INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

An interaural magnification algorithm for enhancement of naturally-occurring level differences

Shadi Pirhosseinloo[1], *Kostas Kokkinakis*[2]

[1] Department of Electrical Engineering and Computer Science, University of Kansas, USA
[2]Department of Speech-Language-Hearing, University of Kansas, USA

shadi@ku.edu, kokkinak@ku.edu

Abstract

In this work, we describe an interaural magnification algorithm for speech enhancement in noise and reverberation. The proposed algorithm operates by magnifying the interaural level differences corresponding to the interfering sound source. The enhanced signal outputs are estimated by processing the signal inputs with the interaurally-magnified head-related transfer functions. Experimental results with speech masked by a single interfering source in anechoic and reverberant scenarios indicate that the proposed algorithm yields an increased benefit due to spatial release from masking and a much higher perceived speech quality.

1. Introduction

In many everyday listening situations, a listener's goal is to hear out a specific sound of interest (target) from amongst a mixture of other interfering sounds. Despite the fact that all of these individual sounds are summed up into a single acoustic waveform, the binaural hearing system can very efficiently separate between different voices in a noisy environment and solve what has been coined as the binaural (or two-eared) cocktail-party problem [1]. This separation is accomplished by the use of binaural cues, such as interaural differences in level and time. *Interaural level differences* (ILDs) are the differences in the overall intensity or level of the signals received at the two ears. A signal with higher intensity at the left ear is perceived as a sound source located to the left of the listener. *Interaural time differences* (ITDs) refer to the different arrival times of signals at each ear due to the spatial separation of the two ears. A signal that reaches the left ear earlier than the right ear will be perceived as a sound source located to the left of the listener. Typically, ILDs are more informative regarding azimuth locations at frequencies above 3 kHz and ITDs below 1.5 kHz [2].

In order to better describe these effects, consider the typical scenario where the target source is located right in front of the listener, whereas another source (interferer) is located to the right of the listener. In this scenario, the sounds reaching each ear are transformed in a directionally dependent manner through filters associated with each ear. These linear time-invariant transfer functions can accurately capture the direction dependent effects of the head on the signals received at the two ears and are commonly known as the *head-related transfer functions* (HRTFs). To generate a binaural signal, the HRTFs are convolved with an input acoustic signal, generating a stereo signal with binaural cues associated with a source from a specific azimuth relative to the listener. For a source in front of the listener, there is very little difference in either the magnitude or the phase responses for both ears. For another source placed to the right of the listener, the magnitude of the source in the right ear is greater than the one on the left, while the time delay in the left ear is longer (i.e., the sound arrives at the right ear sooner than in the left ear). In order to obtain better speech intelligibility even at poor *signal-to-noise ratios* (SNRs), normal-hearing listeners often take advantage of these perceived differences in magnitude and the fact that in most cases the target and competitors are spatially separated. This benefit, known as *spatial release from masking* (SRM), is fairly robust in normal-hearing listeners and has been well-established in the literature with many researchers demonstrating that speech perception is markedly better when the speech source is spatially separated from the interfering noise rather than co-located (e.g., see [3]).

To increase discriminability between two competing sound sources, an elegant approach is to artifically increase the deviation of the competing sound source from the midline by frequency scaling of the space filters or the head-related transfer functions [4]. The rationale is that this processing would ultimately enhance one's ability to use auditory spatial cues in psychophysical tasks and in understanding speech in a noisy environment. This processing algorithm is referred to as the *interaural magnification* (IM) approach [4]. A variant of this approach, has been previously applied to the spectra of the signals received by the two ears instead of the spatial filters [5]. The authors demonstrated a significant increase in binaural masking level difference particularly in listeners with hearing impairments. Other similar studies in human auditory perception, have shown that listeners can eventually adapt to such unnatural (altered or re-mapped) auditory spatial cues, which can, in theory, provide better than normal localization ability (e.g., see [6], [7]).

In this paper, we propose an interaural magnification algorithm for speech enhancement in noise and reverberation. Motivated by the method[1] originally proposed in [4], we examine the effect of interaural magnification on speech intelligibility and spatial release from masking in a listening to speech-in-noise task. The complex acoustic mixture perceived binaurally is processed by magnifying the interaural level difference cues corresponding to the interfering sound source. This leads to a lateral spreading of the interfering source, which ultimately increases speech perception. Additionally, the performance of the proposed algorithm for noisy and reverberant speech enhancement is assessed through an articulation-based intelligibility model and the perceptual evaluation of speech quality metric.

