The role of high frequencies in speech localization

Virginia Best and Simon Carlile
Department of Physiology, University of Sydney, Sydney, NSW, 2006, Australia

Craig Jin and André van Schaik
School of Electrical and Information Engineering, University of Sydney, Sydney, NSW, 2006, Australia

(Received 25 January 2005; revised 14 April 2005; accepted 14 April 2005)

This study measured the accuracy with which human listeners can localize spoken words. A broadband (300 Hz–16 kHz) corpus of monosyllabic words was created and presented to listeners using a virtual auditory environment. Localization was examined for 76 locations on a sphere surrounding the listener. Experiment 1 showed that low-pass filtering the speech sounds at 8 kHz degraded performance, causing an increase in polar angle errors associated with the cone of confusion. In experiment 2 it was found that performance in fact varied systematically with the level of the signal above 8 kHz. Although the lower frequencies (below 8 kHz) are known to be sufficient for accurate speech recognition in most situations, these results demonstrate that natural speech contains information between 8 and 16 kHz that is essential for accurate localization. © 2005 Acoustical Society of America.

DOI: 10.1121/1.1926107

PACS number(s): 43.66.Qp [AJO]

I. INTRODUCTION

Human speech is a dynamic acoustic stimulus, varying in both frequency and intensity over time in a complex manner. A large body of research exists that is concerned with how speech recognition depends on the identification of these acoustic patterns, as well as how recognition is influenced by the presence of noise or other interfering sources. It is well known that the spatial layout of sources in a noisy environment has a large impact on how well they can be segregated, and differences in spatial location can aid in the focusing of selective attention on a talker of interest (see Zurek, 1993; Bronkhorst, 2000). However the spatial perception of human speech sounds in quiet has actually received little attention in the literature.

The ability of subjects to locate a sound in space is dependent on the integration of information from a number of acoustic cues (see Carlile, 1996; Blauert, 1997 for reviews). A sound source located away from the midline results in differences in the arrival time and the level of the sound at each ear. Thus for a sound source positioned around a listener, there will be a set of interaural cues that the auditory system can use to estimate the direction of the incoming source. However, as the ears are placed more or less symmetrically on the head, these interaural cues are ambiguous and result in the so-called “cone of confusion” where a single interaural difference corresponds to the surface of a cone centered on the interaural axis. The three-dimensional asymmetry of the outer ear results in a location-dependent spectral filtering of sounds which can provide spectral cues to resolve the ambiguities in the interaural cues. The dynamic nature of speech provides the spatial auditory system with particular challenges. In particular, it is likely that spectral fluctuations inherent in the source spectrum would hinder the extraction of the spectral cues introduced by the head and pinnae.

Speech is widely used as a stimulus in auditory research, especially in the important areas of hearing loss and communication in noisy environments. Importantly, it has become apparent that such studies have frequently used low-pass filtered (8 kHz or below) speech stimuli. The explanation for this is that most of the important sound components for speech recognition, such as formants, occur well below 8 kHz. However, naturally produced speech does contain significant amounts of energy above 8 kHz. Broadband segments can extend up to around 20 kHz [an example is shown in Fig. 1(b)]. It is also known that energy in this high-frequency region (above 8 kHz) is important for accurate auditory localization of nonspeech signals (Bronkhorst, 1995; King and Oldfield, 1997; Carlile and Delaney, 1999; van Schaik et al., 1999). As the impact of these high-frequency cues on speech localization is unclear, the primary aim of the current experiment was to compare localization of broadband and low-pass filtered speech stimuli.

A few studies have examined speech localization in the horizontal plane using single words presented in virtual auditory space and found that although the lateral angle (the angle away from the median plane) was estimated as accurately as for nonspeech broadband stimuli, there was an increase in front–back confusions (Begault and Wenzel, 1993; Ricard and Meirs, 1994). Using a larger range of locations, Gilkey and Anderson (1995) compared the localization of click trains to recorded single words. They examined positions distributed randomly on a sphere (including elevations between −45° and 90°). They reported that performance was comparable in the left–right dimension, but poorer in the front–back dimension and in elevation for speech stimuli. The kinds of errors reported for speech stimuli are typical cone of confusion errors, arising commonly in situations where the spectral cues to sound source location are somehow compromised.
While this previous work indicates that cone of confusion errors are increased for speech localization compared to localization of simple non-speech stimuli, the present study explored directly the relationship between speech bandwidth and localization performance for a large range of spatial locations.

