Prediction of virtual sound source elevation improved by including input source spectral shape in the prediction equation

Ella MANOR[1]; William L MARTENS[1];

[1] University of Sydney, Australia

ABSTRACT

Presented with familiar broadband sound sources, human listeners are expected to base reports of source elevation primarily upon spectral variation due to directional dependence of the head-related transfer function (HRTF). Using four recorded speech sounds as sources, this study tested whether reported elevation of virtual sound sources might be systematically influenced by the spectral shape variation that naturally occurs in human speech. A set of 36 virtual sound stimuli were created by processing those four sound sources with nine binaural HRTFs collected at angles spaced at 10-degree intervals on the frontal plane, ranging from 40 degrees below the horizontal plane to 40 degrees above. Perceived source elevation was reported via a graphical user interface by each of five listeners who showed substantial individual differences in response distribution. Overall the collected reports were not highly correlated with the angles at which HRTFs had been measured; however, after the perceptual data were adjusted by subtracting from each source elevation report the overall mean elevation reported by each listener, these centred elevation angles were well predicted by a regression equation that included two terms. The first term was the elevation angle at which the binaural HRTFs were measured, while the second term captured source spectral bandwidth variations between the four sound sources processed using the nine HRTFs.

Keywords: Speech, Elevation judgment, Spectral variation, HRTFs I-INCE Classification of Subjects Number(s): 01.9

1. INTRODUCTION

It is indisputable that the pinnae provide acoustical information important in enabling listeners to detect sound source elevation. The classic example of this is that localisation accuracy on the median plane is greatly reduced if the cavities of the pinnae are occluded (1). It is also widely accepted that spectral cues due to the convolutions of the pinna are responsible for determining elevation of sources well removed from the median plane, such as those on sagittal planes offset from the median. It may be that the pinnae play a smaller role in the lateralisation of sound sources arriving from directions arrayed on the horizontal plane, where inter-aural level difference (ILD) and inter-aural time difference (ITD) provide effective cues for positioning sources along the inter-aural axis (the imaginary line passing through the listener's ears). Early models of the binaural cues to direction explained the lateralisation of tonal stimuli on the basis of these simple intensive and temporal differences between the ears (see, for example, (2)). But for sound sources arriving from directions above or below the horizontal plane, the spectral features of a listener's head related transfer functions (HRTFs) are held to provide crucial information for that listener to distinguish between source elevation angles for which the same ILD and ITD are presented at the listener's ears.

But how do listeners know what spectral variation is present between speech sound sources before those sources are subjected to HRTF-based processing? It is a mystery that a listener can only resolve when that listener makes assumptions about which spectral variation is due to the HRTFs and which spectral variation exists between the processed sources (3). Since in the current study the HRTF-processed sources included only recorded human speech sounds, the source spectral variation should have been familiar to the listeners participating, and therefore they might be expected to base their reports of source elevation primarily upon spectral variation due to directional dependence of the employed HRTFs. The extent to which the source spectral variation influences elevation reports may reveal the manner in which spectral details contribute to the

[1] ella.manor@sydney.edu.au

formation of spatial auditory images under constrained conditions, which are more representative of a use case for binaural technology than short bursts of white noise (the stimulus of choice in many localisation studies).

Previous studies have shown that the apparent elevation of a virtual sound source will be more closely matched to its associated HRTF measurement elevation as the source bandwidth is increased and includes components above 7 kHz. This spectral dependence is distinct from the dependence of source elevation on pitch, in which high-pitched sounds usually are reported by listeners to be localised high in space, whereas low-pitched sounds are perceived as originating lower in space (4). If the auditory spatial imagery resulting from the headphone simulation includes cues essential to elevation perception, then the reported elevation angles should match the spatial relationships between the stimuli. Of course, the listeners may not all agree closely on the absolute direction of the spatial images they experience when listening to the simulation. Indeed, there is no set of absolute elevation locations at which the images should be experienced, given that each listener has their own HRTFs that may be more or less similar to the generic HRTFs used in this study (which were HRTFs measured using the manikin called KEMAR). But if KEMAR's HRTFs encode directional information in a manner characteristic of the anthropometrically median human male (5), then reported elevation angles should be highly correlated with the angles at which KEMAR's HRTFs were measured, monotonically increasing with measured elevation, even if offset a bit due to difference in pinna-size between KEMAR and the listener.