[1]Theoretically, this interaural magnification procedure is equivalent to artificially enlarging the diameter of the listener's head. Such an enlarged head would in principle magnify both naturally-occurring interaural amplitude differences and interaural time differences [8].

http://dx.doi.org/10.21437/Interspeech.2016-1049

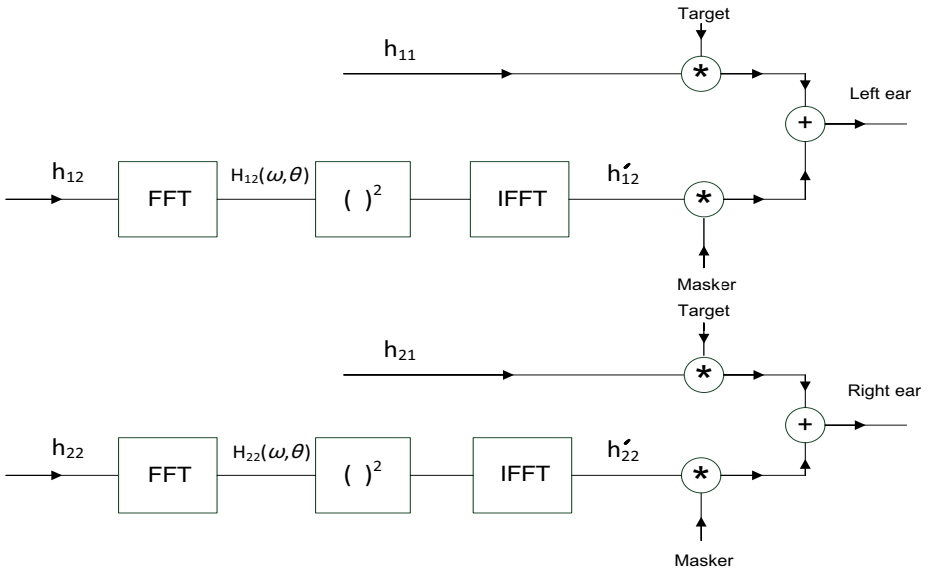

Figure 1: Block diagram of the proposed interaural magnification algorithm.

2. Algorithm Formulation

In this section, we analyze the interaural magnification algorithm, which can enhance the interaural level differences in a two input-signal and two-output system configuration shown in Fig. 1. First, the mixture signals perceived in each ear (binaural inputs) are generated by convolving the target source and the interferer with the binaural direction-dependent HRTFs corresponding to a particular azimuth location. The frequency-domain inputs $Y_L(\omega,\theta)$ and $Y_R(\omega,\theta)$ representing the left ear and right ear signal, respectively, at frequency ω and azimuth angle θ, can be written as follows:

$$Y_L(\omega,\theta) = H_{11}(\omega,\theta)S(\omega,\theta) + H_{12}(\omega,\theta)N(\omega,\theta) \quad (1)$$

$$Y_R(\omega,\theta) = H_{21}(\omega,\theta)S(\omega,\theta) + H_{22}(\omega,\theta)N(\omega,\theta) \quad (2)$$

where $S(\omega,\theta)$ denotes the source signal and $N(\omega,\theta)$ is the interferer. Furthermore, $H_{11}(\omega,\theta)$, $H_{12}(\omega,\theta)$, $H_{21}(\omega,\theta)$ and $H_{22}(\omega,\theta)$ denote four linear-time invariant filters corresponding to the HRTFs in our experiment. Note that in (1)–(2) the convolution operations are transformed into efficient multiplication operations by setting the frame size of the fast Fourier transform (FFT) to be much longer than the filter length. Focusing on the interfering source, we define the ratio $H = H_2/H_1$, also referred to as the *interaural transfer function* (ITF), which is equal to [9]:

$$\mathrm{ITF}(\omega,\theta) = \frac{H_{22}(\omega,\theta)}{H_{12}(\omega,\theta)} \quad (3)$$

The interaural level difference can be extracted from the ITF as follows:

$$\mathrm{ILD}(\omega,\theta) = 20 \cdot \log_{10}(|\mathrm{ITF}(\omega,\theta)|) \quad (4)$$

