ON THE ROLE OF LOCALIZATION CUES IN BINAURAL SEGREGATION OF REVERBERANT SPEECH

John Woodruff and DeLiang Wang

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

ABSTRACT

Approaches to binaural and stereo speech segregation have often assumed that localization information can be used as a primary cue to achieve segregation of a target signal. Results produced by these systems degrade significantly in the presence of room reverberation. In this work, we present an alternative framework to achieve localization of groups of time-frequency units. We show that grouping across time and frequency allows the use of localization as an important cue for sequential grouping of time-frequency objects. We analyze the level of time-frequency grouping needed to achieve accurate object localization and show preliminary binaural segregation results using the proposed framework. Results indicate that both localization and segregation performance can be improved by grouping across time and frequency.

Index Terms — Binaural sound localization, speech segregation, reverberation, computational auditory scene analysis.

1. INTRODUCTION

Humans frequently encounter situations in which they must attend to an individual's speech in the presence of competing sound sources and room reverberation. Those with normal hearing are able to segregate a desired signal from the background noise and discern vital information regarding its content and location. Computational auditory scene analysis attempts to recreate this phenomenon in machines using known principles of human perception [16]. A system that can reliably isolate the content of an individual source signal embedded in competing signals would be of tremendous use in hearing prostheses and other speech processing applications.

Much attention has been paid to the use of localization cues in multi-channel audio recordings. Beamforming attempts to improve SNR of a source using directional information [3, 8]. Other approaches perform a time-frequency decomposition of the mixture signals and use between channel level and time delay differences in each time-frequency (T-F) unit to estimate an output signal that originates from a particular direction [8, 12, 14, 18]. These systems use localization information as a primary cue to achieve source segregation, and show rapid performance degradation as reverberation is added to the recordings.

Darwin has suggested that it is unlikely that the auditory system uses spatial information as a primary means of attending to individual sources, as individual localization cues are highly unreliable in reverberant environments [6]. He argues that much of the psychophysical evidence supports an alternative framework in which sound source localization may be a product of source segregation, rather than a primary cue used to achieve it. In this account, monaural cues (e.g. harmonicity, common onsets, common amplitude and frequency modulation) are used by the auditory system to form auditory *objects*, or groups of T-F units. The auditory system can then localize these objects to create auditory space and as one major mechanism for sequentially grouping objects into a single auditory *stream*.

In this work, we explore a computational framework for binaural speech segregation in reverberant environments. The proposed ideas represent a significant departure from the way source location has been utilized within a binaural segregation system. We argue that using monaural cues to perform initial T-F grouping will allow for more robust localization of these T-F groups, enabling a system to use the spatial information regarding the T-F groups as a means for sequential organization.

We first provide background details in Section 2. Section 3 outlines the methods used for localization of T-F groups and discusses the levels of T-F grouping explored in this study. Section 4 presents localization performance and binaural segregation results at different levels of T-F grouping. We conclude with a discussion of our preliminary findings and outline directions of future research.

2. BACKGROUND

In this study we use the ROOMSIM package [5] to generate impulse responses that simulate binaural input at human

ears. We generate a library of left and right ear impulse responses for direct sound azimuth angles between 0° and 90° (spaced by 5°) and reverberation times between 0 and 1.6 seconds. In all tests we use monaural speech signals drawn from the TIMIT database, pass the signals through a left and right impulse response for a desired azimuth angle and room reverberation time, and sum the resulting signals in the case of multi-source mixtures.

We pass the mixture signals through a gammatone filter bank with center frequencies spaced according to the ERB scale [12]. Each auditory filter is then processed using a model of neural transduction as described in [2].

The two primary cues used in localization of sound sources in the free field are interaural time difference (ITD) and interaural level difference (ILD) [1]. ITD is calculated as the time delay that produces a maximum in the normalized running cross-correlation between the left and right mixture signals. ILD corresponds to the energy ratio in dB between the two signals at each time instance. We calculate ITD and ILD for each sample and frequency channel as described in [8].

3. LOCALIZATION OF TIME-FREQUENCY GROUPS

As discussed in the introduction, reverberation hinders the direct use of ITD and ILD as a primary means of source segregation. Our proposed computational framework uses localization cues in a fundamentally different way. This framework requires that groups of T-F units be formed using monaural cues in the mixture signals. Once formed, individual localization cues may be pooled over the entire region, providing more accurate localization judgments.

