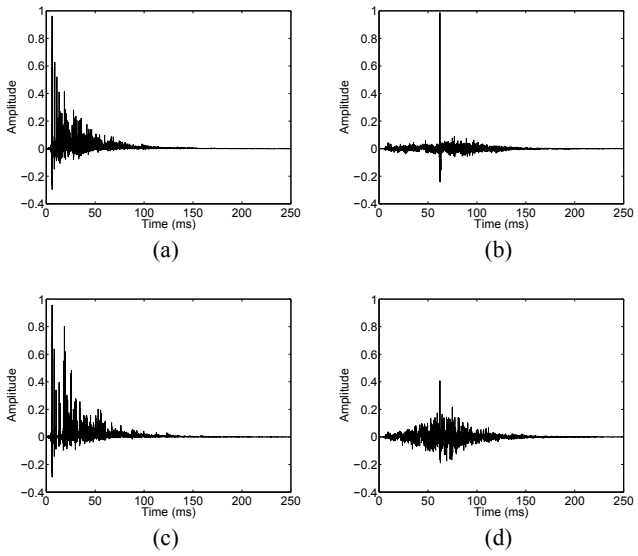

Fig. 1. Effects of inverse filtering on room impulse responses. (a) A room impulse response function generated by the image model in an office-size room of the dimensions 6 by 4 by 3 meters with reverberation time $T_{60} = 0.3$ s. Reflection coefficients are 0.73 for all the walls, the ceiling, and the floor. The source and the microphone are at (4, 0.9, 1) and (2, 1, 1), respectively. (b) The result of convolving the impulse response in (a) with the estimated inverse filter. (c) A different room impulse response function in the same room but with the source location at (0.6, 2.4, 1). (d) The result of convolving the impulse response in (c) with the estimated inverse filter.

3.1. Learning Grouping Cues

The idea of supervised learning is to estimate a grouping cue which is the posterior probability of a T-F unit being target dominant given a set of harmonic related acoustic features $\mathbf{x}_{c,m}$ of time frame m and channel c. Thus, we define the grouping cue $C_g(c,m)$ as:

$$C_g(c,m) = P(H_0|\mathbf{x}_{c,m}). \quad (3)$$

H_0 is the hypothesis that a T-F unit is target dominant and H_1 otherwise. This probability plays a crucial role in unit labeling in our segregation system (see Section 3.2).

To extract acoustic features, the input mixture is first analyzed using a 128-channel gammatone filterbank [10] whose center frequencies are quasi-logarithmically spaced from 50 Hz to 8 kHz [11]. In addition, envelopes are extracted for channels with center frequencies higher than 800 Hz using a Teager energy operator followed by a third-order Butterworth filter with cutoff frequency of 800 Hz. The

outputs of all the channels are further high-pass filtered to 64 Hz to remove the distortions due to very low frequencies. At a given time step m, the correlogram $A(c, m, \tau)$ for channel c with a time lag τ is computed using a window of 20 ms in every 10 ms interval. The range for time delay τ is from 32 to 200 corresponding to the plausible pitch range of 80 to 500 Hz. Following [12], we then construct the feature vector as:

$$\mathbf{x}_{c,m} = \{A(c,m,\tau_m), \text{int}(\bar{f}(c,m)\tau_m), \bar{f}(c,m)\tau_m - \text{int}(\bar{f}(c,m)\tau_m) \\ A_E(c,m,\tau_m), \text{int}(\bar{f}_E(c,m)\tau_m), \bar{f}_E(c,m)\tau_m - \text{int}(\bar{f}_E(c,m)\tau_m)\}, \quad (4)$$

where τ_m is the pitch period for time frame m, $A(c, m, *)$ and $A_E(c, m, *)$ are autocorrelation functions, $\bar{f}(c, m)$ and $\bar{f}_E(c, m)$ are the estimated average instantaneous frequencies corresponding to channel c of time frame m. The first three features are based on the gammatone filterbank responses while the last three based on the envelopes of the responses (described earlier in the section). Essentially, the first and the fourth features capture the periodicity for each T-F unit; the second and the fifth features give the number of harmonics; the third and the sixth features represent the distance between the current pitch and the nearest harmonic.

