THE SEGREGATION OF SPATIALISED SPEECH IN INTERFERENCE BY OPTIMAL MAPPING OF DIVERSE CUES

Jingbo Gao and Anthony I. Tew

Audio Lab, Department of Electronics
University of York
York YO10 5DD, UK
jg757@york.ac.uk, tony.tew@york.ac.uk

ABSTRACT

We describe optimal cue mapping (OCM), a potentially real-time binaural signal processing method for segregating a sound source in the presence of multiple interfering 3D sound sources. Spatial cues are extracted from a multi-source binaural mixture and used to train artificial neural networks (ANNs) to estimate the spectral energy fraction of a wanted speech source in the mixture. Once trained, the ANN outputs form a spectral ratio mask which is applied frame-by-frame to the mixture to approximate the magnitude spectrum of the wanted speech. The speech intelligibility performance of the OCM algorithm for anechoic sound sources is evaluated on previously unseen speech mixtures using the STOI automated measures, and compared with an established reference method. The optimized integration of multiple cues offers clear performance benefits and the ability to quantify the relative importance of each cue will facilitate computationally efficient implementations.

Index Terms **—** Speech segregation, neural networks, ratio mask

1. INTRODUCTION

In everyday life, the human hearing system displays a remarkable ability to attend selectively to a single sound source in a mixture of competing concurrent sources, often referred to as the cocktail party effect [1]. In practice, the interfering sounds will exhibit a range of characteristics: for example, they may be stochastic or periodic in nature or a mixture of the two, they may be relatively uncorrelated with the target, as is usually the case with multiple talkers, or highly correlated, as in the case of reverberation. These variations may create conflicting requirements when attempting to segregate the target sound.

Methods involving the estimation of a multiplicative mask in the frequency domain show great promise. The mask is used to extract the target speech from the mixture. [3], [5], [6] describe and develop a sound segregation model based on an ideal binary mask (IBM): essentially, when the target speech is estimated to be dominant in a time-frequency unit (TFU), that mask element is set to unity, otherwise it is set to zero. The difficulty here is how best to estimate the IBM when only the mixture is available.

An important benefit of listening to speech with both ears is the intelligibility improvement it yields in acoustically challenging conditions [2]. In one of the few papers to discuss target segregation in a binaural mixture [3], mask estimation is informed by the interaural time difference (ITD) and interaural intensity difference (IID) spatial cues. The importance of these cues in localization has long been recognized [19], but acoustic signals are rich with other cues to assist with source segregation, not only spatial, but also from the acoustic environment and originating in the sources themselves. The use of these additional cues has the potential to improve segregation performance. In considering this possibility, however, it is desirable to be able to determine the strength of each one, so that weak cues are not integrated, possibly wasting limited computational resources.

A difficulty with the IBM is that it becomes increasingly sparse as the signal-to-noise ratio (SNR) of a target speech source reduces, causing intelligibility to deteriorate, albeit gracefully. In these circumstances, the ideal ratio mask (IRM) has been shown to perform well in automatic speech recognition tasks [7], [8]. Mask values between 0 and 1 are allowed and indicate the probability of target dominance in each TFU. The IRM leads to better objective intelligibility than IBM [17], in part because errors in estimation are likely to have less perceptual impact. [17] is distinctive in its use of a deep neural network (DNN) to optimally estimate the IRM for a monaural mixture. In further pioneering work on monaural mixtures, Wang *et al.* address the question of which features most effectively inform the mask estimation process [20].

In this paper, we apply a simple two-layer feed-forward artificial neural network (ANN) to segregate speech in binaural mixtures and perform an input importance analysis. We demonstrate the viability of the approach using binaural cues such as interaural phase difference (IPD), interaural

level difference (ILD) and their deltas, together with other information extracted from the binaural mixture. Using models, we evaluate the performance of the algorithm in terms of target intelligibility and demonstrate the smooth variation with frequency of the relative importance of six particular inputs.

