DETECTION AND LOCALIZATION OF SPEECH IN THE PRESENCE OF COMPETING SPEECH SIGNALS

Brian D. Simpson
Douglas S. Brungart
Nandini Iyer

Air Force Research Laboratory
2610 Seventh Street
WPAFB, OH 45433
brian.simpson@wpafb.af.mil
douglas.brungart@wpafb.af.mil
nandini.iyer@wpafb.af.mil

Robert H. Gilkey[1]
James T. Hamil[2]

[1]Wright State University
[2]General Dynamics
Dayton, OH
gilk@wright.edu
james.hamil@wpafb.af.mil

ABSTRACT

Auditory displays are often used to convey important information in complex operational environments. One problem with these displays is that potentially critical information can be corrupted or lost when multiple warning sounds are presented at the same time. In this experiment, we examined a listener's ability to detect and localize a target speech token in the presence of from 1 to 5 simultaneous competing speech tokens. Two conditions were examined: a condition in which all of the speech tokens were presented from the same location (the 'co-located' condition) and a condition in which the speech tokens were presented from different random locations (the 'spatially separated' condition). The results suggest that both detection and localization degrade as the number of competing sounds increases. However, the changes in detection performance were found to be surprisingly small and there appeared to be little or no benefit of spatial separation for detection. Localization, on the other hand, was found to degrade substantially and systematically as the number of competing speech tokens increased. Overall, these results suggest that listeners are able to extract substantial information from these speech tokens even when the target is presented with 5 competing simultaneous sounds.

1. INTRODUCTION

Auditory displays have been employed in a variety of applications, from simple alarms and warnings in automobiles to advanced virtual audio display technologies in aircraft cockpits. A common issue in the design of these displays is the tradeoff between the desire to present the listener with as much information as possible and the concern that the listener will be unable to process and interpret the auditory information if too many sounds are presented at the same time. This can be a particularly important issue in speech-based auditory displays that present information via prerecorded voice samples rather than more abstract sounds. This paper presents the results of an experiment that evaluated listeners' ability to detect and localize speech-based audio tokens in a display where multiple competing tokens are presented at the same time.

While many types of auditory displays could potentially be used to present multiple simultaneous warning sounds, we decided to focus initially on speech displays. Speech displays have the advantage that they are intuitive and thus can be understood with little or no training on the part of the operator. In addition, they lack the ambiguity that so often typifies many nonspeech auditory symbologies and they can be used to convey almost any kind of information. However, there are a number of potential disadvantages to using speech as the basis for an auditory display. First, speech intelligibility can degrade rapidly in noisy environments (see, e.g., [1]), which can result in an operator misinterpreting or completely missing a critical signal. Whereas such difficulties may be overcome in nonspeech displays by careful manipulation of the stimulus parameters to accommodate such environments without distorting the meaning of the stimulus, such is not necessarily true in the case of speech. Another disadvantage of speech is that most of the energy in speech signals is concentrated in the lower frequency region (i.e., below 6 kHz), which means that speech signals may lack the high-frequency information needed to support accurate sound localization, particularly in regards to elevation determination and front/back discrimination ([2]). This issue is important because the ability to convey spatial information independent of the semantic content of a speech stimulus is desirable for future spatial auditory displays in which the location of the speech signal itself may convey critical information.

A possible problem with the use of speech displays is that the listener may be unable to extract information from the most relevant auditory warning when more than one warning sound is presented at the same time. Such warning sound "collisions" can result in display stimuli that are distorted or obscured, and this can lead to reduced detectability of critical signals, lowered recognition rates, and a general degradation of stimulus localizability. Despite the importance of these issues, the guidelines employed for implementing speech-based auditory displays have traditionally relied on laboratory research, most of which has employed relatively simple stimulus situations in which a single source or small number of sources are presented simultaneously. Little is known about the detectability and localizability of speech in the presence of a large number of competing speech phrases. The goal of this study was to examine both the detection and localization of a speech signal as a function of the number of sources present and the relative locations of these sources.

2. METHODS

2.1. Participants

A total of 7 paid volunteer listeners (3 males and 4 females, 20-25 years of age) participated in the experiment. All had normal hearing (i.e., bilateral thresholds $<$ 15 dB HL from 125 Hz to 8000 Hz) and all had participated in previous experiments involving both detection and localization.