We use a MLP to transform $\mathbf{x}_{c,m}$ into $C_g(c, m)$. At each frame for each channel, the input to the MLP is the feature vector extracted in (4). The desired output is set to be 1 if the target energy is stronger than interference energy within a T-F unit and 0 otherwise. The prior knowledge of the energies are obtained from the premixing target and interference signals. We train 128 MLP's for 128 channels, each having the same network structure of 20 nodes in the hidden layer. The number of nodes is justified using ten-fold cross-validation. The transfer functions of the hidden and output layers are hyperbolic tangent sigmoid and linear, respectively. Each MLP is trained using Levenberg-Marquardt backpropagation [13] for 100 epoches.

3.2. Monaural Speech Segregation

The proposed monaural segregation algorithm follows the framework of segmentation and grouping in [11]. Due to smeared harmonic structures in reverberant speech, we found that both the periodicity criterion in the low-frequency range and the amplitude modulation criterion in the high-frequency range [11] are no longer reliable for grouping. In this condition, our proposed grouping cue as discussed in Section 3.1 is more robust and therefore a criterion based on it can be a reasonable substitute. Consequently, we define a new criterion for unit labeling: A T-F unit u_{cm} is labeled as target speech if the posterior probability of it being target dominant ($P(H_0|\mathbf{x}_{c,m})$) is greater than the probability of interference dominant ($P(H_1|\mathbf{x}_{c,m})$). According to (3) and note that $P(H_0|\mathbf{x}_{c,m}) + P(H_1|\mathbf{x}_{c,m}) = 1$, this criterion can be written as

$$C_g(c, m) > 0.5. \quad (5)$$

Because the grouping cue is derived from the feature vector that captures both resolved and unresolved harmonics, the above criterion should work well for unit labeling in both low- and high-frequencies.

In the proposed algorithm, the above criterion is first used to label the units of the segments generated by the initial segmentation based on temporal continuity and cross-channel correlation. Segments are then grouped into target or background streams according to this labeling. Secondly, new target segments are formed by iteratively merging T-F units that are labeled as target dormant but not yet grouped into the target stream in the previous grouping. Only those segments longer than 50 ms are grouped into the target stream. Finally, the target stream is further refined by iteratively absorbing the neighboring units that do not belong to either stream but are labeled as target dominant according to (5).

After the grouping stage, all T-F units belonging to the target stream are labeled 1 and the other units labeled 0. Finally, a binary mask is formed and the segregated target speech is resynthesized from this estimated mask for evaluation.

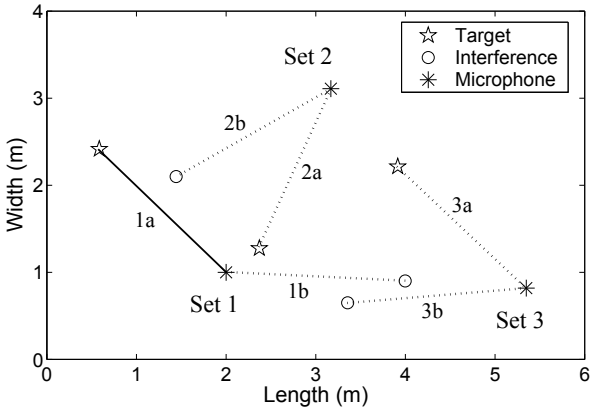

Fig. 2. Room configurations with three sets of (target, interference, microphone) locations randomly placed in the enclosure. Lines indicate direct transmission paths from sources to the microphone within each set.

4. EXPERIMENTAL RESULTS

In order to systematically evaluate the proposed system, we simulate six rooms with different dimensions and reverberation times ranging from 0.1 s to 0.6 s. The first column of Table 1 shows the dimensions and the reverberation time for each room. Then, as already mentioned in Section 2, three sets of (source1, source2, microphone) locations are randomly generated in each room, source1 being target and source2 interference. Fig. 2 depicts an example of configurations in room 3 with $T_{60} = 0.3$ s. Each (source, microphone) pair is characterized by a room impulse response calculated using the image model [14]. Corpus is built by mixing a set of ten voiced male utterances used in [15] as target and five different types of interference (including white noise, babble noise, rock music, a male utterance and a female utterance). In particular, mixtures are created according to (2), in which $h_T(t)$ and $h_I(t)$ are room impulse responses of (target, microphone) and (interference, microphone) respectively. The signal-to-noise ratio (SNR) is set to 0 dB for each mixture.

