LEARNING TO MAXIMIZE SIGNAL-TO-NOISE RATIO FOR REVERBERANT SPEECH SEGREGATION

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

Monaural speech segregation in reverberant environments is a very difficult problem. We develop a supervised learning approach by proposing an objective function that directly relates to the computational goal of maximizing signal-to-noise ratio. The model trained using this new objective function yields significantly better results for time-frequency unit labeling. In our segregation system, a segmentation and grouping framework is utilized to form reliable segments under reverberant conditions and organize them into streams. Systematic evaluations show very promising results.

Index Terms— Computational auditory scene analysis, monaural speech segregation, objective function, room reverberation, supervised learning.

1. INTRODUCTION

Room reverberation happens in everyday listening and it creates an additional challenge to speech segregation. Most current studies approach the problem using localization cues [1, 2] from more than one microphone, which is less desirable than a monaural solution in many applications [3], e.g., hearing aid design and noise removal for automatic speech recognition. This study is concerned with monaural segregation of reverberant voiced speech.

Pitch, or harmonic structure, has long been studied as a prominent characteristic of speech signals and offers a major cue for a listener to separate target speech from other sounds [4, 5]. The pitch cue has been applied successfully in monaural CASA algorithms under anechoic conditions (e.g., in [6]). However, the harmonic structure is distorted by reverberation as reflections of each harmonic combine with the direct sound. As a result, the performance of pitch-based CASA systems suffers in room reverberation [7]. To tackle this problem, the study in [8] estimates an inverse filter of the room impulse response to counteract the smearing effect of reverberation on speech spectrum. However, the inverse filtering method is very sensitive to even small changes in room configuration [9, 10].

In [10], we proposed a supervised learning approach to achieve robustness against reverberation effects in the computational auditory scene analysis (CASA) framework [7]. A multilayer perceptron (MLP) is trained for each channel of a gammatone filterbank to estimate a harmonic-related grouping cue within each time-frequency (T-F) unit from a set of pitch-based auditory features. A grouping cue encodes the posterior probability of a T-F unit being target dominant given observed features. This approach is shown to be more robust than the inverse filtering method.

In this paper, we propose a segregation system by employing more robust low-level grouping cues and improving the means by which the cues are utilized. By analyzing the goal of maximizing SNR in segregation, we formulate an objective function for MLP training which takes into account of unit-wise errors in a generalized form of mean squared error (MSE). Since it is a continuous function of model parameters, an error backpropagation technique can be devised in order to maximize SNR. In addition, we employ a new segmentation method to more reliably compute auditory segments in reverberant environments. Specifically, we use cross-channel correlation and temporal continuity for segmentation in the low-frequency range because they are observed to be relatively robust to reverberation. In the high-frequency range, we apply onset-offset detection [11] to capture intensity variation and form segments by matching pairs of detected onsets and offsets. It is expected that onset cues are robust to room reverberation in the light of the precedence effect, which refers to the perceptual importance of a direct sound or signal onset. The grouping stage then organizes segments into streams by combining grouping cues.

The paper is organized as follows. In the next section, we derive the new objective function with the goal of maximizing SNR performance. Section 3 describes the overall system. In Section 4, we first show the effect of the new objective function in contrast to conventional MSE minimization, and then evaluate our segregation system. We conclude the paper in Section 5.

2. RELATING OBJECTIVE FUNCTION TO SNR

T-F unit labeling plays an important role in a CASA based segregation system. Reliable unit labeling can be obtained through supervised learning approaches [12]. For frequency channel c and time frame m, a grouping cue $C_g(c,m)$, later used to label T-F unit u_{cm}, encodes the posterior probability of u_{cm} being target dominant given auditory features $\mathbf{x}_{cm}$. The desired value of the grouping cue $C_g(c,m)$ is defined to be 1 if u_{cm} is dominated by the target stream and 0 otherwise, consistent with the notion of the ideal binary mask [13] which labels a T-F unit as target if and only if target energy is greater than interference energy within that unit. Thus, the ideal binary mask provides the desired values of $C_g(c,m)$.