2. ALGORITHM DESCRIPTION

2.1. Cue selection

It is well known that binaural cues play a key role in determining the direction of arrival of the sound sources in the human auditory system. At frequencies below about 1.5 kHz, the dominant cue is ITD (or IPD), whereas above this frequency ILDs have been shown to play a greater role [9]. We use the short-time Fourier transform (STFT) to compute these cues on a per-frame basis. Two complex-valued spectra $X_l(m,b)$ and $X_r(m,b)$ are obtained for the left and right channels, where the integers m and b denote the time frame index and frequency band, respectively. Then IPD and ILD are extracted according to:

$$IPD(m,b) = \phi\{X_l(m,b)\} - \phi\{X_r(m,b)\} \quad (1)$$

$$ILD(m,b) = 20log10\left(\frac{|X_l(m,b)|}{|X_r(m,b)|}\right) \quad (2)$$

where ϕ is the unwrapped phase of the spectra.

We also choose to extract the first differences (rates of change) of IPD and ILD. We do so to demonstrate the ability of the OCM method to determine the strength of these cues and to combine them optimally with the IPD and ILD cues, such that they are likely to improve the estimate for both an IBM and an IRM:

$$\Delta IPD(m,b) = IPD(m,b) - IPD(m-1,b) \quad (3)$$

$$\Delta ILD(m,b) = ILD(m,b) - ILD(m-1,b) \quad (4)$$

For example, consider ΔIPD in the case that the target speech alone is active and is located at 0° azimuth (straight ahead of the listener). Then both *IPD* and ΔIPD will be stable and approximately 0 and the IRM value will be 1. If, instead, two interferers are active, one at +30° and the other at -30° azimuth, and the target is inactive, the IPD will vary erratically and will occasionally be close to zero. Now, however, ΔIPD will generally be non-zero and the IRM value will be 0. In this scenario therefore, when *IPD* is zero, the value of ΔIPD is a strong indicator of the mask value and can be considered important. A similar argument applies to ΔILD.

The fifth ANN input is TFU magnitude, which we anticipated would prove to be of moderate importance through its ability to detect source inactivity at the frequency in question. E.g., when the magnitude is close to 0, little target speech energy is present and the mask ratio should be forced to 0, whereas when interferers and/or the target are active the ratio may lie between 0 and 1.

As a further demonstration of the OCM method we also investigate the importance of the interaural coherence (Coh) as an ANN input. Coh is used in many separation and dereverberation algorithms [10], [11]. In this paper we consider anechoic sources only and so it is likely that Coh will contribute little to the estimation of the mask values. Coh is obtained recursively from the binaural mixture in the manner described in [10].

2.2. Time-frequency mask

The ideal ratio mask $R(m,b)$ is found by using the priori energy ratio:

$$R(m,b) = \frac{|T(m,b)|^2}{|T(m,b)|^2 + |I(m,b)|^2} \quad (5)$$

where $T(m,b)$ and $I(m,b)$ are the spectra of the target speech and interference, respectively. In addition, the ideal binary mask is defined as:

$$B(m,b) = \begin{cases} 0 & R(m,b) > 0.5 \\ 1 & R(m,b) \leq 0.5 \end{cases} \quad (6)$$

2.3. Neural network-based mask estimator

Early research [12], [13] has shown that a multi-layer feed-forward network is a universal function approximator. In addressing our mapping problem we found the performance of two-layer and three-layer networks to be very similar. Since the training process of a three-layer network is more time-consuming, we employ a two-layer feed-forward network to estimate the mask value for each frequency point.

To improve neural network generalization, we trained at least 20 ANNs for each frequency point and selected the five with the lowest mean square estimation errors using test data. Averaging the outputs of these 5 ANNs reduces the mean square error (MSE) by approximately 3% compared with the MSEs of each individual ANN. ANN inputs were IPD, ILD, ΔIPD, ΔILD, Coh, and the TFU magnitude and the expected output was either IBM or IRM.

