AUDITORY ATTENTION BASED ON DIFFERENCES IN MEDIAN VERTICAL PLANE POSITION

J.W. Worley & C.J. Darwin

Experimental Psychology, University of Sussex
Brighton BN1 9QG, UK
johnwo@biols.susx.ac.uk

ABSTRACT

The first experiment described here asks whether listeners are able to selectively attend to one of two sentences differing in median vertical plane (MVP) location using a paradigm developed for azimuthal attention [1]. It also asks whether their ability to use MVP cues improves with a difference in fundamental frequency (Fo) between the two sentences. Listeners attend to one of two simultaneous same-talker utterances and report which of two target words ("speech" or "phrase") occur in the attended sentence. The sentences are played from matched loudspeakers in an anechoic chamber from different MVP positions. When the sentences are both played on the same constant Fo, listeners report the target word almost perfectly with a 31° vertical separation. For smaller separations, performance is worse but improves with increasing difference in Fo between the sentences. This improvement is not due to the Fo difference improving listeners' ability to use MVP cues, but rather to their using continuity of Fo difference to track the target sentence. When the sentences are low-pass filtered, listeners are less able to use MVP cues and so there is a greater relative use of Fo continuity. The second experiment showed that both the ability to selectively attend to one of two sentences and the ability to localise a single sentence were worse with headphone presentation after convolution with individually-optimised library HRTFs than with free-field presentation. However, the low-pass filtering at 5 kHz gave relatively little additional degradation in either task.

1. INTRODUCTION

Although it is well-established that listeners can attend selectively to speech sound sources that come from different azimuthal positions, even when their location is cued simply by interaural time differences [1], it is less clear whether listeners can also attend selectively to speech sounds that come from a given position in the median vertical plane (MVP). If such attention is possible, then an interesting question is raised as to how listeners are able to associate the high-frequency spectral cues to MVP location with the appropriate low frequencies which give the content of the speech.

Tracking sound sources which differ only in MVP position is likely to be more difficult than tracking sound sources that differ in azimuth. Good and Gilkey [2] had listeners localise a click train at 239 possible locations in the presence of a distractor noise centred straight ahead. As the signal-to-noise ratio (SNR) was decreased location accuracy for sources lying on the horizontal plane decreased monotonically. However, the location accuracy for a source that differed in MVP location was more strongly influenced by the noise.

One way in which listeners might group together the high-frequency spectral region which is the main source of MVP cues with the lower frequency content-bearing region is to use a common fundamental frequency. Just as it is possible for listeners to group together the higher formants with the first formant across resolved and unresolved harmonics by using a common Fo [3], so it might be possible to group together using a common Fo the frequency region below 3kHz with that above it for the purposes of assigning the appropriate MVP locations to simultaneous speech utterances.

2. EXPERIMENT 1

This experiment asks two questions: first, can listeners use a common MVP location to follow a particular utterance; second, is this ability improved by having a difference in Fo between the utterances.

In order to investigate listeners' ability to follow a sound source rather than to investigate intelligibility, we use a very simple task with maximally-predictable materials. Listeners have to follow one of two sentences which are played on every trial, and to identify which of two possible target words occurred in the attended sentence. In this experiment they can solve this task using MVP location.

We are also interested in whether differences in Fo between the sentences improve listeners' ability to use the MVP cue. Consequently, we vary systematically the Fo difference between the two sentences. However, since we know that listeners can use a constant difference in Fo to help them perform the tracking task [1], we introduce a "switched" condition in which the Fo of the two target words is switched. In these switched trials there is always a difference in Fo between the two sentences and target words, so that any advantage that this Fo difference gives to establishing MVP location would remain, however, we can then remove the confounding effect of listeners tracking Fo by averaging their responses to the switched and unswitched conditions. If a difference in Fo improves listeners' ability to use MVP location in this task, then this average score should increase with the difference in Fo. A secondary reason for including the switched conditions, is that it allows us to compare the relative effectiveness of the MVP cue and the Fo difference as tracking cues.

2.1. Materials

Two sentences, "Could you please write the word speech down now" and "You will also hear the sound phrase this time," were spoken with a flat intonation contour by a native British English speaker (CJD). The recordings were made in a sound proof booth onto digital audiotape, then digitized at 44.1kHz. It was necessary to align the two target words ('speech' and 'phrase') within the carrier sentences in order for their onset and offset to coincide exactly. The durations of the target words were equalized by adding and removing

pitch periods. About 40ms of silence was added to the beginning of the sentence "could you..." to align the two target word onsets across sentences. The target words started 1.25 s from the onset of the sentences.