We use an MLP to learn the grouping cue $C_g(c,m)$ from the pitch-based features $\mathbf{x}_{cm}$. Training usually minimizes an objective function (i.e., error function) defined as the square distance between desired and actual outputs. Our previous study [10] uses a conven-

tional MSE objective function, defined as

$$J_c = \frac{1}{M}\sum_m (d_c(m) - y_c(m))^2 \quad (1)$$

where $d_c(m)$ and $y_c(m)$ are desired (binary) and actual outputs, m frame index, M the total number of frames, and c channel index. The model using the above objective function performs reasonably well [10]. However J_c treats all T-F units equally. Such treatment may not be optimal—a T-F unit with higher energy contributes more to the overall SNR than a unit with lower energy. In other words, minimizing J_c does not necessarily lead best SNR performance.

In order to derive an objective function that directly relates to the goal of maximizing SNR, we start by analyzing the SNR definition. Since the computational goal of our proposed segregation system is to identify T-F regions that are target dominant, we use the same SNR measure in [6], which regards the resynthesized signal from the ideal binary mask as ground truth

$$\text{SNR} = 10\log_{10}\frac{\sum_t s_I^2(t)}{\sum_t (s_I(t) - s_E(t))^2}. \quad (2)$$

Here $s_I(t)$ and $s_E(t)$ are signals resynthesized from the ideal binary mask and an estimated mask, respectively. Consider the SNR in a single channel as training is independently conducted within individual channels. To maximize the overall SNR we maximize SNR in each channel. Rewrite (2) for a single channel as

$$\text{SNR}_c = 10\log_{10}\frac{\sum_m d_c(m)\cdot E_c(m)}{\sum_m (d_c(m) - Y_c(m))^2 \cdot E_c(m)} \quad (3)$$

where $E_c(m)$ represents the mixture energy within u_{cm}, calculated as the sum of squares of the unit response. $Y_c(m)$ is an actual binary label, binarized from $y_c(m)$. From (3), it is intuitively clear that minimizing the denominator maximizes SNR_c. Therefore, we define the new objective function J_c' as

$$J_c' = \sum_m (d_c(m) - y_c(m))^2 \cdot E_c(m) \Big/ \sum_m E_c(m). \quad (4)$$

Note that the function J_c' is modified from the denominator in (3) in order to make it differentiable, needed for applying gradient descent learning. The denominator in (4) is added for the purpose of normalization (cf. (1)). It is worth mentioning that J_c' is a generalized form of MSE, with each squared error weighted by normalized energy within the corresponding T-F unit.

3. SYSTEM DESCRIPTION

The input signal is decomposed into a T-F representation using a gammatone filterbank [7]. In each T-F unit, a set of pitch-based features are extracted and the grouping cue for unit labeling is estimated using a trained MLP model. In segmentation and grouping stage, T-F units are merged into segments based on cross-channel correlation in low frequency regions and onset/offset analysis in high frequency regions. Target and background streams are then generated by grouping segments from labeled units and refined in the final segregation step. A binary mask is thus estimated and the reverberant target is segregated from the original mixture by retaining those T-F regions labeled as 1 in the mask and discarding the rest.

3.1. Feature Extraction

To extract pitch-based features, an input mixture is passed through a 128-channel gammatone filterbank whose frequencies are quasi-logarithmically spaced from 50 Hz to 8 kHz. The response of a filter channel is further transduced by the Meddis model of auditory nerve transduction, denoted by $h(c,t)$. Then, the normalized correlogram $A(c,m,\tau)$ is computed using a window of 20 ms with 10 ms overlapping. c is channel index, m is frame index and τ is time lag. For u_{cm}, a 6-dimensional feature vector is extracted in similar form in [10, 12]:

$$\begin{aligned}&\mathbf{x}_{cm} = \\ &\Big\{A(c,m,\tau_m), \big[\bar{f}(c,m)\tau_m\big], \Big|\bar{f}(c,m)\tau_m - \big[\bar{f}(c,m)\tau_m\big]\Big|, \\ &A_E(c,m,\tau_m), \big[\bar{f}_E(c,m)\tau_m\big], \Big|\bar{f}_E(c,m)\tau_m - \big[\bar{f}_E(c,m)\tau_m\big]\Big|\Big\}.\end{aligned} \quad (5)$$