2.2. Apparatus

The Auditory Localization Facility in the Air Force Research Laboratory at Wright-Patterson Air Force Base was used for the collection of behavioral data. This facility consists of an anechoic chamber, the walls, floor, and ceiling of which are covered with 1.1-m thick fiberglass wedges to reduce echoes. A 4.3-m geodesic sphere (see Figure 1), which has 277 Bose 11-cm Helical-Voice-Coil, full-range loudspeakers mounted on its surface, is housed in the chamber. The loudspeakers that were utilized in this study (239 in total) surrounded the listener (360 in azimuth and from -45 to +90 in elevation) and were directed toward the listener's head, which was positioned at the center of the sphere. (Those loudspeakers below -45 U/D were not utilized in this experiment because the direct path to the listener from these loudspeakers was, in some cases, obstructed.) This large set of locations reduced the potential for a listener to make categorical, rather than absolute, localization responses, as may be the case when more restricted sets of sound source locations are tested. Mounted directly in front of each loudspeaker on the sphere is a square cluster of four LEDs.

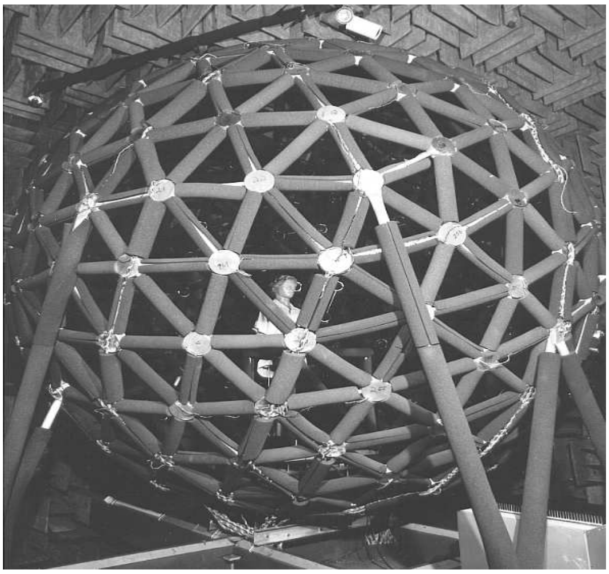

Figure 1: *The Auditory Localization Facility at Wright-Patterson Air Force Base. See text for details.*

2.3. Stimuli

The auditory stimuli employed in this experiment were 50 phonetically balanced (PB) monosyllabic words drawn from a single list of the PB50 word list corpus. This list was spoken by each of 12 talkers (6 male and 6 female) for a total of 600 unique speech tokens. The speech tokens were broadband (.2kHz - 16kHz), and were level normalized. They were also processed with the Pitch Synchronous Overlap and Add (PSOLA) algorithm in PRAAT to change their durations to exactly 500 ms.

On each trial, a target was defined by a specific speech token (i.e., a specific word spoken by a specific talker). On target-present trials, the target speech token was accompanied by the presentation of between 0 and 5 competing speech tokens. Relative to the target speech, each competing speech token was spoken by either the same talker, a different talker but of the same sex, or a different-sex talker. On target-absent trials, between 1 and 6 non-target speech tokens were presented. The individual talker characteristics were similar to those in target-present trials (i.e., all same talker, all same sex, or 1 talker that was a different sex than the other talkers), and the speech tokens were selected such that one of the tokens came from the target talker.

The individual speech tokens were convolved with the inverse transfer function from the appropriate loudspeakers in order to remove the effects of the loudspeaker frequency responses, and were then sent from an experimental control computer to a Mark of the Unicorn (MOTU 24 I/O) digital-to-analog converter. Each signal was then sent to a separate channel from a bank of power amplifiers (Crown Model CL1). These amplified signals were then directed to a custom-built signal-switching system (Winntech) before each individual signal was routed to the appropriate loudspeaker. On half of the trials, the speech tokens were spatially separated from one another, with the constraint that the angular separation between all active loudspeakers was at least 45° (the 'spatially separated' condition), and on half of the trials all speech tokens were presented from the same loudspeaker (the 'co-located' condition).