In order to determine the optimum number of neurons for the ANN at each frequency point, 20 ANNs with only one neuron were trained for 32 frequency points out of the 257 described in section 3.1, equally spaced on the equivalent rectangular bandwidth scale. Their MSE was calculated at the testing stage and this was repeated for ANNs with from 2 to 20 neurons and the resulting set of MSEs is plotted in Figure 1. Substantial improvements in MSE are observable up to 10 neurons, beyond which there is very little further improvement. ANNs with 15 neurons

exhibit a MSE which is within 0.1% of the asymptotic value and this network topology is used throughout our OCM model.

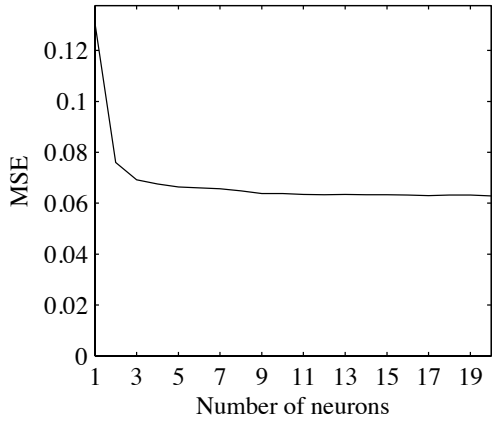

Fig. 1: MSE performance of 20 ANNs averaged across frequency as a function the number of neurons.

3. TRAINING AND ANALYSIS

3.1. Experimental set up

Training speech was recorded by four males and two females in the University of York anechoic chamber. The HRIRs of SYMARE subject HA01 [14] were used to spatialize the sound sources in the binaural mixture. Following [3] as a starting point, we configured our model with three sources. All sources were in the horizontal plane, with the target at 0° azimuth and the interferers at -30° and 30° azimuth, respectively. ANN input data was generated from contiguous pairs of STFT frames (512-point Hann window with 50% overlap).

STFT frames from the binaural mixture, consisting of 257 frequencies from 0 Hz up to and including 8kHz were used to generate data for the six ANN inputs defined in section 2.1 and the corresponding ideal outputs, IBM and IRM. The training data consisted of 35,000 items for each frequency bin. 80% of the items were used to train the ANNs and the rest were used for selecting the best five ANNs, as described in section 2.3. There is a left/right pair of ANNs for each frequency and hence 514 ANNs in total (in this paper, we consider only the left channel). The ANNs were trained by using the backpropagation algorithm with MSE as the cost function. The ANNs were trained using the York Advanced Research Computing Cluster [16].

3.2. ANN input importance analysis

Examining the level of contribution by each input to the mask value estimated by the ANN reveals the relative importance of each input. Several methods of importance estimation have been reported [18, 21, 22]. Here we employ Garson's method [18] to analyze each of the inputs. We present the results of analyzing the ANNs which were trained to estimate ratio mask values. Similar results are displayed by the ANNs which estimate binary masks.

Fig. 2 plots the envelopes of six histograms, showing the number of ANNs in 10 bands of normalized relative importance values for each input type. The results indicate that IPD and ILD contribute most strongly and in equal measure to ratio estimation, followed by ΔIPD, ΔILD and Mag. As expected, interaural coherence is determined to be least important in this anechoic situation.

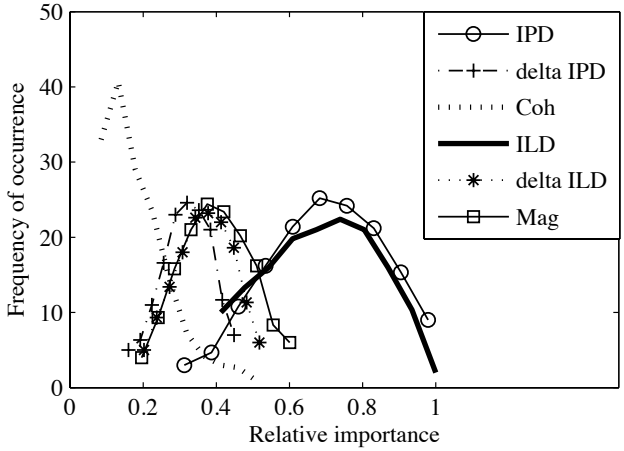

Fig. 2: Number of ANNs in each band of relative importance values, for the inputs IPD, ΔIPD, ILD, ΔILD, coherence (Coh) and TFU magnitude (Mag).