As depicted in Fig. 1, the filters h_{ij} corresponding to the impulse responses between the jth source and the ith ear are first converted to the frequency-domain. Secondly, the HRTFs describing the acoustic transfer functions are magnified by power of n and are then converted back to the time-domain. Finally, the enhanced outputs are calculated by the convolution of the target signal and the interferer signal with the processed or interaurally-magnified HRTFs. The magnified HRTFs are estimated according to:

$$H'_{12}(\omega,\theta) = [H_{12}(\omega,\theta)]^n \quad (5)$$

$$H'_{22}(\omega,\theta) = [H_{22}(\omega,\theta)]^n \quad (6)$$

where $H'_{12}(\omega,\theta)$ and $H'_{22}(\omega,\theta)$ are the modified HRTFs in the frequency-domain at frequency ω and azimuth angle θ and exponent n denotes the magnification power, which is equal to two in this paper. The processed (enhanced) outputs can be estimated by the interaurally-magnified HRTFs as follows:

$$Y'_L(\omega,\theta) = H_{11}(\omega,\theta)S(\omega,\theta) + H'_{12}(\omega,\theta)N(\omega,\theta) \quad (7)$$

$$Y'_R(\omega,\theta) = H_{21}(\omega,\theta)S(\omega,\theta) + H'_{22}(\omega,\theta)N(\omega,\theta) \quad (8)$$

where $Y'_L(\omega,\theta)$ and $Y'_R(\omega,\theta)$ are the modified signals for the left and right ear, respectively. The modified interaural transfer function $\mathrm{ITF}'(\omega,\theta)$ and the modified interaural level difference $\mathrm{ILD}'(\omega,\theta)$ are defined as:

$$\mathrm{ITF}'(\omega,\theta) = \left[\frac{H'_{22}(\omega,\theta)}{H'_{12}(\omega,\theta)}\right] = \left[\frac{H_{22}(\omega,\theta)}{H_{12}(\omega,\theta)}\right]^n \quad (9)$$

$$\mathrm{ILD}'(\omega,\theta) = 20 \cdot n \cdot \log_{10}(|\mathrm{ITF}(\omega,\theta)|) \quad (10)$$

According to Eq. (10), the interaural level difference corresponding to the noise source, is multiplied by a factor of n , which is expected to increase the lateral spreading of the interfering source and improve the overall benefit due to spatial release from masking.

3. Experimental Results

3.1. Stimuli

The performance of the proposed interaural magnification algorithm was evaluated on a test set of 10 speech signals comprised of a single randomly selected male-spoken sentence. A

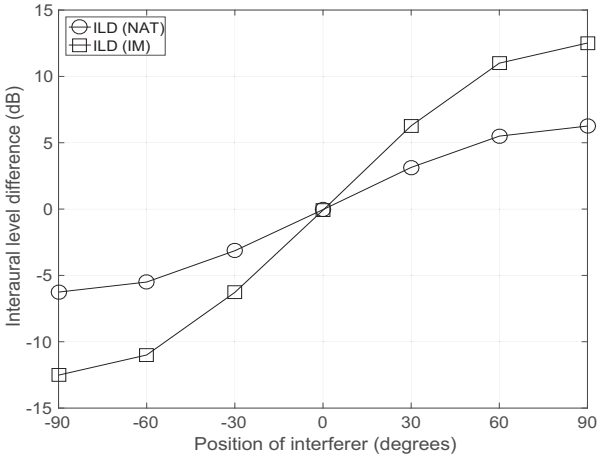

Figure 2: Natural and interaurally magnified ILDs in the anechoic scenario.

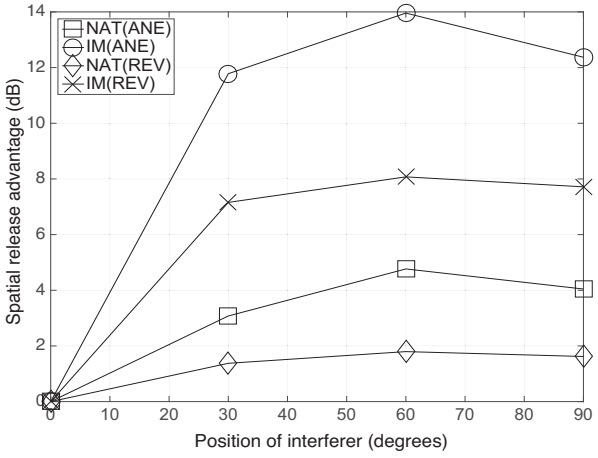

Figure 3: Modeled spatial release from masking (SRM) advantage for noise placed in the frontal hemifield after interaural magnification in an anechoic and a reverberant setting.

female interferer was used with a root-mean-square value equal to the target source, such that the input SNR = 0 dB. The duration of each speech signal was approximately 3 s. All signals were recorded at a sampling rate of 22,050 Hz. To generate the speech test stimuli, we used the IEEE database, which consists of phonetically balanced sentences, with each sentence being composed of approximately 7 to 12 words [10]. All signals had the same onset and were normalized to their maximum amplitude before convolving with the HRTFs.