2.4. Procedure

During the experiment, each listener stood on a platform in the middle of the Auditory Localization Facility. The listeners' task was to determine whether or not a particular speech token was present (the detection phase) and then, if present, to determine the location of that speech token (the localization phase). At the start of each block of trials, the listener was required to turn to face a reference loudspeaker located directly in front of her/him on the horizontal plane and boresight a hand-held tracking device (the 'wand'; Intersense IS900), which was subsequently used to record both detection and localization responses. An LED cluster, co-located with this reference loudspeaker, was then activated briefly to indicate the start of a trial. This was followed by a cuing interval, during which the target speech token was presented (Note: in order to avoid biasing the listeners' localization responses with a directional cue, the cued target speech token was presented from the 4 horizontal-plane polar loudspeaker locations simultaneously, resulting in a diffuse image). A subsequent 500-ms silent interval was followed by the observation interval, during which the stimulus (between 1 and 6 simultaneous speech tokens) was presented.

The listener first judged whether the target was present or absent. If the target was judged to be present, the listener was required to indicate the perceived location of the target by pointing the wand at the appropriate loudspeaker and pressing a button; the orientation of the wand was indicated by activating the LED cluster at the loudspeaker to which the listener was pointing (i.e., the wand served as an LED 'cursor'). Note that, on these trials, this single localization response also served as a positive detection response. If the target speech token was judged to be absent, the listener depressed a button on the wand to indicate a ‘target-absent’ response. If, however, the target was present but was judged to be absent (i.e., a ‘miss’), the listener was nevertheless required to make a localization response. No constraints were imposed on head movements throughout the trial, but the listener was required to re-orient to the reference loudspeaker before the start of each subsequent trial. Trial-by-trial feedback was provided regarding the correctness of the detection response and the true location of the target speech token.

In each block of 48 trials, 2 trials were run in each combination of number-of-competing tokens (0-5), spatial configuration (spatially separated and co-located) and target state (present or absent). The *a priori* probability of a target-present trial was 0.5. Only one talker characteristics condition (same-talker, same-sex, different-sex) was tested in each block, and 16 blocks were run in each of these conditions, for a total of 48 blocks per listener.

3. RESULTS AND DISCUSSION

Figure 2 shows the percentage of correct detections of the target speech token as a function of the number of simultaneous competing speech tokens. The left panel shows performance in the spatially separated condition, and the right panel shows performance in the co-located condition. Within each panel, the open circles show performance for the different-sex target condition, the black squares show performance for the same-sex target condition, and the gray triangles show performance for the same-talker condition.

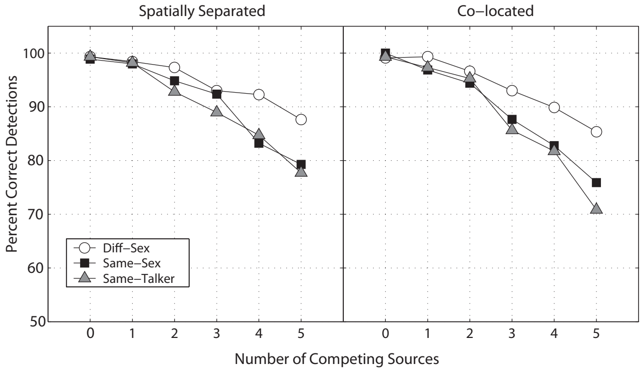

Figure 2: *The percentage of correct detections plotted as a function of the number of simultaneous competing speech tokens. The lefthand panel depicts the data from the trials in which the speech tokens were spatially separated from one another (the spatially separated condition), and the righthand panel depicts the data from trials in which all speech tokens emanated from the same loudspeaker (the co-located condition). The parameter in each panel is the specific talker characteristics condition tested (different sex target condition, same sex target condition, same talker target condition*

As would be expected, all of the curves in Figure 2 show that the listeners were able to correctly detect the presence of the target speech token 100% of the time when it was the only token presented. It can also be seen that overall detection performance decreased as the number of simultaneously presented competing speech tokens increased. However, the rate at which detection performance decreased was remarkably slow. Even in the worst case tested, where the target speech token was presented in the context of five simultaneous competing speech tokens spoken by the same talker in the same location (gray triangles in righthand panel), listeners were able to correctly detect the presence of the the target more than 70% of the time. This suggests that the detection of a known monosyllabic target word in the presence of simultaneous masking words is a remarkably robust process that may be possible even in very adverse listening environments containing multiple similar sounds.