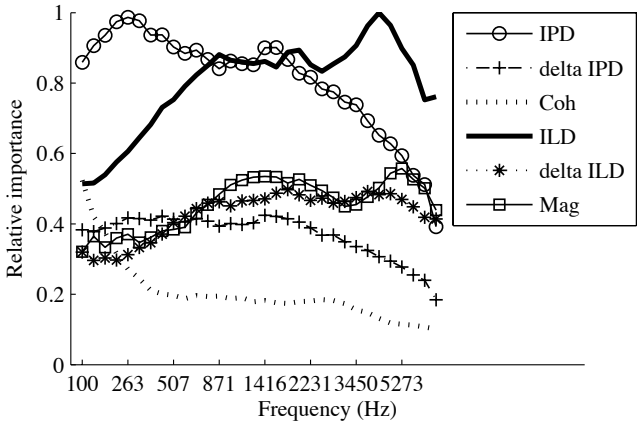

Fig. 3: Relative importance of the inputs IPD, ΔIPD, coherence (Coh), ILD, ΔILD and TFU magnitude (Mag) for all ANNs from 100 Hz to 8000 Hz.

Fig. 3 breaks the analysis down further and shows how the importance of each input varies across frequency. Since IPD is closely related to ITD, this result supports the well-known observation that ITD is the dominant cue at

frequencies below about 1.5 kHz and ILD dominates at higher frequencies. ΔIPD and ΔILD follow a similar pattern, even though they contribute less overall. The importance of TFU magnitude increases slightly with frequency. This could be because there tend to be more periods of inactivity at higher frequencies, increasing the opportunities for this input to contribute to the output estimate. As expected, interaural coherence is little used, except at low frequencies where correlation between neighboring frames may exist.

3.3. Target speech segregation performance

The system's ability to segregate target speech that the ANNs had not been exposed to during training was evaluated. The short-time objective intelligibility (STOI) measure [15] was evaluated for SNRs of -8 dB, -5 dB and 0 dB. The test data employed in [3] was used throughout. The STOI model yields scores between 0 and 1, where a higher score indicates higher intelligibility.

Sys.	STOI		
	-8 dB	-5 dB	0 dB
	Interferer: cocktail party noise, speech utterance		
MIX	0.320	0.392	0.537
EBM2	0.508	0.551	0.766
EBM6	0.567	0.634	0.784
IBM	*0.741*	*0.808*	*0.881*
ERM6	**0.607**	**0.683**	**0.795**
IRM	*0.870*	*0.889*	*0.918*
	Interferer: rock music, speech utterance		
MIX	0.470	0.542	0.676
EBM2	0.728	0.768	0.728
EBM6	0.766	0.817	0.880
IBM	*0.822*	*0.863*	*0.895*
ERM6	**0.801**	**0.847**	**0.889**
IRM	*0.874*	*0.890*	*0.908*
	Interferer: two different speech utterances		
MIX	0.499	0.577	0.708
EBM2	0.799	0.838	0.894
EBM6	0.838	0.865	0.904
IBM	*0.866*	*0.885*	*0.910*
ERM6	**0.843**	**0.873**	**0.905**
IRM	*0.888*	*0.901*	*0.917*

Table 1: Comparison between different systems across different SNRs for a three-source configuration with the target at 0° and two interferers at 30° and -30°, respectively. Numbers in bold are the best real results for each test condition and numbers in italics are the ideal results.

The upper section of Table 1 presents the results from the STOI speech intelligibility model for two cocktail party interferers at -30^0 and $+30^0$ azimuth. The performance of each of the following three masks appears in the rows:

- EBM2 (binary mask, using two inputs; IPD and ILD only which is similar to ITD and IID in [3]);
- EBM6 (binary mask, using all six of the inputs described in section 2.1) and
- ERM6 (ratio mask using the same six inputs).

