Speech intelligibility enhancement through binaural signal processing

Jorge Mejia[1,2], Harvey Dillon[1,2] and Simon Carlile[3]

1 - National Acoustic Laboratories Sydney, Australia

2 - The Hearing CRC, Australia

3 - Auditory Neuroscience Laboratory, the University of Sydney, Australia

Spatial separation of target speech from distracting sounds greatly assists the listener to segregate the sounds, and hence better understand the target speech. Consequently, listening can occur in poorer signal-to-noise ratios (SNRs). Bilateral beamformers, which combine microphone output signals from both sides of the head, can improve SNR, but in the process remove interaural difference cues, and hence remove the ability to segregate the target from distracting sounds on the basis of spatial separation. This spatial cue removal decreases the speech intelligibility benefits provided by the beamformer. Some techniques aim to retain the spatial cues in a beamformer output but in the process constrain its directional efficiency. An alternative technique proposed by Mejia *et al*., [WIPO Pub. No: WO/2007/137364 (2007)] exploits the perceptual suppression of early reflections (known as the precedence effect) through combination of omni-directional precedent sounds with highly directional processed sound. The enhancement produces intelligibility scores much higher than those produced by bilateral beamformer outputs in the absence of precedent sounds. This paper will describe the strategy of spatial enhancement and discuss the outcome from a subjective study intended to evaluate the technology.

INTRODUCTION

Bilateral beamformers operate by weighting and linearly combining microphone output signals from the left and right sides of the head to produce a super-directional output. Known strategies include fixed array weights, e.g. manifold vector rotation or alternatively adaptive weights, e.g. least mean square error (see Brandstein *et al*., 2001). However, these techniques remove the localization cues e.g. interaural time delay and level differences, naturally available to the both ears of listeners. In order to provide bilateral beamformer processing benefits while preserving the localization cues, Greenberg *et al*. (1992) examined a combination of bilateral beamformer processing applied to high frequency signals, and bilateral amplification (e.g. independent processing on each side of the head) applied to low frequency signals. The aim was to provide an signal-to-noise output benefit to high frequency sounds and to retain the localization cues to low frequency sounds. It was assumed

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation, 2010.

that the localization cues provide a greater benefit to low frequencies sounds than high frequency sounds. However when the scheme was evaluated under a relatively challenging listening condition, which included mild reverberation and a single competing sound, the algorithm appeared to retain the localization cues with an accuracy of 70% while providing a modest signal-to-noise benefit, not exceeding 3 dB over conventional bilateral amplification.

The number of microphones available in head-wearable bilateral beamformer systems is limited mostly due to aesthetics in the relatively small surface area available around the head, e.g. ears, headbands or spectacles (eyeglasses). Most systems based on Behind-the-Ear devices are limited to typically two, or at most, three microphones on either side of the head. As a result, the benefits reported in the literature from different beamformer techniques may relate to the conditions under which the different beamformer algorithms are examined, e.g. reverberation, number of sounds present in the environment and spatial separation between sound sources. Despite different outcomes, single output bilateral beamfomer techniques appear to provide the largest signal-to-noise benefit reported in the literature. Thus, it is worth asking if it is possible to reassert the localization cues to listeners while retaining the signal-to-noise benefit provided by the beamformer outputs.

A mechanism to provide spatial separation without significantly altering the signal-to-noise available to listeners in free field listening situations has been recently studied by Freyman *et al*. (1989) and others. The idea relates to the perceptual suppression of identical and successive sound presentations. When two identical sounds are presented from two different spatial locations most listeners report a fused sound image medial to the spatial placement of the sound sources. However if one sound is delayed relative to the second sound most listeners report a sound image in the location of the leading sound source. Thus applying localization dominance by leading sounds, Mejia *et al*. (2007) proposed a novel signal processing scheme aiming to reassert the localization cues in bilateral beamformer processing. The following presents a summary of this scheme with a subjective study examining its performance in real complex listening situations.