Of course there are clear differences in interaural spectra that might be used to distinguish between elevations, and these are not affected by source spectral variations. Also, there is a strong ipsilateral increase in HRTF gain at 8 kHz that is associated with a shifting of the sound source above or below ear-level that has been shown to bias elevation reports (6), but this cue requires the source to have appreciable energy in the 8 kHz region. The speech sound sources employed in the current study do not have much energy above 5 kHz. Therefore, it will be the spectral cues to elevation that exist below 5 kHz that will be of most interest here. As the four speech sound sources selected for the current study differ in their first two formant frequencies in this lower frequency region, it is possible that these spectral peaks will align with spectral cues that are effective in discriminating between elevations. Alternatively, it could be the overall 'spectral tilt' variations that listeners will use to make distinctions between source elevations (7). Blauert (6) hypothesised that pinna cues could be evaluated by the auditory system in terms of "directional bands" that map spectral patterns into perceived direction, and the importance of these spectral cues is most significant in providing directional information. Humanski & Butler (8) examined the "quality" of HRTF spectral features that could be contributing to elevation identification by measuring the correlation between HRTF spectral-band magnitude and direction. A high correlation indicated that there was low ambiguity in the relation, and hence the magnitude in that band could be taken as a good indicator of direction, regardless of the magnitude of other bands in the same HRTF. They argue that a band magnitude that is not an "overt" peak relative to adjacent bands in a given HRTF may nonetheless be a "covert" peak when compared to bands at the same frequency in spatially adjacent HRTFs. In fact, from a combined analysis of behavioural and spectral data, they conclude that "covert" peaks contribute more to median plane localisation performance than do "overt" peaks.

The current study employed spectral moments analysis (9), which is a statistical method used to analyse the power spectrum of a signal and to provide information regarding overall spectral shape using the Fourier transform. In digital signal processing the central moments are derived from the spectral moments, providing statistical interpretation for the different moments. Spectral moments consist of four moments, including the centroid, bandwidth, skewness and kurtosis. The first moment is derived from the frequency-weighted mean of the critical-band distribution also known as the spectral centroid, which is often associated with the 'brightness' of the sound (10). The second moment is derived from the dispersion centred at the spectral centroid, providing information of how wide or narrow the bandwidth of the spectrum is. The third moment, known as skewness, is the measure of asymmetry in the distribution. The fourth moment, known as kurtosis measures the width of the peak in the distribution (i.e., the distribution's "peakedness" or "flatness"). The spectral centroid and bandwidth are considered by the majority of similar studies to provide the most important power spectra characterisation of speech type signals (9). Nonetheless, all four spectral moments of the stimuli were analysed using stepwise regression, along with the measured elevation angles as a primary predictor.

By limiting the range of sound source directions to those on a restricted portion of the frontal plane, this investigation attempted to isolate the spectral cues involved in elevation perception. In this study subjects were asked to report their perceived elevation judgments for speech stimuli that were processed to simulate source arriving from various elevation angles, ranging between -40 degrees and 40 degrees on a constant azimuth angle of 90 degrees (meaning directly from the left side). The first question of interest is that regarding

whether sound sources that differ in spectral features are localised at different elevation angles despite the fact that they are transformed by identical HRTFs? If the majority of subjects reports similar trend of reported elevation, then perhaps this particular evaluation is not only dependent upon the use of a listener's "own" HRTFs, but indicates that spectral features associated with the stimuli influence perception of elevation for listeners in general. Of course, it is of even more interest in the current study which spectral features of the binaural stimuli will influence elevation reports as the HRTF-processed source spectra vary, particularly at the low frequencies within the vocal formant range. While other studies have shown the influence of reducing the spectral bandwidth of the processed sources (e.g., (11)), the current study focuses only upon broadband sources that vary in spectral shape. The findings of this study will be useful to any researchers looking for an explanation as to why elevation localisation errors occur (particularly when using generic HRTFs) and how simulated sound sources might be manipulated in binaural processing to prevail over these errors.