The performances of the ideal binary and ratio masks appear in rows IBM and IRM, respectively. The results show that intelligibility improves with the inclusion of a richer input set and that the extended data has been integrated successfully into the mapping. They also indicate that the ratio mask performance is superior to that of the binary mask. Similar results are obtained for the other interferer conditions using the rock music and the two different speech utterances, respectively. These results are in line with expectation and demonstrate the successful integration of the richer input cues for mask estimation.

4. DISCUSSION

We have selected six cues for the purpose of demonstrating the OCM method. The variety of segregation cues can be easily expanded to include properties of the acoustic environment and source characteristics. Using the methods outlined in section 3, it is possible to determine the varying relative importance of the input cues for estimating the mask under a range of typical acoustic conditions. In hearing aid design, for example, this knowledge will allow maximum benefit to be drawn from the limited computational resources available and will assist in the development of segregation algorithms which are able to adapt in a continuous fashion as acoustic conditions change. A priority for further research is a binaural analysis of input importance in the presence of early reflections and reverberation.

5. CONCLUSIONS

Optimal cue mapping (OCM) is a straightforward ANN-based binaural ratio mask estimation method for speech segregation. Through modeling, we have demonstrated OCM's ability to integrate binaural cues and other acoustic features to segregate speech of high perceptual quality and intelligibility in the presence of multiple interfering anechoic sounds. OCM also has the potential to assist in establishing the relative importance of acoustic spatial cues, of properties of the acoustic space and of sound source features in a binaural mix. Work is in progress to expand the approach to accommodate increasingly complex acoustic environments.

6. REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, pp. 975–979, 1953.

[2] A. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acust. united with Acust.*, 2000.

[3] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, no. 4, p. 2236, 2003.

[5] D. Wang, "On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis," *Speech Sep. by humans Mach.*, pp. 181–197, 2005.

[6] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking.," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2336–47, Apr. 2009.

[7] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 58–67, Jan. 2006.

[8] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, no. 11, pp. 1486–1501, Nov. 2006.

[9] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, Cambridge, MA, MIT Press (1997).

[10] A. Westermann, J. M. Buchholz, and T. Dau, "Binaural dereverberation based on interaural coherence histograms.," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 2767–77, May 2013.

[11] A. Alinaghi, "Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation," *Acoust. Speech Signal Process.*, pp. 684–688, 2013.

[12] B. Irie and S. Miyake, "Capabilities of three-layered perceptrons," *Neural Networks, 1988., IEEE Int. Conf. on. IEEE*, pp. 641–648, 1988.

[13] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, pp. 359–366, 1989.

[14] C. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. van Schaik, A. I. Tew, C. T. Hetherington and J. Thorpe, "Creating the Sydney-York morphological and acoustic recordings of ears database," *IEEE Transactions on Multimedia*, vol 16, no. 1, pp. 31-46.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[16] York Advanced Research Computing Cluster (YARCC), available https://wiki.york.ac.uk/display/RHPC/YARCC+-+York+Advanced+Research+Computing+Cluster [Accessed: 29 Sep 2014]

[17] Y. Wang, A. Narayanan, and D. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[18] G. D. Garson, "Interpreting neural-network connection weights," *AI expert6.4*, pp. 46-51, 1991

[19] Lord Rayleigh, "On our perception of sound direction," *Phil. Mag.*, vol. 13, pp. 214-232, 1907.

[20] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions in Audio Speech and Signal Processing*, vol. 21, no. 2, pp 270-279.

[21] S. L. Özesmi and U. Özesmi, "An artificial neural network approach to spatial habitat modelling with interspecific interaction," *Ecol. Modell.*, vol. 116, no. 1, pp. 15–31, Mar. 1999.

[22] M. Scardi and L. W. Harding, "Developing an empirical model of phytoplankton primary production: a neural network case study," *Ecol. Modell.*, vol. 120, no. 2–3, pp. 213–223, Aug. 1999.