Anechoic head-related impulse responses were used to simulate a non-reverberant listening condition. To simulate a more realistic scenario, a second set of reverberant head-related impulse responses were measured inside a typical office with reverberation time equal to RT_{60} = 300 ms, which is a typical value for a moderately reverberant environment. Both sets of impulse response measurements were conducted in the University of Oldenburg (e.g., see [9]). For each listening scenario, a total of four sound source locations were calculated for sound sources located 1 m away from the center of the listener in the azimuthal plane for every angle from 0^0 (i.e., straight ahead) to $+90^0$ to the right of the listener in 30^0 increments. In all cases, the target source was placed directly in front of the listener at 0^0, such that the ITF corresponding to the source is 1.

3.2. Spatial release from masking benefit

The spatial release from masking benefit facilitates the suppression of competing sounds in a noisy environment, based on contributions from two specific mechanisms: (1) better-ear listening or head-shadow and (2) binaural unmasking or binaural squelch, which rely on interaural level and time differences. Target and interferers at different locations produce different ILDs, such that one ear (contralateral) has always a better SNR than the other (ipsilateral), and therefore listeners can attend to the ear offering the better SNR. Furthermore, differences in the ITDs generated by the target and interfering source facilitate binaural unmasking, in which the auditory system is able to squelch to some extent the noise source.

The proposed interaural magnification algorithm was validated in both the simulated anechoic (ANE) and reverberant (REV) conditions with the binaural model described in [11, 12]. This is an articulation-based intelligibility model, which can predict better-ear listening and binaural unmasking effects contributing to overall SRM, by relying on prior information of the exact spatial location through access to binaural HRTFs. The model was run with naturally-occurring (NAT) level cues and also with ILDs subjected to interaural magnification (IM) as described in (10). These IM-processed ILDs are plotted in Fig. 2 for different spatial locations in the anechoic scenario (-90^0 to $+90^0$ in 30^0 steps). The theoretical values obtained for the SRM benefit are plotted as a function of the azimuth in Fig. 3. In all of the spatial locations (except for the 0^0 azimuth) and for both room conditions tested (ANE and REV), the model predicted a substantial benefit, likely stemming from the amplification of the level differences attained between the two ears.

3.3. Perceptual evaluation of speech quality

The overall quality of the enhanced output binaural signals described in (7)–(8) was also assessed with the perceptual evaluation of speech quality (PESQ) score [13]. The PESQ employs a sensory model to compare the original (unprocessed) with the enhanced (processed) signal, which is the output of the IM algorithm, by relying on a perceptual model of the human auditory system. The PESQ score has been shown to exhibit a high correlation coefficient (Pearson's correlation) of $r = 0.91$ with subjective listening quality tests [14].

The PESQ measures the subjective assessment quality of the dereverberated speech rated as a value between 1 and 5 according to the five grade *mean opinion score* (MOS) scale. Here, we use the PESQ measure with parameters optimized towards assessing overall speech signal distortion, calculated as a linear combination of the average disturbance value $D_{\rm ind}$ and the average asymmetrical disturbance values $A_{\rm ind}$ [13, 14]

$$\text{PESQ} \;=\; a_0 + a_1\, D_{\rm ind} + a_2\, A_{\rm ind} \qquad (11)$$

such that

$$a_0 = 4.5,\;\; a_1 = -0.1 \;\;\text{and}\;\; a_2 = -0.0309 \qquad (12)$$

By definition, a high value of PESQ indicates low speech signal distortion, whereas a low value suggests high distortion with considerable degradation present. The PESQ score is presumed to be inversely proportional to the amount of masking and is expected to increase as spatial release from masking increases (e.g., see [11]).

PESQ (anechoic)	0°	+30°	+60°	+90°
Left ear input	2.84 (±0.21)	3.11 (±0.17)	3.28 (±0.16)	3.17 (±0.17)
Right ear input	2.84 (±0.21)	2.60 (±0.15)	2.49 (±0.30)	2.62 (±0.17)
Left ear output	2.83 (±0.19)	4.23 (±0.09)	4.44 (±0.06)	4.30 (±0.06)
Right ear output	2.83 (±0.19)	3.39 (±0.11)	3.33 (±0.12)	3.49 (±0.09)
PESQ (reverberant)	0°	+30°	+60°	+90°
Left ear input	2.98 (±0.17)	3.12 (±0.16)	3.23 (±0.16)	3.15 (±0.18)
Right ear input	2.88 (±0.18)	2.79 (±0.19)	2.63 (±0.27)	2.69 (±0.27)
Left ear output	2.98 (±0.17)	4.03 (±0.08)	4.16 (±0.06)	4.05 (±0.08)
Right ear output	2.88 (±0.18)	3.42 (±0.10)	3.36 (±0.12)	3.37 (±0.12)

Table 1: PESQ input and output values for each azimuth location averaged over 10 IEEE sentences. The standard errors of the mean are inside the parentheses.