The first three features are derived from $h(c,t)$, suitable for detecting resolved harmonics in low-frequency channels. Given the pitch period τ_m at frame m, $A(c,m,\tau_m)$ is a quantitative measure of how the observed signal in u_{cm} is consistent with τ_m. The average instantaneous frequency $\bar{f}(c,m)$ is estimated from the zero-crossing rate of $A(c,m,\tau)$. When multiplying $\bar{f}(c,m)$ with τ_m, the product provides an alternative way of periodicity comparison and supplements the autocorrelation measure in the feature set. So, the next two features are extracted out of this product: the second feature, the nearest integer $[\cdot]$ to the product, indicates a harmonic number, and the third feature, the distance $|\cdot|$ between the product and the nearest integer, represents the deviation between the two periods. To detect unresolved harmonics, the last three features are based on the envelope of the hair cell output $h_E(c,t)$ and the corresponding normalized correlogram $A_E(c,m,\tau)$. Here, the purpose is to extract amplitude modulation (AM) for high-frequency channels and $h_E(c,t)$ better reveals the periodicities of these harmonics. To extract AM, we perform band-pass filtering with the passband from 50 to 550 Hz, which corresponds to the plausible pitch range of the target speech. When extracting these features, the pitch period τ_m needs to be specified. To remove the influence of pitch errors on the segregation system, we obtain *a priori* pitch contours from the premixed reverberant target speech using Praat [14].

3.2. MLP Training and Labeling

In order to reliably detect both resolved and unresolved harmonics, we train one MLP for each channel. Each MLP has the same network topology with 6 input nodes, 20 hidden nodes and 1 output node. The number of hidden nodes is chosen based on ten-fold cross-validation. The transfer function of the hidden and output layers are both hyperbolic tangent sigmoid. The backpropagation algorithm is adapted to learn MLP parameters. In theory, each of the weights in (4) acts as a constant factor in the partial derivative of J_c'. So the delta rule can be easily rewritten. It should be noted that the normalization term in (4) is necessary to ensure the convergence of the modified backpropagation algorithm [15]. During training, we use J_c' in conjunction with a generalized Levenberg-Marquardt backpropagation algorithm [16] which achieves fast convergence by avoiding the computation of the Hessian Matrix.

The trained MLP estimates the posterior probability directly. Thus, a T-F unit u_{cm} is labeled as 1 if its posterior probability of being target dominant ($C_g(c,m)$) is greater than the posterior probability of interference dominant (1-$C_g(c,m)$). That is,

$$C_g(c,m) > 1/2. \quad (6)$$

3.3. Segmentation and Grouping

To improve segmentation in reverberant speech, we apply two different strategies in different frequency regions. Specifically, in low frequency (below 800 Hz) we merge T-F units into segments based on cross-channel correlation and temporal continuity. Since high-frequency channels are more susceptible to room reverberation, segmentation using cross-channel correlation based on $h_E(c,t)$ is not effective. Signal onsets, on the other hand, are largely unaltered by room reverberation because the direct sound arrives earlier than its echoes. Therefore, we propose that high-frequency regions be segmented using onset and offset detection [11]. This method first smooths signal intensity over time in individual frequency channels to reduce insignificant fluctuations and then over frequency to enhance synchronized onsets and offsets. It then detects onsets and offsets from smoothed intensity in each channel. Segments are formed by matching pairs of onset and offset fronts, which are the vertical contours connecting onset and offset candidates across frequency. In order to achieve a compromise between over- and under-segmentation, a multiscale integration is applied from a coarse scale to a fine scale. Along the scale change, new segments are created and existing segments are better localized. Finally, the segments obtained from different frequency regions are combined to form a complete segmentation.

With unit labels obtained in Section 3.2 together with T-F segments, we group each segment into the target stream if the energy corresponding to its T-F units with target labels (1s) dominates, i.e., greater than the energy of the T-F units with non-target labels (0s). Finally, to group more units into the target stream, we expand each segment in the target stream by iteratively recruiting its neighboring units that are labeled as target and do not belong to any segment. Consequently, a binary mask is formed and the segregated target speech can be resynthesized for performance evaluation [7].