The target word in the attended sentence was always coupled with the other target word in the distracter sentence. Therefore both sentences and both target words were heard on each trial.

The two sentences were resynthesized on a monotone by applying a pitch synchronous overlap algorithm (PSOLA) (Moulines and Charpentier, 1990) at Fo's of 115, 122, 126, 130, 133, 137, and 146Hz. To ensure correct target word alignment small adjustments were made to the silent interval

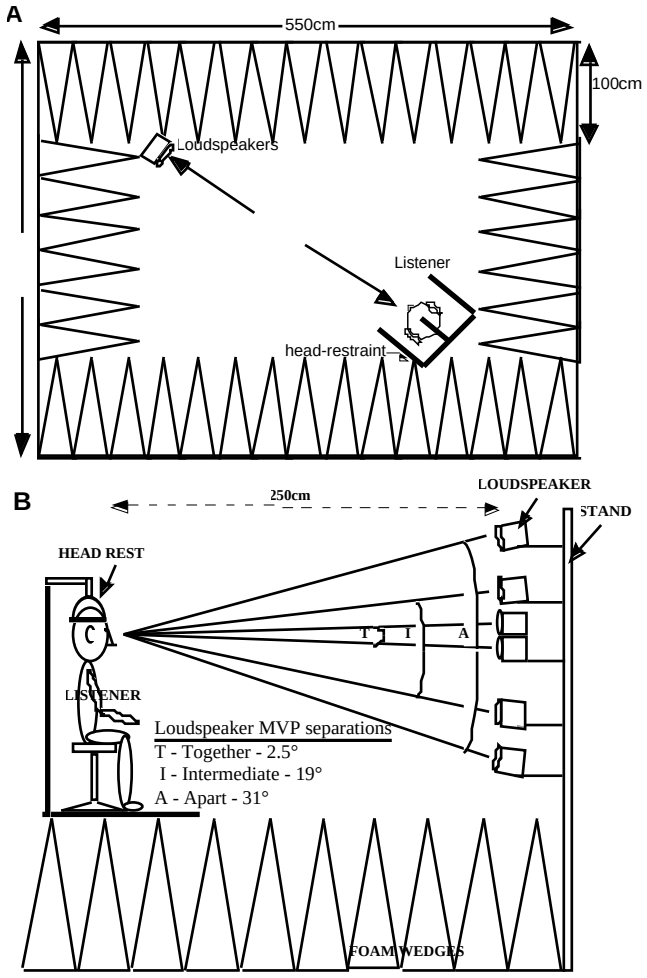

Figure.1 *A) Aerial view of anechoic chamber B) Lateral view of loudspeaker MVP separations*

before target word onset. This procedure compensated for duration changes produced by the PSOLA resynthesis.

Since the spectrum of speech falls off at approximately 6dB/octave, we pre-emphasis filtered the sentences at +6dB/octave in order to boost the high frequency region that provides the main MVP cues.

For the low-pass filtered condition the sentences were low-pass filtered with a 5 kHz cut-off, using a FIR filter (1001 samples @ 44.1kHz), with a 100Hz slope.

2.2. Procedure.

The 10 participants were native British English speakers between the ages of 20 and 35. All had pure-tone thresholds within the normal range for frequencies between 125 Hz and 8 kHz.

The participants were tested individually in a single-skinned anechoic chamber (Fig.1) at B&W Loudspeakers, Steyning, West Sussex. The listener's head was restrained throughout each block of trials to allows us to directly compare the results of these free-field experiments to later ones using virtual sources.

The sentences were presented over two loudspeakers (Drive units: Vifa M610 MD09-04) from an Apple Macintosh 7100 using an Audiomedia II soundcard, through an amplifier (Aura Evolution VA 10011s), which maintained an average sound level of 70dB (SPL) at the listening position. The frequency response of each loudspeaker was flat within ±3dB for the frequency range 0.2 to 18kHz and the loudspeakers were matched to each other within ±1dB across this frequency range. The MVP separation of the loudspeakers was achieved by means of an adjustable stand. The MVP separations were: Fully apart (31°) an intermediate position (19°) and together (2.5°) (Fig.1).