BINAURAL PROCESSING

Spatial enhancement technique

The novel strategy shown in Fig. 1 comprises of two signal paths. The first signal path performs an unconstrained combination of microphone output signals to produce a single super-directional signal, denoted as x. Algorithms to achieve this are readily available in the literature (e.g. Cardoso *et al*., 1989 and Greenberg *et al*., 1992). This super-directional signal is delayed by more than one millisecond but by less than ten milliseconds, to produce a super-directional delayed signal. The second signal path weights the left and right signals independently to produce subsidiary left and right

signals, denoted as y_L and y_R accordantly. The weights applied to the left and right signal paths are identical and approximately equals to the ratio between the noise present in the bilateral beamfomer output signal to the noise present in the subsidiary signals. Subsequently, the subsidiary left and right signals are combined with the super-directional delayed signal to produce a binaural output signal, denoted as O_L and O_R.

Principle of operation

The scheme operates by assuming that the noise or off-axis distracting sounds present in the subsidiary signal paths are identical to the sounds leaking into in the super-directional delayed signal path. In effect, the output presented to listeners consists of identical successive presentations of distracting noise sounds, with leading sounds having natural localization cues available to both ears of listeners. Thus, due to the precedence effect (i.e. perceptual suppression successive sound presentation) listeners perceive the natural reassertion of localization cues. Arguably, the subsidiary signal path reintroduces noise in the system adversely affecting the output SNR available in the super-directional delayed signal. Fortunately, the masking effect between two identical and successive sound presentations occurs even when the level of the leading sound is below the level of the corresponding sounds in the delayed super-directional signal, thus minimally altering the SNR available at the binaural output.

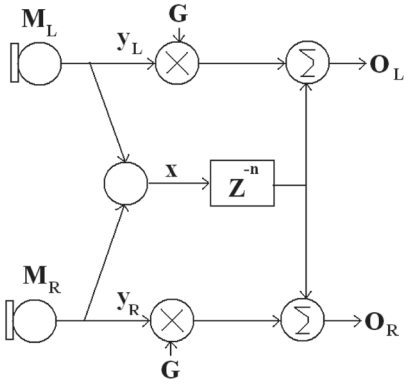

Fig. 1: Spatial enhancement processing based on microphones located on each side of the head. (From Mejia *et al*., 2007.)

VALIDATION METHODS

The experimental validation examined here is based on speech intelligibility measures as well as spatial quality of sounds relative to bilateral amplification (conventional directional microphones operating on each side of the head independently).

Subjects

Eleven normal hearing listeners and eight hearing-impaired listeners with mild sensorineural hearing losses participated in this experiment. The corresponding hearing loss profiles averaged within groups are shown in Fig. 2. The subjects ranged from 22 to 64 years of age. The subjects were required to attend one session lasting approximately two hours and they were paid a small contribution for their attendance to cover travelling cost.

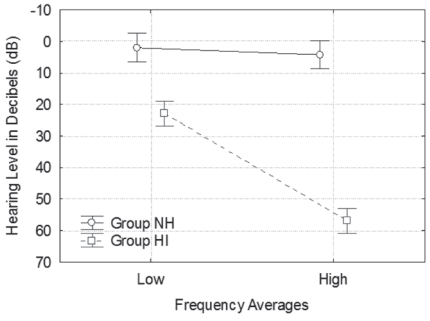

Fig. 2: Hearing loss profiles for normal and hearing-impaired groups. The data is shown as 250, 500 and 1000 Hz low-frequency averages, and 2000, 3000 and 4000 Hz high-frequency averages.

Signal processing conditions

Three signal-processing strategies were examined. The first was referred to as bilateral amplification, which was based on directional cardioid responses using two microphones located on each side of the head. The second was referred to as a bilateral beamformer, which was based on the combination of the cardioids responses from each side of the head to produce one super-directional output. The thirds was referred to as spatially enhanced processing, which was based on a combination of delayed, by three milliseconds, bilateral beamformer output signals with left and right omni-directional preceding signals.