II. EXPERIMENTAL METHODS

A. Subjects and environment

Five subjects participated in the experiments (3 male, 2 female, aged 24–34) and all had previous experience in auditory localization experiments. Stimuli were delivered over headphones using virtual auditory space technology (see Sec. II B) at a level of approximately 75 dB SPL. Acoustical recordings for the generation of virtual stimuli as well as localization testing took place in an anechoic chamber situated at the University of Sydney (dimensions 4 m × 4 m × 4 m).

B. Generation of virtual auditory space

Virtual auditory space (VAS) is generated by recording the acoustic filter functions of the auditory periphery of individual listeners [the head-related transfer functions (HRTFs)] and convolving these filter functions with sounds subsequently presented over headphones to create a realistic, externalized percept. HRTFs were measured using a “blocked-ear” recording technique. This approach involves embedding a small recording microphone in an earplug secured flush with the distal end of the ear canal (Møller et al., 1995). The recordings were performed in the anechoic chamber with the subject at the center. A loudspeaker, mounted on a robotic arm, delivered the stimuli from locations on an imaginary sphere of 1 m radius around the listener’s head. The automated recording procedure recorded an HRTF for the right and left ear at 393 locations evenly distributed around the sphere. The position of the subject’s head was stabilized using a chin rest and head orientation was monitored using an electromagnetic tracking system to ensure that it was fixed. Impulse responses were measured using Golay code pairs of length 1024 and a sampling rate of 80 kHz (Golay, 1961). The HRTFs are composed of a location-independent component (LIC) and a location-dependent component. The LIC for a given ear is estimated as the rms magnitude spectrum of the entire set of recordings for that ear. The LIC is then deconvolved from the recorded HRTFs to obtain the directional transfer functions (DTFs). The primary reason for performing such a manipulation is to remove measurement artifacts including speaker and microphone transfer functions and the detailed acoustic effects associated with the precise location of the microphone.

The desired acoustical stimulus is then convolved with the right and left ear DTFs appropriate for a particular sound location and presented via in-ear headphones (Etymotic ER-2). These headphones are designed to have a flat frequency response up to 10 kHz, and were considered appropriate for the purposes of this study on the basis of their ability to produce high-fidelity VAS (see Sec. II D). All stimuli were windowed by applying a raised cosine to the first and last 10 ms before convolution with DTFs.

C. Speech stimuli

As a publicly available broadband speech corpus (i.e., not filtered at 8 kHz or below) could not be found, the Harvard list (Egan, 1948) was recorded by a male actor with extensive vocal training from the National Institute of Dramatic Arts (Sydney). This corpus consists of 20 phonetically balanced word lists each containing 50 monosyllable words. These were recorded using a Brüel and Kjær 4165 microphone and a Brüel and Kjær 2610 amplifier. The speech was digitized at a sample rate of 80 kHz using an anti-aliasing filter with a 30 kHz cut-off. The first 5 of the 20 lists were used in the present experiment. The duration of the words in these lists ranged between 418 and 1005 ms with an average duration of 710 ms.

Because this experiment examines the influence of high-frequency spectral information on localization, it was very important that background noise be eliminated. This was accomplished by: (i) recording the speech stimuli in the anechoic chamber, (ii) seating the actor (in a quiet chair) close to the microphone; (iii) setting the amplifier so that signals were within and spanned the maximum range of the analog-to-digital converter; (iv) turning off unnecessary equipment. Figure 1 shows a spectrogram of a silent period and a recorded speech signal (the word “sludge”) under two different filtering conditions (8 and 16 kHz low-pass). The spectrograms show that the recordings were made with high signal-to-noise ratio and that background noise was not a concern.

D. Localization testing

The localization testing paradigm in place in the laboratory (Carlile et al., 1997) was used to validate the VAS.
stimuli as well as to measure localization for the specific test stimuli under examination in this study.