The listeners were told that they would hear two simultaneous sentences, and their task was to attend to the "could you please..." sentence and to indicate whether it contained the word “speech” or “phrase”. Responses were recorded by pressing the "s" or "p" keys on the keyboard. On each trial the listeners heard both target words and both carrier sentences.

The carrier sentences and target words differed in Fo by 0, 1, 2, or 4 semitones. The zero ΔFo condition paired sentences with an Fo of 130Hz. The one semitone ΔFo used 126 Hz with 133 Hz, two-semitone 122 Hz and 137 Hz and four semitone 115 Hz and 146 Hz.

The relation between the Fo of the target word and the Fo of the carrier sentence was either congruent (unswitched) or incongruent (switched). For the ‘unswitched’ conditions, the Fo of the target word was the same as the Fo of the attended carrier sentence with the same MVP position. For the ‘switched’ conditions, the Fo of the target word was the same as the Fo of the unattended carrier sentence, but the target word kept the same MVP position as the attended sentence.

Each sentence pairing was presented 5 times giving a total of 130 trials in each block. A block is defined as having a specific MVP separation (2.5°, 19° or 31°) and bandwidth (full vs. LP filtered). The presentation order of the sentence pairings was randomized within each block, and the presentation order of the blocks was counterbalanced. Prior to the start of the experiment, listeners were familiarized with the sentence that was to be attended to both in isolation and in the presence of the distracter sentence.

2.3. Results

2.3.1. Full bandwidth conditions

The percentage of trials on which listeners reported the target word with the same MVP location as the attended carrier sentence are shown in Fig 2. (The nominal zero separation corresponds to an actual separation of 2.5° so this scoring method is still defined for those conditions.)

At the widest separation listeners performed the task almost perfectly in the normal (unswitched) condition, across all values of Fo difference, indicating that they can use MVP location to track the appropriate target word. They show only a slight decrease in the switched condition with increasing difference in Fo, as continuity of Fo acts against the MVP cue. The small size of this decrease indicates the

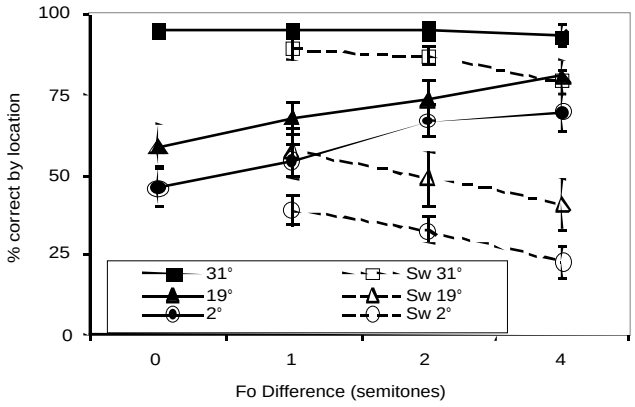

Figure 2 *Percent target words from the same location as the attended sentence as a function of the Fo difference between the two sentences, and their difference in elevation. In the switched conditions the target word had the Fo appropriate to the unattended sentence but still came from the MVP position of the attended sentence.*

relative strength of the 31° MVP cue. Because performance in the unswitched condition has asymptoted, the average of the unswitched and switched scores actually decreases with increasing Fo difference.

At 19° MVP separation, listeners' use of the difference in Fo is more apparent. With no difference in Fo, they perform marginally above chance on the basis of MVP location alone. As the difference in Fo increases, performance in the unswitched condition improves substantially while performance in the switched condition correspondingly decreases; both changes indicate the extent that listeners are tracking the target word by Fo. There is no evidence that the average of the switched and the unswitched conditions increases with increasing difference in Fo, and so no evidence that the difference in Fo is improving listeners' use of MVP cues.

At the 2.5° separation, listeners are (not surprisingly) unable to track the target word when there is no difference in Fo. Increasing the difference in Fo gives rise to the same pattern of change as found with 19° separation.

2.3.2. Low-pass filtered (5 kHz) conditions

When the experimental sounds were low-pass filtered at 5 kHz, the results shown in Figure 3 were obtained.

The overall pattern of results is similar to the broadband condition in figure 2, but with performance in the 31° and 19° separation conditions reduced in the unswitched conditions indicating weaker MVP location cues, together with a correspondingly greater relative contribution in the switched conditions from Fo continuity. As with the unfiltered conditions, there is no evidence that the average of the switched and the unswitched conditions increases with increasing difference in Fo.