Implementation

Input wave files, sampled at 22050 samples-per-second, were time sliced into approximately 12 ms frames and transformed into the frequency space using a Discrete Fourier Transform (DFT) which comprises of 2^9 taps in length. The DFT strategy was based on a well-known delay-and-add method with 50% overlap, and using a hamming widow as weighting function. All directional processing was calculated in the DFT domain and subsequently transformed to the time domain by an inverse DFT computation. Finally, outputs were presented to listeners over HD251 headphones. The sound levels were also frequency compensated according to NAL-RP (see Dillon, 2001).

Intelligibility assessment

The intelligibility assessment task selected corresponded to a version of the Coordinate Response Measure (CRM) corpus (Bolia *et al*., 2000). The CRM speech corpus consist of a call name, a colour and a number token, all embedded within a carrier phrase. A typical example is "Ready Baron go to Blue Five now", where Baron is the call name, and Blue and Five are the colour and number tokens needed to be identified by the listener. In this experiment the call name for the target remained constant, whereas the colour and number tokens were randomly assigned. On the other hand, the call name, the colour and number tokens were randomly assigned for the distracting sounds. Recordings from a target female talker and two different female talkers were chosen for the disctractor sound sources. The recordings were performed in a relatively quiet room, covered with carpet material on walls and flooring. The target speech sound was presented over a loudspeaker located at 0° azimuth location relative to the orientation of the KEMAR head. The distracting speech sounds were presented over loudspeakers located at 90° and 270° relative to the orientation of the KEMAR head. All three loudspeakers were 1.2 m from the centre of KEMAR. This exceeds the critical distance in the room. The sound presentations were recorded with two microphones mounted on behind-the-ear (BTE) hearing aid devices and located on each side of the head. During the experiment, the input SNR (dB) was adjusted until the 72% correct responses were obtained.

Preference assessment task

Listeners were asked to choose, in a forced-choice task, whether the bilateral amplification scheme or the spatially enhanced processing had the better spatial quality of sounds. However due to the inherently superior SNR produced by the spatially enhanced scheme (due to the bilateral beamformer output) its output signal-to-noise was adjusted for every listener to match the SNR produced by the bilateral amplification strategy. In an AB-comparison, listeners were asked to assess the ease of listening, spatial naturalness and provide an overall preference rating.

RESULTS

Intelligibility assessment

The Speech Reception Thresholds (SRT) scores at 72% correct responses obtained for the different processing schemes, and different groups are shown in Fig. 3. The analysis of variance indicates a statistically significant difference between the amplification conditions (ANOVA, $p < 0.001$). The figure indicates that the normal-hearing group benefited from the bilateral beamformer processing by 5 dB SNR over the bilateral amplification. Furthermore the spatially enhanced strategy produced a further 4 dB SNR mean improvement in SRT scores. This improvement was significant (Scheffe's post-hoc, $p < 0.05$). On the other hand, the bilateral beamformer benefited the hearing-

impaired listeners by more than 10 dB SNR over bilateral amplification. However, the intelligibility scores were not significantly different between bilateral beamformer and spatially enhanced processing (Scheffe's post-hoc, $p < 0.05$).

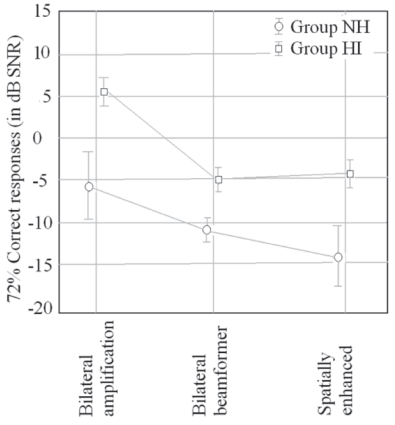

Fig. 3: SRT scores for the four different processing schemes examined.

Preference assessment

The accumulated preference rates in the matched output SNR conditions are shown in Fig. 4. The normal hearing group provided no absolute preference for either processing strategy. However, the scores suggest that the ease of listening and naturalness were significantly different between conditions (ANOVA, $p < 0.001$). On the other hand, the hearing-impaired group provided no preference for any of the processing strategies (ANOVA, $p > 0.05$).