Briefly, in the validation test, the ability of a subject to localize a burst of broadband noise in free-field anechoic space was compared with his/her localization performance for the same type of stimulus presented in VAS. The test required the subject to stand in darkness in the center of the anechoic chamber and indicate the perceived direction of a series of 150 ms broadband noise bursts presented from 76 random locations evenly distributed on an imaginary sphere surrounding his/her head. For the free-field component of the test, sound sources were presented from a loudspeaker that could be positioned at any location on this sphere. For the VAS component, stimuli were presented over headphones at the corresponding “virtual” locations on this sphere. To respond, the subject was required to turn and face the perceived location of the target and point his/her nose at the source. An electromagnetic tracking receiver mounted on the top of the head (Polhemus: IsoTrack) was used to measure the orientation of the head and thus provided an objective measure of the perceived direction. Five such tests were completed by each subject in the two conditions. If performance under the two stimulation conditions was accurate (spherical correlation coefficient of at least 0.85, see Sec. II E), the subject was considered a sufficiently accurate localizer and the VAS a valid simulation. This was the case for all subjects used in this study.

Localization performance under the test stimulus conditions for Experiments 1 and 2 was measured using the same paradigm as above. Each experiment consisted of three stimulus conditions and Experiment 1 was completed by each subject before the commencement of Experiment 2. Experiment 1 compared the ability of subjects to localize broadband noise, broadband speech, and 8 kHz low-pass filtered speech. In Experiment 2, subjects localized speech stimuli in which the level of the high-frequency information was systematically varied.

For each condition, five localization responses were obtained for each of the 76 stimulus locations. For each of the five replicates, the spoken word signals were chosen randomly from one of the five balanced lists. As there were 76 locations and only 50 words per list, 26 of the words were played twice per condition. For each condition, the same word stimuli were played from the same location on the sphere, thus keeping the stimulus set identical between conditions. Within each experiment, trials from the three conditions were interleaved and stimuli were presented in a randomized fashion. Each of the two experiments consisted of 1140 localization trials in total (3 conditions × 76 locations × 5 replicates), which were divided into 15 localization tests. Thus in total, each subject completed 30 tests (approximately 8 h of listening time) over a period of about two months.

Subjects understood that they did not have to identify the spoken words, only localize them. However, when questioned after testing, they did not report any difficulty in recognizing the relatively common words.

E. Data analysis

In order to gauge the overall performance of subjects in the different experimental conditions, the spherical correlation coefficient (SCC; Fisher et al., 1987) was calculated. Its use with localization data is described in detail elsewhere (Leong and Carlile, 1998), but in brief, it describes the degree of correlation between target and response locations based on their directional cosines (1=perfect correlation; 0=no correlation).

To investigate the pattern of responses more closely, the localization data were analyzed in terms of lateral and polar angles (Fig. 2). Analysis of localization responses involved plotting actual target lateral (or polar) angle against perceived lateral (or polar) angle. Furthermore, lateral and polar angle errors were calculated for each trial and pooled in order to compare performance across stimulus conditions. Finally, in order to compare results to those of previous studies, it was useful to calculate the percentage of cone of confusion errors made by subjects in each condition. These were de-
fined as large polar angle errors (>90° in magnitude) and as such included the commonly reported front–back and up–down confusions.

III. EXPERIMENT 1: SPEECH LOCALIZATION

A. Conditions

Experiment 1 consisted of three stimulus conditions. In condition A, which acted as a control, the stimulus was a 150 ms white noise burst that was windowed by applying a raised cosine to the first and last 10 ms. In condition B, stimuli were broadband speech signals. Both noise and speech signals were band-passed between 300 Hz and 16 kHz. Importantly, the duration of noise and speech stimuli were not matched. However, if this were to have an effect, it would be expected that the longer duration of the speech stimuli might confer an advantage for localization over the control broadband noise condition. Note also that as the stimuli were presented in virtual auditory space, any head movements made during stimulus presentation would not have provided any dynamic localization cues. However, subjects were instructed to keep their head still during the stimulus to ensure that the perceived position did not change in relation to the measurement apparatus. In condition C, stimuli were speech signals low-pass filtered at 8 kHz before DTF filtering. Low-pass filtering was performed in the time domain using brick-wall FIR filters (60 dB down in the stop band). The effect of the low-pass filtering is demonstrated in Fig. 1 where it can be seen that certain parts of speech signals contain substantial high-frequency information that is removed with the 8 kHz low-pass filtering.