Comparing the different curves within each panel of the figure, it is apparent that similarity between the target voice and competing voices does have an impact on the ability to detect the target. When the stimulus contained four or five competing speech tokens, detection performance was consistently 8-10 percentage points better when the target voice was different in sex than the competing tokens (open circles) than when it was the same sex as the competing tokens. On the surface, one might attribute this difference to the fact that the listener in the different-sex condition only needs to listen for the presence of a talker of a particular gender (e.g., a female voice in the presence of male voices) rather than for the actual key word spoken by that talker. However, the stimuli in this experiment were balanced so that the target-absent trials in the different-sex conditions contained the same mix of genders as the target present trials (for example, one female talker and five male talkers in the six-talker condition) and always contained a speech token from the cued talker . Thus the greater detection performance obtained for the difference-sex condition, shown in Figure 2, cannot be attributed to a detection strategy based solely on the recognition of a female target in the presence of male maskers. The most likely explanation is that the listeners in the different-sex condition were able to immediately focus their attention on the word spoken by the odd-sex talker in the stimulus, and that this made it substantially easier for them to determine if the word spoken by that odd-sex talker matched the cued target token.

Comparing the left and right panels of Figure 2, we see one of the most surprising results of the experiment: the listeners performed nearly as well in the co-located condition as they did in the spatially separated condition. More specifically, performance in the co-located condition was sufficiently good such that very little additional release from masking was seen when the tokens were spatially separated. These results appear to be inconsistent with previous results in the literature demonstrating that spatial separation does, in fact, yield improved detection performance [3] and speech *intelligibility* [4]. However, the results are in fact consistent with the notion that the spatial release from masking is very small when performance in the baseline condition (in this case, the co-located condition) is sufficiently good [5]. A closer look at the data, however, indicates that detection performance in the co-located condition degrades more rapidly than performance in the spatially separated condition as the number of competing speech tokens increases. That is, the spatial release from masking is increasing as the number of competing sounds increases. This trend suggests that much larger releases from masking might be found if the number of competing sounds extended beyond 6.

The results from the localization task are shown in Figure 3 for all listeners in all cases where the speech tokens were spatially separated and the target was correctly detected. Each row depicts the data for a single spatial dimension (left/right, L/R; front/back, F/B; up/down, U/D), as the number of competing talkers varies

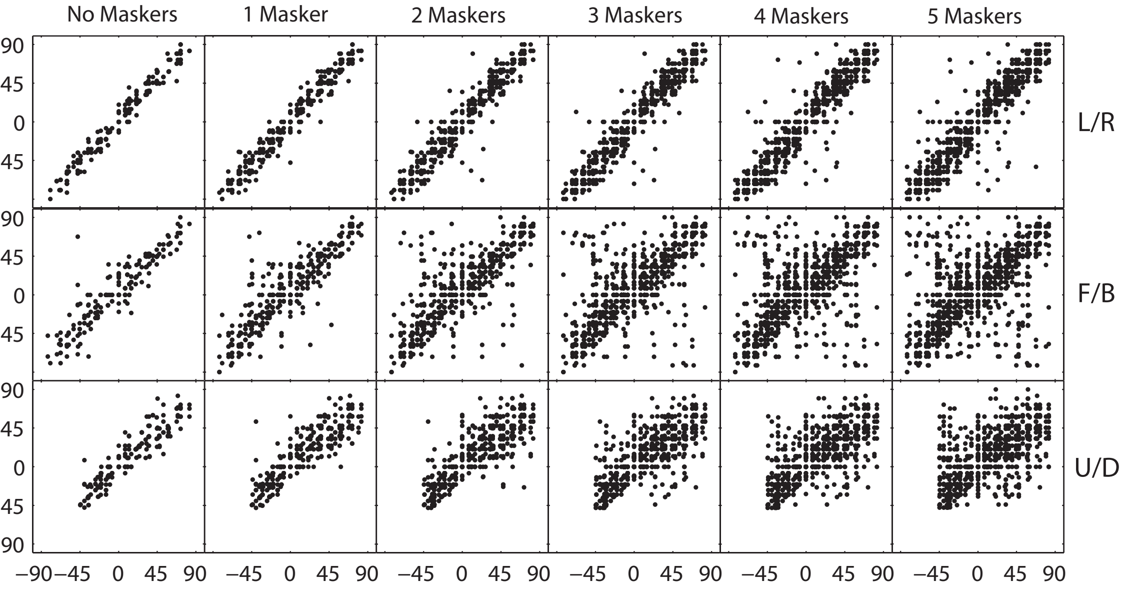

Figure 3: *Localization responses plotted as a function of the actual source locations for all listeners in the left/right dimension (top row), front/back dimension (middle row) and up/down dimension (bottom row). The number of competing sources increases as you move from the left-most panel to the right-most panel. Perfect performance would result in all responses falling along the positive-slope diagonal.*