3.1. Time-frequency grouping

Time-frequency grouping refers to the process of joining T-F units together that are thought to primarily contain energy from a single source. Grouping is performed in order to pool data across time and/or frequency, allowing for more reliable labeling (target/interference) decisions. Numerous authors have examined methods of time-frequency grouping using monaurally available cues in both anechoic and reverberant environments [4, 9, 10]. Since it is outside the scope of this study to propose mechanisms for monaural grouping of T-F units, we assume groupings of T-F units can be formed using cues unrelated to localization. For all of the simulations discussed below, we make use of the ideal binary mask (IBM) [15] as a mechanism for grouping T-F units generated by the same source signal. The IBM is a binary labeling of a mixture's time-frequency decomposition where pre-mixing signals are used to measure whether each T-F unit is target dominant (labeled 1) or interference dominant (labeled 0).

In the development of a system with the proposed framework, some fundamental questions arise. First, should grouping be performed across both time and frequency? How much grouping is necessary if one hopes to accurately localize the objects formed? How should one pool the individual within-object cues to form a reliable judgment of source location?

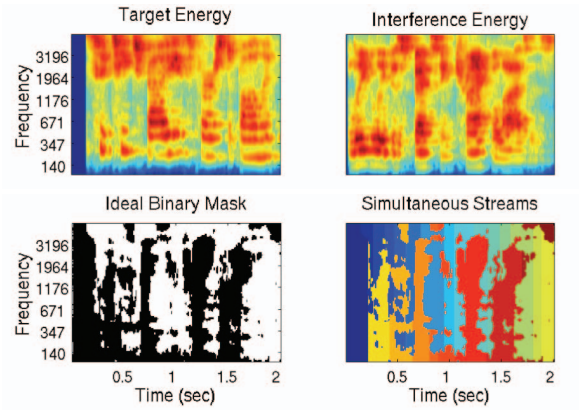

Fig. 1. Illustration of the formation of T-F objects. (top left) Cochleagram of target signal. (top right) Cochleagram of interference signal. (bottom left) Ideal binary mask for target signal where 1 is white and 0 is black. (bottom right) Simultaneous streams created with IBM and manual labeling of syllable boundaries.

To analyze these questions we perform object localization at several time-frequency grouping levels. As a baseline for comparison, we estimate a source signal's azimuth in each individual T-F unit. Second, we group T-F units across time into T-segments [9], or contiguous regions of time dominated by one source signal. Third, we group units dominated by one source across time and frequency into contiguous T-F segments. Finally, we create *simultaneous streams*, or (not necessarily contiguous) groups of T-F units dominated by the same syllable in a given speaker's utterance. In all experiments presented here, we make use of the IBM and manually labeled syllable boundaries to form the T-F groups described. Figure 1 shows the time-frequency decomposition of two utterances, the IBM formed from a 0 dB mixture pair and the simultaneous stream T-F groups.

The different grouping levels correspond to different amounts of assumed monaural grouping. At the T-F unit level, monaural cues are not utilized at all and localization is responsible for all T-F grouping. At the simultaneous stream level, it is assumed that monaural cues are able to perform substantial simultaneous grouping and localization provides primarily sequential grouping information.

3.2. Azimuth estimation methods

At each level of grouping, we perform localization using three different methods. All methods involve calibration where we pass white noise through the anechoic impulse

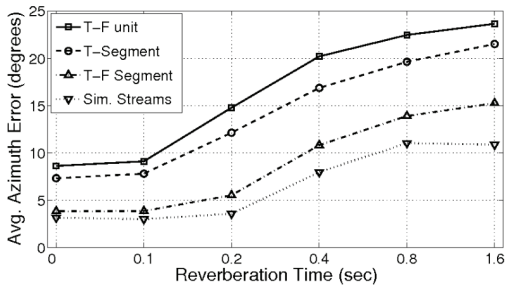

Fig. 2. Average azimuth estimation error at four T-F grouping levels: T-F unit, T-Segment, Segment and Simultaneous Stream.