Given that the computational objective of our proposed system is to identify T-F regions that are target dominant, we adopt the same SNR measure in [11] using the resynthesized speech from the ideal binary mask as the ground truth

$$SNR = 10\log_{10}\left(\sum_t s_I^2(t) \Big/ \sum_t \left(s_I(t) - s_E(t)\right)^2\right), \quad (6)$$

where $s_I(t)$ and $s_E(t)$ are signals resynthesized from the ideal binary mask and the mask estimated by our proposed system, respectively.

We compare the performance of our proposed system to that of Roman and Wang [6]. In their system, an inverse filter is first esti-

Table 2. Comparisons of SNR gain (in dB) between the proposed system and the Roman-Wang system.

	Roman-Wang						Proposed					
$T_{60}(s)$	0.1	0.2	0.3	0.4	0.5	0.6	0.1	0.2	0.3	0.4	0.5	0.6
Set 1	8.28	8.05	7.99	7.39	7.83	5.70	8.83	9.63	9.26	8.63	9.57	7.24
Set 2	9.09	8.10	7.97	7.85	6.53	6.92	9.48	9.58	9.15	8.39	7.76	7.98
Set 3	8.28	6.56	7.13	7.82	8.09	5.97	9.26	7.65	8.82	9.32	9.25	7.44
Average	8.55	7.57	7.70	7.69	7.49	6.19	9.19	8.95	9.08	8.78	8.86	7.55

mated by maximizing the kurtosis of the inverse-filtered LP residual of the reverberant speech, which is generated by the room impulse response between the target and the microphone from set 1 in room 3 (shown as the solid line in Fig. 2). The obtained inverse filter is then applied to mixtures from different configurations. Similarly, in our proposed system, we only use mixtures generated from set 1 in room 3 for the training purpose. To remove the influence of pitch estimation errors to the segregation systems in comparison, we extract *a priori* pitch contours from premixing reverberant targets for both systems.

Segregation performance of the proposed system in terms of SNR gain is summarized in Table 2. Each column in the table shows the average SNR gains across all test utterances in three different sets of (target, interference, microphone) configurations in a room. The last row shows the average SNR gains over different configurations in each room. As observed in Table 2, the proposed system achieves considerably higher SNR gains across all different room reverberation times compared to the Roman-Wang system. This improvement largely comes from the unit labeling in high frequencies because previously there was no reliable way to handle unresolved harmonics in high frequency channels for reverberant signal. Through training, both resolved and unresolved harmonics can be captured and units in both low- and high-frequencies can be reliably labeled.

For different room configurations, there is also a clear trend of SNR drop as the room reverberation time increases. This indicates that the level of reverberation is still an important factor that decides the performance. The sensitivity of a system to different reverberation conditions can be measured by the standard deviation of the averaged SNR gains across all T_{60}'s. The proposed system has a standard deviation of 0.59, which is lower than that of the Roman-Wang system, which is 0.76. This confirms the conclusion that our system generalizes better to different T_{60}'s.

5. CONCLUSIONS

In this paper, we have proposed a supervised learning solution to monaural speech segregation in reverberant conditions. A grouping cue is estimated using MLP's for T-F unit labeling in the grouping stage. Our evaluation shows that the proposed system yields considerable performance improvement over a previous approach in terms of SNR gain and sensitivity to different reverberation conditions. The key advantage of our system lies in the use of supervised learning which circumvents the application of inverse filtering and makes our system more generalizable to a variety of room configurations.

Acknowledgements. This research was supported in part by an AFOSR grant (F49620-04-1-0027) and an NSF grant (IIS-0534707). We thank Y. Li for comments on an earlier draft of this paper.

6. REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. NY: Springer, 2005.

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

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

[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, 2005, pp. 181–197.

[5] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, pp. 684–697, 1999.

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

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

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

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

[10] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Appl. Psychol. Unit, Cambridge Univ., Cambridge, UK, APU Rep. 2341, 1988.

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

[12] G. Hu, "Monaural speech organization and segregation," Ph.D. dissertation, Biophysics Program, The Ohio State University, 2006.

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

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

[15] M. P. Cooke, *Modeling auditory processing and organization*. Cambridge, UK: Cambridge Univ. Press, 1993.