4. RESULTS

To simulate typical room acoustics, we use the image model which is commonly applied for efficient simulation of the acoustic properties of enclosures [17]. In such a model, a pair of physical locations, corresponding to the source and the microphone, decide RIR in a fixed room. In order to simulate both convolutive and additive distortions, we randomly specify the locations of the target and one interfering source and one more location for the microphone. More specifically, we start with anechoic target speech $s(t)$ and anechoic interference $n(t)$. We then generate a simulated room and randomly create a set, $\{\mathbf{r}_T,\mathbf{r}_I,\mathbf{r}_M\}$, representing locations of the target, the interference and the microphone inside the room, respectively. From these locations, two RIR's—$h_T(t)$ and $h_I(t)$—are calculated by the image model. Consequently, a reverberant mixture $r(t)$ is constructed by

$$r(t) = h_T(t) * s(t) + \alpha \cdot h_I(t) * n(t) \tag{7}$$

where "*" denotes convolution. We use α as a coefficient in order to set mixture SNR to 0 dB. The goal of our system is to segregate the reverberant target $h_T(t) * s(t)$ from the mixture $r(t)$.

In order to systematically evaluate the proposed system under different reverberant conditions, we simulate six acoustic rooms with different sizes and their reverberation times (T_{60}'s) range from 0.1 to 0.6 s in steps of 0.1 s. In each room, we randomly create three sets of locations as mentioned above, resulting in three sets of $\{h_T(t), h_I(t)\}$ and three sets of reverberant mixtures created by (7). Our evaluation uses Cooke's corpus [18], which contains 100 noisy utterances constructed by mixing 10 anechoic voiced utterances (target speech) and 10 different types of interference. We generate a total of 1,900 mixtures, with the original 100 mixtures in anechoic and $6 \times 3 \times 100$ mixtures in reverberant conditions.

4.1. MLP Labeling

Given that the computational objective of our segregation system is to identify T-F regions that are target dominant, we adopt the SNR measure defined in (2) to assess the segregation performance using the resynthesized speech from the ideal binary mask as the ground truth. SNR gain is defined as the improvement over the initial SNR before segregation.

To assess the advantage of J' over J in MLP learning, we segregate target speech using a binary mask formed by unit labeling only (i.e., without segmentation and grouping) and evaluate segregation performance in terms of SNR gain. MLP is trained on one set of 100 reverberant mixtures in the room whose $T_{60} = 0.3$ s, which is the same training set used in [10]. Note that we only trained on one of the three random configurations in one room and test on all configurations under all reverberant conditions. Both J and J' objective functions are used in training and their performances are compared in Table 1. The trained MLP using J' performs uniformly better than the one using J, providing more than 1 dB gain on average. Such an improvement is significant for speech segregation systems and is purely brought about by training itself with almost no additional computational cost.

Table 1. Comparison of SNR gain in dB between MSE and generalized MSE (J and J') in MLP training.

	Room Cond. – T_{60}(s)						
MLP's	0.0	0.1	0.2	0.3	0.4	0.5	0.6
J	10.1	10.1	9.5	**9.5**	8.9	9.1	8.0
J'	11.6	11.6	10.6	**10.9**	9.9	10.0	8.4

It is also observed that the difference is greater where training and test are done in the same room (shown in bold numbers). When training and test conditions do not match, both MLPs have performance degradation due to feature variations with changing acoustic environments. This introduces a generalization problem. Next, we evaluate our proposed system in three different scenarios which place different levels of demand on generalization.

4.2. Segregation Evaluation

When reverberation time is known, we train on one set of 100 reverberant mixtures and test the resulting model in the same room. The dotted line in Fig. 1 represents this case. The performance curve depicts the SNR gain of seven separate systems, each trained at a different T_{60}. The observed performance drop with increasing reverberation likely reflects the nature of the ascending difficulty of segregation. In other words, segregation in highly reverberant conditions is probably a harder task than in low reverberant conditions.