B. Results

Table I shows the SCC calculated for each subject under the three stimulus conditions. The SCC was highest in the broadband noise condition, fell slightly in the broadband speech condition, and fell considerably in the low-pass speech condition.

Table I. Spherical correlation coefficients (SCCs) for each of the five subjects in Experiment 1 (and the mean). Each of the three rows contains values for the three stimulus conditions. Each SCC is calculated on the basis of five repetitions at each of 76 stimulus locations (i.e., 380 trials in total). See the text for details of this statistic.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB noise</td>
<td>0.82</td>
<td>0.88</td>
<td>0.91</td>
<td>0.90</td>
<td>0.91</td>
<td>0.88</td>
</tr>
<tr>
<td>BB speech</td>
<td>0.75</td>
<td>0.87</td>
<td>0.87</td>
<td>0.88</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td>LP speech</td>
<td>0.40</td>
<td>0.79</td>
<td>0.66</td>
<td>0.76</td>
<td>0.66</td>
<td>0.65</td>
</tr>
</tbody>
</table>

FIG. 3. Scatter plots showing lateral angle data (pooled across all subjects) for Experiment 1. The three panels contain data for the three stimulus conditions: broadband noise (left), broadband speech (middle), and low-pass filtered speech (right). Target lateral angle (abscissa) is plotted against response lateral angle (ordinate) and the size of the dots represents the number of responses clustered at a point.
response errors, particularly for stimuli presented in the lower frontal regions (polar angle range $-90^\circ$–$0^\circ$).

In order to summarize the magnitude of the lateral and polar angle errors, the absolute value of the errors was calculated and the mean was obtained for each subject in each stimulus condition. Mean errors and standard errors of the means (SEMs) are shown in Fig. 5. Lateral angle errors [Fig. 5(a)] were small in general and did not appear to vary systematically across the three stimulus conditions (means of $8.3^\circ$, $8.9^\circ$, and $8.5^\circ$). A Kruskal–Wallis nonparametric ANOVA performed on the group data revealed no significant differences across conditions [$\chi^2(2,5697)=5.35, p=0.069$]. Mean polar angle errors [Fig. 5(b)], however, varied considerably between stimulus conditions. Overall, polar angle errors were consistently smallest in the broadband noise condition (mean $12.3^\circ$), higher in the broadband speech condition (mean $30.9^\circ$), and highest for low-passed speech (mean $46.4^\circ$). A Kruskal–Wallis nonparametric ANOVA per-

FIG. 4. Scatter plots showing polar angle data for Experiment 1. Each row shows data for a different subject as labeled. The three panels contain data for the three stimulus conditions: broadband noise (left), broadband speech (middle), and low-pass filtered speech (right). Target polar angle (abscissa) is plotted against response polar angle (ordinate) and the size of the dots represents the number of responses clustered at a point.
formed on the group data revealed a highly significant effect of condition \( [\chi^2(2.5622)=401.94, p<0.001] \), and post-hoc analysis (Tukey HSD, \( p=0.05 \)) found significant differences between all three conditions.

Table II shows the calculated cone of confusion error rates. For different subjects, the number and distribution of cone of confusion errors across the sound conditions varied, although there were fewer errors on average for the broadband noise condition, an increase in the broadband speech condition, and a much larger number of errors for the low-pass speech condition.

C. Discussion

The findings of the current study were consistent with those of previous studies. Subjects accurately estimated the lateral angle of a source regardless of spectral content, as the available interaural differences provide a robust cue to lateral angle. This has been reported previously for both speech (Ricard and Meirs, 1994; Gilkey and Anderson, 1995) and non-speech stimuli (Gilkey and Anderson, 1995; Carlile et al., 1997). Performance in the control condition confirmed the well-supported notion that broadband flat-spectrum sounds are well localized in polar angle (Carlile et al., 1997; King and Oldfield, 1997). In terms of the localization errors that arise when the monaural spectral cues are ambiguous, the present study confirmed and extended the findings of previous researchers. Our basic finding was that polar angle errors increased somewhat for sound stimuli with time-varying spectra (speech) but particularly when high-frequency content was removed (low-passed speech).