from zero (the left-most panel) to 5 (the right-most panel). As can be seen, localization in the L/R dimension was found to be quite accurate, as can be seen by the proximity of the data points to the positive-slope diagonal, particularly when the number of competing sources was small. Localization in the U/D dimension was worse than the L/R dimension, as indicated by a greater spread of data points around the positive-slope diagonal. Localization in the F/B dimension was worse than both the L/R and U/D dimensions. These results are consistent with previous results in the literature (e.g., [6]). One can also see that, as the number of competing sources increases, localization accuracy degrades systematically in all dimensions, but much more rapidly and to a much greater extent in the F/B and U/D dimensions. These results are summarized in Figure 4, where the mean rms errors in each spatial dimension are plotted as a function of the number of competing speech tokens. In all dimensions, the rms errors increase with the number of competing sounds. However, the errors in the L/R dimension remain relatively low, not exceeding 18° until more than four competing sounds are present in the stimulus. The rms errors are slightly larger in the U/D dimension, and are larger still in the F/B dimension. In fact, the rms errors in the F/B dimension are greater at every point along the curves than those in the L/R and U/D dimensions for the corresponding conditions. It is interesting to note that the similarity between the target voice and the voices of the competing speech tokens makes no difference in the F/B and U/D dimensions, but that localization in the L/R dimension does, in fact, seem to be better when the target is a different-sex than when it is more similar to the competing voices.

Figure 5 combines the L/R, F/B, and U/D localization errors shown in Figure 3 into a single overall measure of angular (great circle) error. As in Figure 2, the two panels show performance in the two spatial conditions of the experiment, and the individual curves within each panel show the different target-masker similarity conditions in the experiment. In the easiest localization conditions, where the target token and/or competing tokens were all presented from the same spatial location (i.e. the no-masker condition in the left panel of the figure and all co-located conditions in the right panel of the figure), the overall angular errors averaged approximately 15°. Note that this is roughly the same angular error reported by [7] for broadband sounds. In part, the relatively high level of performance obtained for the speech stimuli in this experiment can be explained by the use of some exploratory head movements. The 500 ms stimulus duration in this experiment was not long by any means, but it probably afforded the listeners some opportunity to initiate a head movement and thus helped to reduce front-back confusions. The front-back confusions and elevation errors were probably also reduced by the use of broadband speech recordings. Recent studies have shown that sufficient high-frequency information is preserved in broadband speech to support relatively accurate localization [8], despite the fact that most of the energy (and virtually all of the intelligibility information) in speech is concentrated at frequencies below 6 kHz.

As the number of maskers in the spatially separated condition increased (left panel), the average localization error increased in a nearly linear fashion, with approximately a 5-7° increase in angular error for each additional masker added to the stimulus. However, it is worth noting that performance remained well above chance performance (approximately 90° error) even in the worst case tested with five simultaneous maskers. In that case, the overall average error was around 45°, which, although not very accurate, does indicate that the listeners were able to recover some spatial information about the target.