With unknown T_{60}, we train on all different T_{60}'s. Specifically, we form a training corpus with a total of 700 reverberant mixtures by using the first set of mixtures in each room together with anechoic mixtures. The pentagram line in Fig. 1 shows the system performance in this case. This way of training gives a single system re-

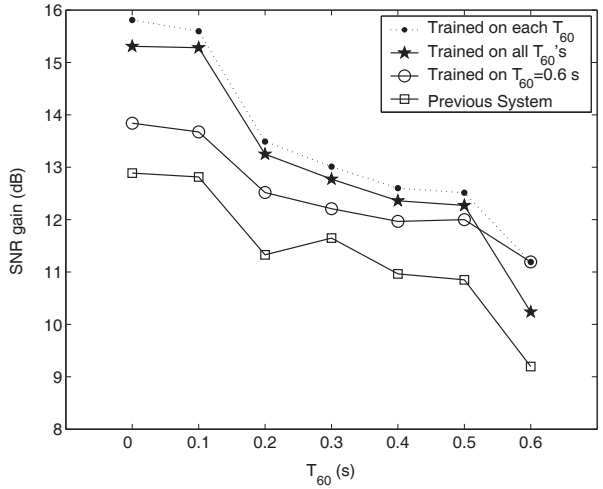

Fig. 1. Voiced speech segregation performance. SNR gain is measured under room conditions with T_{60} ranging from 0 to 0.6 s. The dotted line, the pentagram line and the circle line represent three cases discussed in the text. The performance of the previous system [10] is also presented for comparison.

gardless of reverberant conditions and the performance is only about 0.5 dB worse on average compared to the known room case.

With unknown T_{60}, we can also train on a single T_{60}. If we assume T_{60} is more likely above 0.3 s which is typical of rooms encountered in daily life [19], we can train at $T_{60} = 0.6$ s, the most reverberant condition, because generalization to less reverberation may be better than the other way around. The rationale here is to obtain the best possible classifier under the least favorable condition, often referred to as a MINIMAX solution [20]. The SNR gain of this case is the circle line in Fig. 1. Some degradation is observed, but the system yields relatively good performance at high T_{60}'s.

Finally, we compare the above three curves using the proposed system with the one using [10]. Different from the comparison made in Section 4.1, this is on the overall segregation performance. As can be seen, the proposed system achieves significantly higher SNR gains across all different T_{60}'s than the previous system. This margin reflects the contributions of a more proper training schema and a more effective segmentation and grouping strategy.

5. CONCLUSIONS

This paper develops a supervised learning approach to reverberant speech segregation where a generalized MSE objective function is proposed for MLP training, which directly relates to the goal of maximizing SNR. A multiscale onset and offset analysis is employed for reliable segmentation in high-frequency region. The proposed system is evaluated in three different training scenarios and shows a significant SNR improvement over a previous approach. In the current study, we use *a priori* pitch, calculated from reverberant target speech before mixing, in feature extraction. Although our preliminary evaluation suggests that system performance is not very sensitive to errors in pitch detection, robust pitch estimation in noisy and reverberant conditions is an important topic for future research.

Acknowledgements. This research was supported in part by an AFOSR grant (FA9550-08-1-0155) and an NSF grant (IIS-0534707).

6. REFERENCES

[1] C. Liu, B. C. Wheeler, J. W. D. O'Brien, C. R. Lansing, R. C. Bilger, D. L. Jones, and A. S. Feng, "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," *J. Acoust. Soc. Amer.*, vol. 110, pp. 3218–3231, 2001.

[2] N. Roman, S. Srinivasan, and D. L. Wang, "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4040–4051, 2006.

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

[4] C. J. Darwin, "Listening to speech in the presence of other sounds," *Phil. Trans. R. Soc. B*, vol. 363, pp. 1011–1021, 2008.

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

[6] 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.

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

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

[9] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: robustness results," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 311–319, 2000.

[10] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," in *Proc. IEEE ICASSP*, 2007, pp. 921–924.

[11] G. Hu and D. L. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, pp. 396–405, 2007.

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

[13] 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.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 4.3.14)," 2005. Available at http://www.fon.hum.uva.nl/praat

[15] M. Kukar and I. Kononenko, "Cost-sensitive learning with neural networks," in *European Conference on Artificial Intelligence*, 1998, pp. 445–449.

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

[17] 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.

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

[19] H. Kuttruff, *Room Acoustics*. Taylor & Francis, 2000.

[20] H. L. V. Trees, *Detection, Estimation and Modulation Theory*. NY: Wiley, 1968.