It was interesting to find that subjects were relatively accurate at localizing broadband speech, despite some increase in confusions compared to broadband noise. Speech has a spectrum that is nonflat, and also varies over time, and yet listeners showed a reasonable capacity for distinguishing spectral cues to location from spectral features of the source spectrum. It may be that familiarity with the source spectrum plays an important role here, as suggested by Plenge and Brunschen (1971). These authors showed that familiar voices (with familiar spectra) are localized better in the median plane than unfamiliar voices, presumably because the spectrum due to directional filtering can be calculated. In the current experiment the same talker was used on every trial and thus his voice would have been familiar to listeners. Poorer polar angle estimates might be expected if the talker was varied randomly from trial to trial such that the source spectrum could not be predicted. However even for unfamiliar stimuli with nonflat spectra the spectral cue localization process shows some robustness. For example, subjects can accurately localize “scrambled spectrum” and “ripped spectrum” sounds as long as the spectral manipulations are moderate (Wightman and Kistler, 1989; Wightman and Kistler, 1997; Kulkarni and Colburn, 1998; Macpherson and Middlebrooks, 2003).

The low-pass filtered speech stimuli were included in this study because speech stimuli in the large majority of previous work have been low-pass filtered at 8 kHz or below. The results presented here indicate that low-pass filtered speech stimuli are localized much less accurately in the polar angle dimension than broadband speech stimuli. Out of the literature consulted, only two localization studies used a cut-off frequency higher than 8 kHz. One was the Gilkey and Anderson study (1995), where speech stimuli were band-pass filtered from 400 Hz to 11 kHz. Due to the differences in analyses, it is difficult to compare performance of subjects in that study to the current one. However, it appears, by inspection of their figures, that the number of errors in Gilkey and Anderson’s data is larger than in our broadband speech (300 Hz–16 kHz) condition. This suggests that frequency bands as high as 11–16 kHz are useful for accurate sound localization of speech.

In another study, consonant–vowel (CV) localization was examined in the presence of diffuse background noise (Karlsen, 1999). In that study, CVs were recorded in an anechoic environment with a low-pass cut-off frequency of 10 kHz. Localization was tested in the horizontal plane. The author reported a relatively low percentage of front–back errors which, by inspection of the data, appears to be approximately 15% of the total trials. This is greater than the pro-
portion of cone of confusion errors calculated in the present study for broadband speech (8.4%). This may suggest that frequencies between 10 and 16 kHz can benefit speech localization, or it may be that CVs are more spectrally impoverished than monosyllabic words. In any case, comparisons must be made with caution, as the testing locations and criteria for defining cone of confusion errors differed between the present study and that of Karlsen.

D. Analysis of high frequency content of stimuli

An interesting analysis included in Karlsen’s work (1999) looked at the differences in localization accuracy for different CVs. In other words, he was probing the specific components of human speech that are useful for localization. He reported that subjects were most sure of the location of the /s/ consonant. In other words, he was probing the specific frequencies between 10 and 16 kHz that are the most useful for speech localization, or it may be that CVs are more spectrally impoverished than monosyllabic words. In any case, comparisons must be made with caution, as the testing locations and criteria for defining cone of confusion errors differed between the present study and that of Karlsen.

It was of interest for the present experiment to take this approach and examine whether the specific frequency content of different words was related to the localization performance. Specifically, it was suspected that words with substantial energy above 8 kHz would be well localized and that performance on these words would be most affected by the low-pass filtering. To examine this, a spectrogram of each of the 250 words used was calculated using a routine from the Auditory Toolbox for MATLAB (Slaney, version 2, 1998). The total energy above 8 kHz was summed and taken as a metric (in arbitrary units) of high frequency content. The words were then divided into 10 bins each containing 25 words. For each bin, polar angle errors were calculated for all trials (across all subjects) in which its members were the stimulus. Mean absolute polar angle errors were calculated independently for the broadband and low-pass filtered conditions.

Figure 6 shows the mean errors for the 10 bins. Note that the bins represent words with increasing amounts of high-frequency energy, and that bin 1 has the least and bin 10 has the most. Three main points emerge from this simple analysis. First, it is clear that the curves do not change smoothly as a function of high-frequency energy. This may be because the energy below 8 kHz in these words (which varies independently of the high-frequency energy) has an influence on the amount of error. Second, for the broadband condition, there appears to be an overall decrease in the mean polar angle error with increasing high-frequency content, suggesting that words with more high frequency energy were better localized. The low-passed condition shows a nonsystematic variation, but since these sounds were presented without any of their high-frequency energy, no relationship is expected here between high-frequency content and performance. Third, for all bins, the mean errors are significantly lower in the broadband condition compared to the low-passed condition. Even for the first bin, i.e., the 25 words with the lowest amount of high-frequency energy, the improvement is substantial. Clearly these words have enough energy above 8 kHz to afford the auditory system a better estimate of the location.