2.4. Discussion

This experiment has demonstrated three things. First, listeners are able to attend to one of two sentences more easily when the two sentences come from different MVP locations than when they come from the same location. This ability is degraded when the sentences have been low-pass

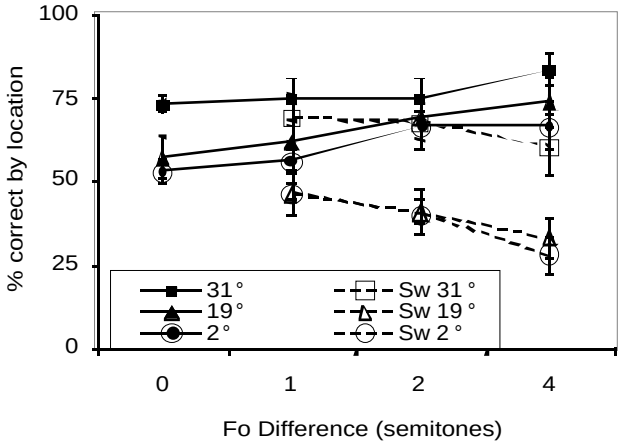

Figure 3 *As Figure 2 but for sentences that have been low-pass filtered at 5 kHz.*

filtered at 5 kHz. Second, listeners can also use a common (monotonous) Fo to attend to one of two simultaneous sentences. This result confirms those from similar experiments using azimuthal location rather than MVP [4]. Third, we have found no evidence that listeners' ability to use MVP is improved by a difference in Fo between the two sentences. This last result is seen in the lack of change of the average of the switched and unswitched conditions across differences in Fo.

The poorer performance with the low-pass filtering probably arises because of the reduced cues to MVP. Figure 4 shows data from another experiment in which 5 listeners heard just the target sentence coming from one of two loudspeakers separated by 31° and had to indicate which loudspeaker it came from. With the full spectrum performance approaches 95% correct, but is substantially poorer with low-pass filtering at 5 kHz.

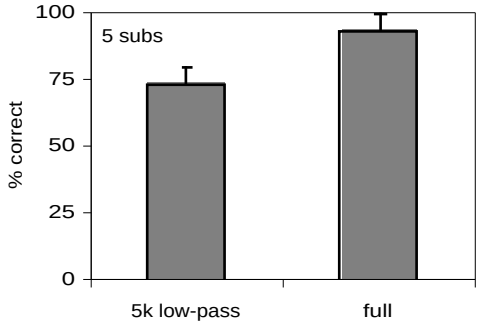

Figure 4 *Percent correct identification of which of two loudspeakers with a MVP separation of 31° a target sentence was played from.*

3. EXPERIMENT 2

This experiment is similar to the previous one, but uses headphone presentation after convolution of the sentences with individually-optimised library HRTFs. The aim of the experiment was to find whether attention to one of two

sentences, differing in virtual location was possible using HRTFs which were not those of the listener.

3.1. Method

This experiment used publically-available HRTFs (from the AUDIS project CD) individually chosen for each of 3 listeners to give the best MVP localization. The same sentences as used in the first experiment were convolved with HRTFs corresponding to MVP separations of 0°, 3°, 10°, 40°, and 80° played to listeners through a pair of Etymotic (ER2A) tubephones. The paired sentences were always played on the same Fo (either 115 or 146Hz) using either the full spectrum or low-pass filtered, as in the previous experiment. The sentences were played either with their full spectrum, or low-pass filtered at 5 kHz.

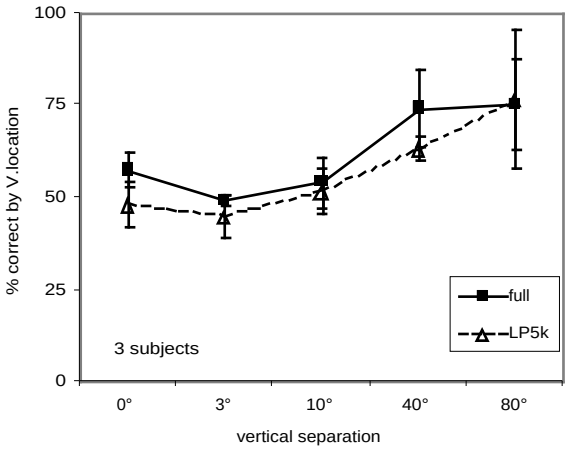

Figure 5. *Percent target words from the same location as the attended sentence as a function of the difference in HRTF MVP separation between the two sentences.*

3.2. Results

Performance in the sentence tracking task with sentences convolved with different MVP HRTFs is shown in Fig. 5 for the full spectrum and 5-kHz low-pass filtered conditions.