responses of angles 0° through 90°, and measure the resulting ITDs and ILDs in all frequency channels. Using the measured data, we calculate two functions. First, we use Kernel density estimation [13] to calculate a likelihood function, $p_c(x_c|\phi)$ for each angle ϕ, and frequency channel c. Here, x_c denotes an ITD-ILD pair for the specified frequency channel. Each density function is most responsive to the ITD-ILD pair that would be measured from a single source at a given angle in anechoic conditions, and decays away from that ITD-ILD pair. From the joint ITD-ILD densities, we create a univariate likelihood function for ITD only by marginalizing the ILD dimension. We denote this by $p_c(\tau_c|\phi)$, where τ_c is an ITD value measured in frequency channel c. Using these calibration schemes, the three azimuth estimation methods explored are as follows:

Method 1: $\hat{\phi} = \arg\max_{\phi} p_c(x_c \mid \phi)$

Method 2: $\hat{\phi} = \arg\max_{\phi} p_c(\tau_c \mid \phi)$

Method 3: Method 2 below 1500 Hz, and Method 1 at or above 1500 Hz

The 1500 Hz cutoff is used in Method 3 because it is roughly the frequency at which phase difference information between left and right signals cannot be uniquely decoded into interaural time difference.

To estimate an angle for an entire T-F group, we take the sum of the log-likelihood values produced by each measured ITD or ITD-ILD pair and classify the entire T-F group as the angle that produces the maximum value. This approach inherently assumes that ITD and ILD measurements are independent when conditioned on a source's azimuth angle.

3.3. Localization results

We use the methods described above to estimate the azimuth angle of the source that was dominant in each T-F unit, T-segment, segment or simultaneous stream. Figure 2 shows the average localization error for the best performing localization method (Method 3) at each of the grouping levels. We can see that each stage of T-F grouping allows for more accurate localization judgments at all reverberation times. A decrease of 13° is seen between T-F unit localization and simultaneous stream localization in the most reverberant conditions. In all cases, Method 3 localization performed best, suggesting that ILD is not as reliable for lower frequency channels. This finding is consistent with human sound localization [1].

4. SEQUENTIAL ORGANIZATION USING LOCALIZATION

As an illustration of sequential organization and segregation performance, we label the T-F groups of an example mixture using pooled localization data. We assume source positions are known and label each T-F group as target or interference using the likelihood functions generated for Method 3. We generate a label for T-F group i using,

$$\arg\max_{k}\left(\sum_{\Gamma_i^{high}} \log\left(p_c(x_c(n)\mid\phi_k\right) + \sum_{\Gamma_i^{low}} \log\left(p_c(\tau_c(n)\mid\phi_k\right)\right) \quad (1)$$

where $\{c,n\}\in\Gamma_i$ denotes a set of frequency channels, c, and time samples, n, for T-F group, i, and $k\in\{0,1\}$ indexes the target (1) and interference (0) sources. Since we are using Method 3, we break Γ_i into a set that contains high frequency channels and a set that contains low frequency channels. Many other labeling approaches are possible and will be explored in subsequent work.

Figure 3 shows the output SNR of the segregated target signal at all four levels of T-F grouping, as well as the output SNR using the IBM. Note that the *largest* T-F group level, simultaneous streams, achieves the *worst* performance in reverberant conditions. This suggests that pooling all of the data within the entire stream does not necessarily create better decisions when many of the individual ITD and ILD measurements are unreliable. It is also important to note that at this level of grouping, a single wrong decision substantially degrades SNR performance.

4.1. Cue Selection

The results in the above example suggest that simple pooling of data throughout the T-F groups is not sufficient for good segregation performance. Others have proposed cue selection mechanisms for localization of single or multiple source signals over entire utterances [7, 17].

Faller and Merimaa propose that *interaural coherence* (IC), the value of the ITD peak in the normalized cross-correlation function is a good indicator of ITD and ILD reliability. In our proposed framework, selecting only those ITD-ILD samples in which the IC value is high (close to 1) may improve the labeling of large T-F groups.

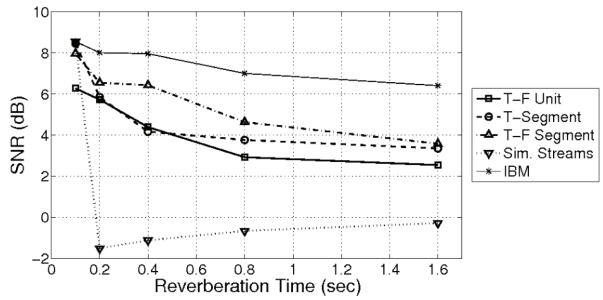

Fig. 3. SNR output for four T-F grouping levels and IBM.