3.4. Discussion

Table 1 compares the performance of the proposed algorithm in terms of PESQ, relative to the performance of the unprocessed binaural inputs for each separate ear. Note that the PESQ metric, is a fairly reliable predictor of speech quality and is known to have the highest correlation with subjective measurements. In terms of overall speech quality and speech distortion, the score for the anechoic (unprocessed) sound source when this is co-located with the masker, averaged across 10 different sentences is equal to 2.84 (left and right), which suggests that a relatively high amount of degradation is present. In contrast, after processing the binaural signals with the proposed IM algorithm, the average scores in the left ear increase to 4.23, 4.44 and 4.30 for azimuths of 30^0, 60^0 and 90^0, respectively. In the reverberant conditions, after processing the binaural signals with the IM algorithm, the average scores in the left ear increase to 4.03, 4.16 and 4.05 for spatial locations corresponding to 30^0, 60^0 and 90^0, respectively. Note than in most of the experimental conditions, the standard deviation for the PESQ results ranges between 0.06 to 0.30. The estimated PESQ scores in both the anechoic and reverberant scenarios, suggest that the proposed algorithm improves the speech quality of the signals considerably, while keeping signal distortion to a minimum.

4. Conclusions

In this study, we have developed and tested an interaural magnification algorithm that can be used for binaural speech enhancement in noise and reverberation. The proposed algorithm operates by magnifying the interaural level differences corresponding to a spatially separated interfering sound source. Experiments carried out with speech signals masked by a single interfering source in both anechoic and reverberant scenarios indicate that the proposed technique is capable of: (1) increasing the spatial release from masking benefit and thus improve the suppression of competing sounds in a noisy environment and (2) improving the speech quality of the signals considerably, while keeping signal distortion to a minimum. A limitation of the proposed technique is that we assume prior knowledge of the head related transfer functions, which listeners use to understand and localize incoming sounds. Thus, for a practical implementation, we would need to pre-measure personalized HRTFs.

5. 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] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Ann. Rev. Psych.*, vol. 42, pp. 135–159, 1991.

[3] H. Glyde, J. Buchholz, H. Dillon, S. Cameron and L. Hickson, "The importance of interaural time differences and level differences in spatial release from masking," *J. Acoust. Soc. Am*, vol. 134, EL147–EL152, 2013.

[4] N. I. Durlach and X. D. Pang, "Interaural magnification," *J. Acoust. Soc. Am*, vol. 80, pp. 1849–1850, 1986.

[5] B. Kollmeier and J. Peissig, "Speech intelligibility enhancement by interaural magnification," *Acta Otolaryngol. Suppl.*, vol. 469, pp. 215–223, 1990.

[6] N. I. Durlach, B. G. Shinn-Cunningham and R. M. Held, "Supernormal auditory localization. I. General background," *Presence*, vol. 2, pp. 89–103, 1993.

[7] B. G. Shinn-Cunningham, N. I. Durlach and R. M. Held, "Adapting to supernormal auditory localization cues I. Bias and resolution," *J. Acoust. Soc. Am*, vol. 103, no. 6, pp. 3656–3666, 1998.

[8] G. F. Kuhn, "Acoustics and measurements pertaining to directional hearing," in *Directional Hearing*, W. A. Yost and G. Gourevitch, Eds. Berlin, Germany: Springer-Verlag, pp. 325, 1987.

[9] N. Kayser, S. D. Ewert, J. Anemuller, T. Rohdenburg, V. Hohhmann and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. Appl. Signal Process.*, pp. 1–10, 2009.

[10] IEEE Subcommittee, "IEEE recommended practice speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[11] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," *J. Acoust. Soc. Am*, vol. 127, pp. 387–399, 2010.

[12] S. Jelfs, J. F. Culling, and M. Lavandier, "Revision and validation of binaural model for speech intelligibility in noise," *Hear. Res.*, vol. 275, pp. 96–104, 2011.

[13] ITU-T P.862, "Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders," ITU-T, 2001.

[14] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.,*, vol. 16, no. 1, pp. 229–238, 2008.