This analysis showed that all of the 250 word stimuli used in this experiment contained high-frequency information (above 8 kHz) that was useful for polar angle localization. This emphasizes that naturally spoken speech is a broadband stimulus, a notion that is frequently overlooked in the literature. Experiment 2 was conducted to test the hypothesis that the preservation of this high-frequency energy, even at a very low level, can benefit the localization of natural speech. To examine whether accuracy is related to the level of this energy, or simply its presence, localization performance was measured with the high-frequency region systematically attenuated.

IV. EXPERIMENT 2: INFLUENCE OF HIGH-FREQUENCY LEVEL ON SPEECH LOCALIZATION

A. Conditions

Experiment 2 consisted of three stimulus conditions. Condition A acted as a control and stimuli were broadband speech signals (identical to those of Experiment 1 condition B). In conditions B and C, the level of the high-frequency information (above 8 kHz) was systematically varied. In condition B the high-frequency region was attenuated by 20 dB and in condition C it was attenuated by 40 dB. This was achieved by low-pass filtering the signals (as in Experiment 1 condition C) with varying attenuation in the stop band. Figure 7 shows the power spectral density plots of a typical word under the different conditions.

B. Results

Table III shows the SCC calculated for each subject under the three stimulus conditions of this experiment as well as the low-pass condition of Experiment 1. The SCC was
highest in the broadband speech condition and in nearly all cases showed a gradual decline with increasing attenuation in the high-frequency region.

The correspondence between target and response lateral angles for the three speech conditions is illustrated in Fig. 8. As the performance data across all subjects showed a very similar pattern, their data were combined for plotting. It can be seen that target lateral angles correspond well with response lateral angles in all conditions, as most of the data fall on or near the "perfect response" diagonal.

As in Experiment 1, the polar angle data were highly individualized. Figure 9 illustrates the correspondence between target and response polar angles. In the broadband speech condition, polar angle estimates were accurate across all subjects, as was seen in Experiment 1. When the high-frequency level was lowered, S1 and S5 showed a dramatic increase in error for the 20 dB down condition, which was exacerbated in the 40 dB down condition. As was seen for low-passed speech in Experiment 1, these two subjects had strong tendency to localize stimuli presented in the lower hemisphere of space (polar angle range −180°−0°) to the upper hemisphere (polar angle range 0°−180°). The other subjects were less affected by the 20 dB drop in high-frequency level, with their response data showing just a small spread. For the 40 dB down condition, this spread in error was more severe and many large errors are evident.

Figure 10 summarizes the magnitude of the lateral and polar angle errors. Mean lateral and polar angle errors (with SEMs) are shown for each subject under the three conditions of this experiment. Data from the low-pass condition of Experiment 1 are also included for comparison. Mean lateral angle errors [Fig. 10(a)] were again small and did not vary systematically across conditions (means of 9.4°, 9.7°, and 9.6° for 0, 20, and 40 dB down). A Kruskal–Wallis non-parametric ANOVA performed on the group data revealed no significant differences across conditions [$\chi^2(2,5697) = 1.68, p = 0.432$]. Mean polar angle errors [Fig. 10(b)], however, varied considerably across conditions. Values were consistently smallest in the control broadband speech condition (mean 26.6°) and increased with increasing attenuation in the high-frequency region (means of 38.0° and 42.3° for attenuations of 20 and 40 dB, respectively). A Kruskal–Wallis non-parametric ANOVA performed on the group data revealed a highly significant effect of condition [$\chi^2(2,5622) = 128, p < 0.001$], and post-hoc analysis (Tukey HSD, $p = 0.05$) found significant differences between all three conditions. Inspection of the low-pass data from Experiment 1 (60 dB down) suggests that this gradual trend continues with further decreases in high-frequency level.

Table IV shows the calculated cone of confusion error rates. On average, the error rate increased with increasing attenuation in the high-frequency region.