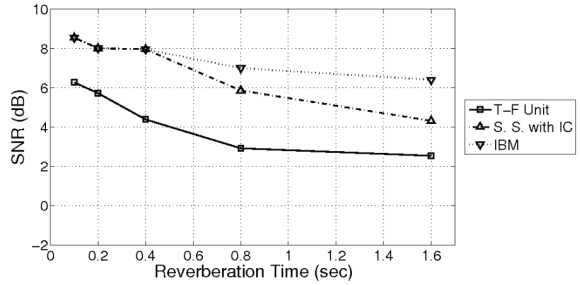

Fig. 4. SNR output for T-F unit level, simultaneous stream level including interaural coherence and IBM.

In Figure 4, we now show the segregation results of simultaneous streams grouping where within each T-F group, we select the most coherent samples such that the remaining signal power in the T-F group is 25% of the power in the whole group. In this case, IC disregards cues that were impairing labeling decisions, and segregation performance improves in highly reverberant conditions.

5. DISCUSSION

We have presented a new computational framework for binaural segregation of speech in reverberant environments. This approach follows findings in psychophysical experiments suggesting that source localization is a product of monaural grouping rather than a primary cue used to achieve segregation. We illustrate that if monaural cues can be used to form T-F objects, the resulting objects can be accurately localized and these judgments can be used as a means of sequentially grouping objects into a cohesive auditory stream. We have shown that simple pooling of all ITD and ILD data within T-F produces more accurate localization, but that with large T-F groups, single wrong decisions result in poor SNR performance. The incorporation of cue selection based on interaural coherence improved segregation results, but one aspect of future work will be to develop a more reliable selection mechanism.

Numerous challenges remain in the creation of a full-fledged system that operates within the proposed framework. Although researchers have made progress on monaural grouping of T-F units in reverberant environments [4, 11], this remains a challenging problem that was not addressed in this study.

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

6. REFERENCES

[1] Blauert, J., *Spatial Hearing*, MIT Press, Cambridge, MA, 1996.

[2] Bernstein, L.R., S. van de Par, and C. Trahiotis, "The normalized interaural correlation: Accounting for NoSπ thresholds obtained with Gaussian and "low-noise" masking noise," *J. Acoust. Soc. Amer.*, vol. 100, pp. 870-876, 1999.

[3] Brandstein, M. and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, 2001.

[4] Brown, G.J. and K.J. Palomäki, "Reverberation," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley/IEEE Press, Hoboken, NJ, 2006

[5] Campbell, D.R., *The ROOMSIM User Guide*, Available at http://media.paisley.ac.uk/~campbell/Roomsim/

[6] Darwin, C. J., "Spatial Hearing and Perceiving Sources," in *Auditory perception of sound sources*, Eds Yost, W. A., Fay R. R. and Popper, A. N., Springer-Verlag, 2008.

[7] Faller, C. and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, pp. 3075-3089, 2004.

[8] Feng, A.S. and D.L. Jones, "Localization-Based Grouping," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley/IEEE Press, Hoboken, NJ, 2006

[9] Hu, G. and D.L. Wang, "Segregation of unvoiced speech from nonspeech interference,", *J. Acoust. Soc. Amer.*, vol. 124, pp. 1306-1319, 2008.

[10] Jin, Z. and D.L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *Tech. Report*, OSU-CISRC-5/08-TR27, The Ohio State University, 2008.

[11] Patterson, R.D., M.H. Allerhand, and C. Giguère, "Time-domain modeling of peripheral auditory processing: A modular architecture and software platform," *J. Acoust. Soc. Amer.*, vol. 98, pp. 1890-1894, 1995.

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

[13] Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, NY, 1986.

[14] Viste, H, *Binaural Localization and Separation Techniques*, Ph.D. Thesis, Swiss Federal Institute of Technology, 2004.

[15] Wang, D.L, "On ideal binary masks as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Kluwer Academic, Boston, MA, 2005.

[16] Wang, D.L., G.J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley/IEEE Press, 2006.

[17] Wilson, K.W. and T. Darrell, "Learning a precedence effect-like weighting function for the generalized cross-correlation framework," *IEEE TASLP*, vol. 14, no. 6, 2156-2164, 2006.

[18] Sawada, H., S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propogation model parameters in frequency-domain blind source separation," *IEEE TASLP*, vol. 15, no. 5, pp. 1592-1604, 2007.