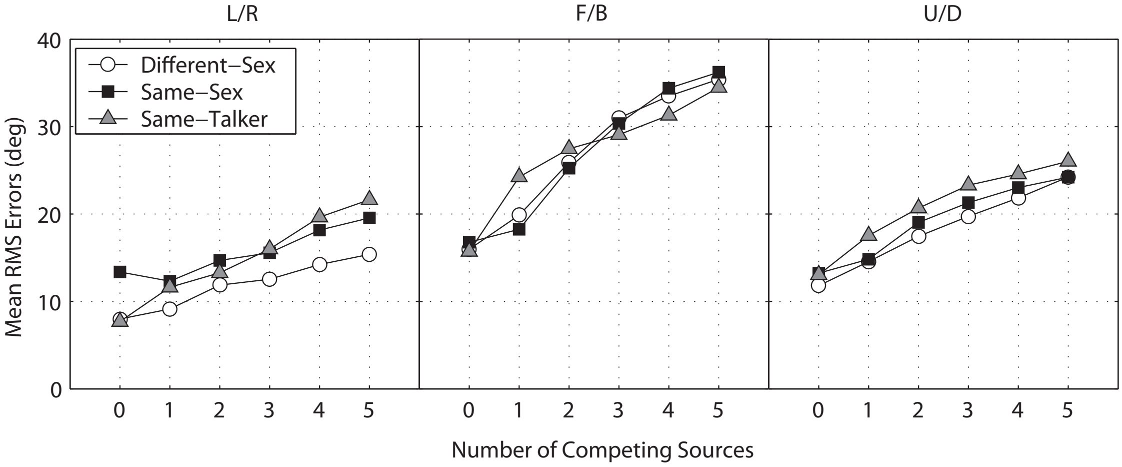

Figure 4: *Mean rms errors are plotted as a function of the number of competing sources in the L/R, F/B, and U/D dimensions.*

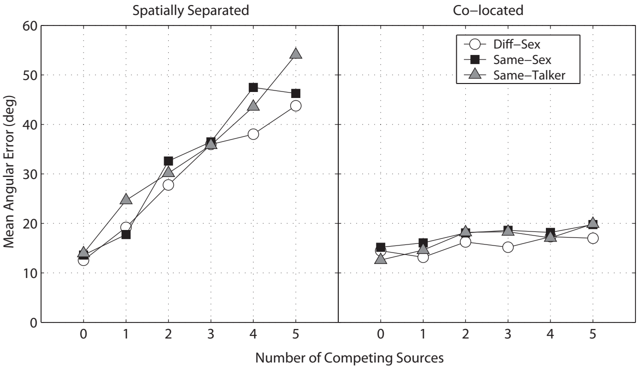

Figure 5: Overall angular errors in the experiment plotted as a function of the number of interfering sounds. These data are plotted in the same format used in Figure 2.

4. CONCLUSIONS

Listeners' ability to detect and localize a target speech token was measured as a function of the number of competing speech tokens and the spatial separation among these tokens. The results show that although performance decreased as the number of competing sources increased, both detection and localization where surprisingly accurate even with 5 competing sources. Additional research is needed to examine how performance degrades when even greater numbers of sources are used, to determine the role of head movements, and to reconcile apparent inconsistencies with previous ”cocktail-party” effect experiments. Of particular interest is the functional relation between detection and localization mechanisms. In this study where the target token is known (via the cuing interval), but the target location is not, spatial separation has little impact on detection performance, apparently supporting a ”what-then-where” strategy. This hypothesis could be systematically examined in a study that varied the uncertainty of the target token and the target location.

5. REFERENCES

[1] G.A. Miller, “The masking of speech,” *Psychological Bulletin*, vol. 44, pp. 105–129, 1947.

[2] R. H. Gilkey and T. R. Anderson, “The accuracy of absolute localization judgments for speech stimuli,” *Journal of Vestibular Research*, vol. 5, pp. 487–497, 1995.

[3] M. D. Good, R. H. Gilkey, and J. M. Ball, “The relation between detection in noise and localization in noise in the free field,” in *Binaural and Spatial Hearing in Real and Virtual Environments*, R.H. Gilkey and T.R. Anderson, Eds., pp. 349–376. Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

[4] R. Drullman and A.W. Bronkhorst, “Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation,” *Journal of the Acoustical Society of America*, vol. 107, pp. 2224–2235, 2000.

[5] I. J. Hirsh, “The influence of interaural phase on interaural summation and inhibition,” *Journal of the Acoustical Society of America*, vol. 20, pp. 592–599, 1948.

[6] F. L. Wightman and D.J. Kistler, “Factors affecting the relative salience of sound localization cues,” in *Binaural and Spatial Hearing in Real and Virtual Environments*, R.H. Gilkey and T.R. Anderson, Eds., pp. 1–23. Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

[7] F.L. Wightman and D.J. Kistler, “Headphone simulation of free-field listening. ii: Psychophysical validation,” *Journal of the Acoustical Society of America*, vol. 85, pp. 868–878, 1989.

[8] V. Best, S. Carlile, C. Jin, and A. van Shaik, “The role of high frequencies in speech localization,” *Journal of the Acoustical Society of America*, vol. 118, pp. 353–363, 2005.