C. Discussion

These results showed that the polar angle localization of speech is related to both the presence and the level of high-frequency information in the stimulus. There are two possibilities to explain the gradual decline in performance with increasing attenuation in the region above 8 kHz. One option

---

**TABLE III. SCCs for each of the five subjects in Experiment 2 (and the mean).** The first three rows contain values for the three new stimulus conditions and the fourth row is a reiteration of the data from the low-pass condition of Experiment 1. Each SCC is calculated on the basis of five repetitions at each of 76 stimulus locations (i.e., 380 trials in total). See the text for details of this statistic.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB speech</td>
<td>0.78</td>
<td>0.88</td>
<td>0.86</td>
<td>0.87</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>20 dB down</td>
<td>0.57</td>
<td>0.82</td>
<td>0.75</td>
<td>0.82</td>
<td>0.75</td>
<td>0.74</td>
</tr>
<tr>
<td>40 dB down</td>
<td>0.56</td>
<td>0.77</td>
<td>0.70</td>
<td>0.78</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>LP speech</td>
<td>0.40</td>
<td>0.79</td>
<td>0.66</td>
<td>0.76</td>
<td>0.66</td>
<td>0.65</td>
</tr>
</tbody>
</table>

---

**FIG. 7.** An illustration of the stimulus conditions employed in Experiment 2. Shown are power spectral density plots of a speech stimulus after low-pass filtering with varying attenuation in the stopband (0, 20, 40, and 60 dB down). Note that 60 dB down was not retested, but is equivalent to the low-pass speech condition of Experiment 1 and is included for comparison.

**FIG. 8.** Scatter plots showing lateral angle data (pooled across all subjects) for Experiment 2. The three panels contain data for the three stimulus conditions: broadband speech (left), 20 dB down low-pass speech (middle), and 40 dB down low-pass speech (right). Target lateral angle (abscissa) is plotted against response lateral angle (ordinate) and the size of the dots represents the number of responses clustered at a point.
is that more and more words in the corpus have their high-
frequency content attenuated to an inaudible level (depend-
ing on the original level) and thus the error rate may be
related to the number of broadband words presented. The
extreme case would be the low-pass speech (60 dB down)
condition, where it is assumed that no words carry high-
frequency information. A second possibility is that most
words retain some high-frequency content in the 20 and
40 dB down conditions, and that the system has an ability to
make use of this low-level energy. Indeed the analysis fol-
lowing Experiment 1 showed that words with very low high-
frequency energy were localized better if this energy was
preserved. If this is the case, then the gradual decline in
performance with increasing attenuation must mean that lo-
calization is impaired at low levels.

It has been observed using nonspeech stimuli that local-
ization accuracy is a function of stimulus level (Harris,
1998). Using broadband noise, it was reported that elevation
perception worsened and front–back confusions increased if
the stimuli were presented at low sound pressure levels
(close to the audibility threshold). Furthermore, Abouchacra
et al. (1998) presented speech phrases in diffuse noise, and
reported that localization accuracy in the horizontal plane
improved as signal-to-noise ratio increased (18%, 89%, and
95% accuracy at −18 dB, +12 dB, and in quiet, respec-
tively). Although these level manipulations were not re-
stricted to the high-frequency region, they demonstrate that
the auditory localization system has an ability to utilize lo-
calization cues that appears to depend on the level of the
signal. Further to this, Hofman and Van Opstal (2003) dem-
onstrated that the spectral cues are binaurally weighted, with
the perceived elevation of a source being more strongly in-
fluenced by the spectral cue at the near ear than that at the
far. It is possible that this weighting is related to the relative

FIG. 9. Scatter plots showing polar angle data for Ex-
periment 2. Each row shows data for a different subject
as labeled. The three panels contain data for the three
stimulus conditions: broadband speech (left), 20 dB
down low-pass speech (middle), and 40 dB down low-
pass speech (right). Target polar angle (abscissa) is plot-
ted against response polar angle (ordinate) and the size
of the dots represents the number of responses clustered
at a point.

smaller ears than others. As the high-frequency spectral cues derive more reliant on high-frequency spectral features for localizations seen with low-pass filtering is that some subjects are and their error patterns. A likely explanation for the differences seen with low-pass filtering of the speech stimuli, their vulnerability to low-pass filtering of the speech stimuli, experiments is the presence of strong individual differences. The subjects varied in their baseline localization accuracy, though the lower frequencies preservation of information above 8 kHz, even at a low level, provided a benefit for polar angle localization. Although the lower frequencies (below 8 kHz) are known to be sufficient for accurate speech recognition in most situations, these results demonstrate that natural speech contains information above 8 kHz that is essential for accurate polar angle localization.