For both filtering conditions, tracking ability is worse for these 3 listeners using HRTFs than it was for the 11 listeners in the free-field conditions of experiment 1. Even with a virtual separation of 80° listeners in the full-spectrum condition do not get more than 75% correct. Low-pass filtering does reduce their performance, but only modestly. Again, this difference is substantially less than with the free-field presentation.

3.3. Discussion

Listeners find it considerably harder to track one sentence rather than another that differs in MVP HRTF than when the sentences are presented in the free field. This conclusion is true even though the HRTFs have been individually selected from a publically-available library. Moreover, low-pass filtering the sentences does not dramatically lower performance. The reason for these two effects may be that the low-frequency part of the spectrum [5] provides cues for speech separation that are more robust across individuals even though they allow poorer MVP localization than the higher frequency cues for individualized HRTFs.

The difference in tracking ability between free-field and virtual conditions is not surprising when considering the reports that sources synthesized with non-individualised HRTFs give poorer MVP location percept than sources synthesized with individualised HRTFs [6], a difference that is partly due to the difference in individual listeners' pinna size [7].

The effectiveness of the library HRTFs in MVP location was tested directly on a group of 13 listeners who had to judge which of three virtual locations (0°, 40° or 80°) a sentence had come from. The sentence was presented either with full spectrum or low-pass filtered at 5 kHz. The results are shown in Fig 6 and show that performance is rather poor (chance = 33%) for the full spectrum condition but is not made much poorer by low-pass filtering at 5 kHz.

These results support the idea that the low-frequency region of HRTFs provides some MVP location information which is weaker but less idiosyncratic than that provided by the high frequency region.

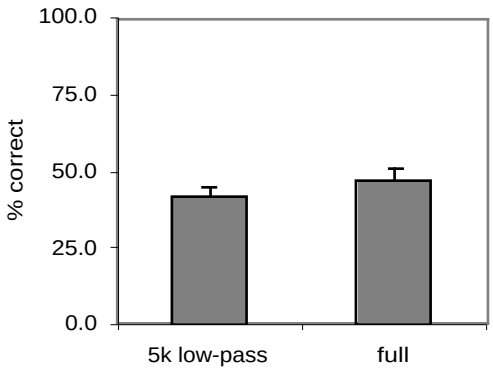

Figure 6 *Percent correct identification of which of three HRTF MVP separations (0°, 40° or 80°) a target sentence was played from.*

4. ACKNOWLEDGEMENTS

John Worley was supported by a BBSRC studentship. We gratefully acknowledge the help of B&W Loudspeakers Ltd, Steyning for access to their anechoic rooms, and to Rob Hukin for technical assistance.

5. REFERENCES

[1] C. J. Darwin and R. W. Hukin, "Auditory objects of attention: the role of interaural time-differences," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 25, pp. 617-629, 1999.

[2] M. D. Good and R. H. Gilkey, "Sound localization in noise: the effect of signal-to-noise ratio," *Journal of the Acoustical Society of America*, vol. 99, pp. 1108-17., 1996.

[3] J. Bird and C. J. Darwin, "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and physiological advances in hearing*, A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis, Eds. London: Whurr, 1998, pp. 263-269.

[4] C. J. Darwin and R. W. Hukin, "Effectiveness of spatial cues, prosody and talker characteristics in selective attention," *Journal of the Acoustical Society of America*, vol. 107, pp. 970-977, 2000.

[5] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *Journal of the Acoustical Society of America*, vol. 109, pp. 1110-22., 2001.

[6] M. Morimoto and Y. Ando, "On the simulation of sound localization," *Journal of the Acoustical Society of Japan*, vol. 1, pp. 167-174, 1980.

[7] J. C. Middlebrooks, "Individual differences in external ear transfer functions reduced by scaling in frequency," *Journal of the Acoustical Society of America*, vol. 106, pp. 1480-1492, 1999.