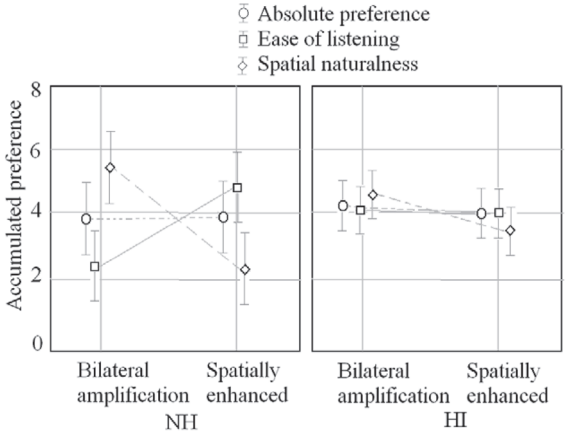

Fig. 4: Accumulated preference following eight successive comparisons, where the larger numbers indicate greater preference.

DISCUSSION

The experiment demonstrated that the bilateral beamformer assisted normal-hearing listeners in poorer listening situations, exceeding 5 dB SNR improvements over bilateral amplification (i.e. conventional directional microphones). Despite the fact the spatial enhancement was perceived as unnatural to normal-hearing listeners, the reassertion of localization cues significantly assisted them in the speech discrimination task, with a further 4 dB SNR improvements. As a result, normal-hearing listeners were able to understand speech at SNR levels, 9 dB poorer than with bilateral amplification. On the other hand, the hearing-impaired listeners received 10dB SNR benefits on the bilateral beamformer over bilateral amplification. This enabled the hearing-impaired listeners to perform as well as normal-hearing listeners using bilateral amplification. However the spatial enhancement did not further assist them in the intelligibility task. These finding are consistent with other studies reporting poorer spatial hearing benefits by hearing-impaired listeners (e.g. Arbogast *et al*., 2002). Finally, the hearing-impaired listeners judged the spatially enhanced scheme to be equal to the bilateral amplification scheme on all criteria, despite the SNR of the input signal being degraded by an average of 10 dB SNR for the spatially enhanced scheme.

CONCLUSIONS

This study had shown that bilateral beamfomer processing provided large SNR benefits to normal and hearing-impaired listeners. The novel reassertion of the localization cues proven not as natural as bilateral amplification, but was remarkably effective to normal-hearing listeners, further improving their speech intelligibility scores in a complex listening situation. On the other hand, hearing-impaired listeners were not able to discriminate between processing schemes when the output SNR was the same, yet the SNR advantage was not further degraded by the spatially enhanced scheme. In summary, the low computational complexity combined with the large SNR advantage, and other spatial hearing benefits significantly improve the benefit-to-cost ratio, hence rendering binaural signal processing more technically feasible and commercially marketable in head wearable devices, such as hearing aids.

REFERENCES

Arbogast, T., Mason, C., and Kidd, G. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

Brandstein, M., and Ward D. (**2001**). *Microphone Arrays: Signal Processing Techniques* (Springer Verlag).

Bolia, R., Nelson, W., Ericson, M., and Simpson, B (**2001**). "Speech Corpus for Multitalker Communication Research." J. Acoust. Soc. Am. **107**, 1068-1066.

Cardoso, J. F. (**1989**). "Source separation using higher order moments." In Proc. ICASSP, 2109-2112.

Dillon, H. (**2001**). *Hearing Aids* (Boomerang Press, Sydney).

Freyman, R. L., Litovsky, R. Y., Balakrishnan, U., and Clifton, R. K. (**1989**). "Buildup and breakdown of the precedence effect," J . Acoust. Soc. Am. Suppl. **1**, 85, S83.

Greenberg, J. E., and Zurek, P. M. (**1992**). "Evaluation of an adaptive beamforming method for hearing aids," J. Acoust. Soc. Am. **91**, 1662–1676.

Mejia, J., Dillon, H., and Carlile S. (**2007**). "A method and system for enhancing the intelligibility of sounds," WIPO Pub. No: WO/2007/137364.

Wallach, H., Newman, E. B., and Rosenzweig M. R. (**1949**). "The precedence effect in sound localization," Am J Psychol. **62**, 315-336.