VI. CONCLUSIONS

A broadband speech corpus (300 Hz–16 kHz) was used to investigate the ability of human listeners to localize monosyllabic words. Experiment 1 showed that low-pass filtering the stimuli at 8 kHz dramatically degraded performance in the polar angle dimension and increased errors associated with the cone of confusion. Experiment 2 showed that the preservation of information above 8 kHz, even at a low level, provided a benefit for polar angle localization. Although the lower frequencies (below 8 kHz) are known to be sufficient for accurate speech recognition in most situations, these results demonstrate that natural speech contains information above 8 kHz that is essential for accurate polar angle localization.

One of the striking features of the data collected in these experiments is the presence of strong individual differences. The subjects varied in their baseline localization accuracy, their vulnerability to low-pass filtering of the speech stimuli, and their error patterns. A likely explanation for the differences seen with low-pass filtering is that some subjects are more reliant on high-frequency spectral features for localization than others. As the high-frequency spectral cues derive from the physical features of the outer ears, subjects with smaller ears (conchae in particular) will produce spectral cues at higher frequencies than subjects with larger ears. This will mean that they are more affected by low-pass cut-offs in the higher frequency regions. Indeed the two subjects who were most affected by the 8 kHz low-pass filtering in these experiments (S1 and S5) were the only two females in the group and had smaller ears than the males.

One of the aims of this work, and the work of others, was to define the features important in optimizing the spatial perception of speech. It has been shown here that the inclusion of high frequencies, which are so often filtered out of speech in playback situations, can greatly improve the spatial perception. The reason this is thought to be important is because spatial perception plays an important role in competing source situations. However, there is also some evidence that the preservation of high-frequency speech information can be of benefit in nonspatial tasks. For example, it has been shown that children and adults require an upper cut-off frequency of 9 kHz to optimally identify fricatives in quiet (Stelmachowicz et al., 2001). Furthermore, the intelligibility of vowel–consonant–vowel sounds presented monaurally (to listeners with mild high-frequency hearing loss) was shown to increase steadily with the low-pass cut-off of the speech and performance did not plateau by the maximum tested cut-off of 8 kHz (Vickers et al., 2001).

TABLE IV. Percentage of COC errors made by each of the five subjects in Experiment 2 (and the mean). The first three rows contain values for the three new stimulus conditions and the fourth row is a reiteration of the data from the low-pass condition of Experiment 1. Each value represents the percentage of trials (out of the total 380) in which a COC error was made. See the text for details of this statistic.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB speech</td>
<td>13.4</td>
<td>3.4</td>
<td>5.3</td>
<td>6.1</td>
<td>5.3</td>
<td>6.7</td>
</tr>
<tr>
<td>20 dB down</td>
<td>20.8</td>
<td>6.6</td>
<td>10.8</td>
<td>10.8</td>
<td>10.8</td>
<td>12.0</td>
</tr>
<tr>
<td>40 dB down</td>
<td>23.4</td>
<td>6.6</td>
<td>12.6</td>
<td>13.9</td>
<td>12.6</td>
<td>13.8</td>
</tr>
<tr>
<td>LP speech</td>
<td>22.6</td>
<td>11.6</td>
<td>15.5</td>
<td>16.6</td>
<td>15.5</td>
<td>16.4</td>
</tr>
</tbody>
</table>

FIG. 10. (a) Mean absolute lateral angle errors from Experiment 2. The six clusters of bars show results for the five subjects as well as the mean across subjects. Mean errors are shown for broadband speech (black bars), 20 dB down low-pass speech (dark gray bars), and 40 dB down low-pass speech (light gray bars). Errors for low-pass filtered speech (60 dB down) from Experiment 1 are also shown for comparison (white bars). Error bars show standard error of the mean. (b) Mean absolute polar angle errors from Experiment 2. All other details as for (a). The asterisk indicates that for the mean data the three new conditions were significantly different from each other ($p<0.05$).