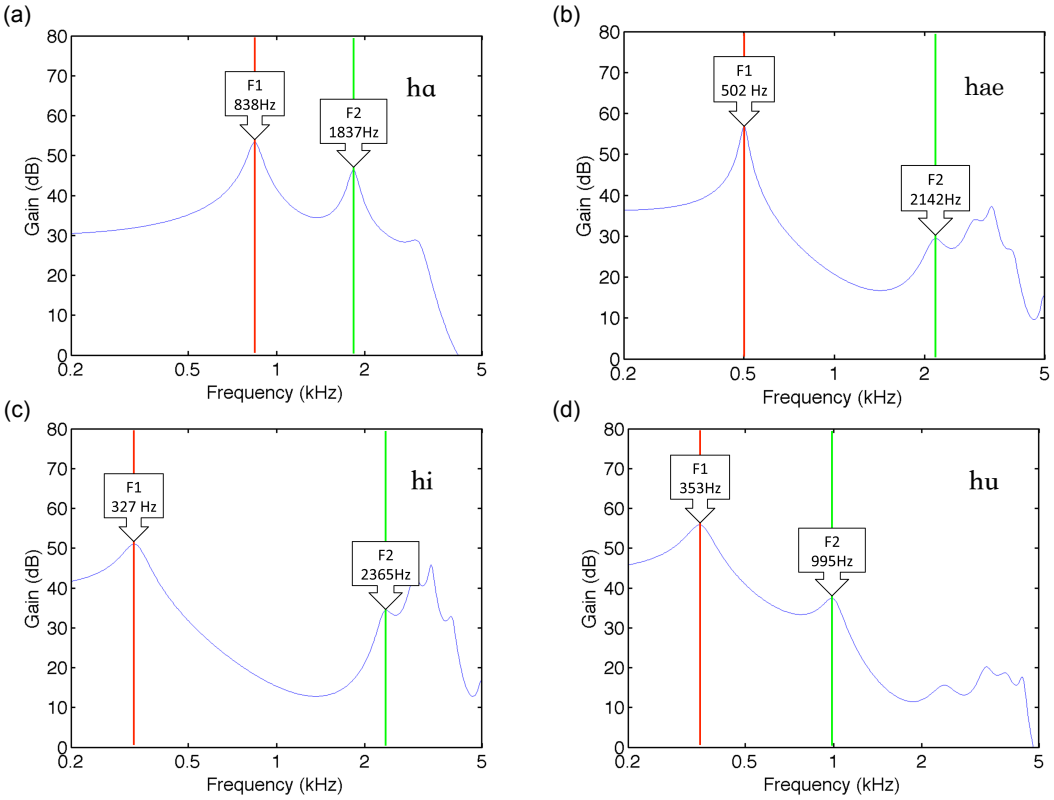

Figure 1 – LPC magnitude spectra for the four CVs used as sound sources in the current study. The smooth magnitude functions were the result of a 50-pole LPC analysis. Each panel shows the magnitude response in dB as a function of frequency in Hz, where the formants are labeled and colour coded (red line for first formant, green line for second formant). A comparison between the four CVs analysis results (as shown in panels a, b, c, and d) demonstrates a variation in spectral shape and in formants frequencies.

2. METHODS

2.1 Stimuli

The stimuli used in the experiment were short-duration speech sounds recorded in an anechoic chamber and reproduced over headphones after digital processing. The speech sounds were American English consonant-vowel (CV) syllables spoken by a male, and were digitally recorded using an omni-directional microphone placed at a distance of two metres from the talker. The four 16-bit audio recordings were obtained at a sampling rate of 44.1 kHz. Each CV began with /h/ and terminated with one of four vowel sounds. The vowels were /ɑ/ as in "hot", /æ/ as in "hat", /i/ as in "heat", and /u/ as in "hoot". The duration of each CV was less than 500 msec. The duration of the signal stored in the digital audio data file for each was exactly 500 msec and the original audio data was gated with a raised cosine to give a rise time of 5 msec and a fall time of 50 msec. Spectrum analysis of these recordings demonstrates the spectral variation that naturally occurs in human speech due to formants. Consequently, the /hi/ and /hæ/ CVs sounded brighter than the /hɑ/ CV, and these three together sounded considerably brighter than the /hu/ CV. The long-time-average LPC spectra of the

CVs are illustrated in Figure 1. These smooth magnitude functions were the result of a 50-pole LPC analysis intended to reveal the resonant structure for the CVs while glossing over the fine spectral detail of the periodic glottal source. It should be clear that most of the energy in these sound samples exists below 5 kHz. Indeed, an examination of the LPC spectra at higher frequencies than those shown in Figure 1 confirms that all employed vowel sound sources reached at 8 kHz a level at least 60dB below their peak level, which always occurred at the first formant frequency.

2.2 Processing

The set of stimuli comprised the factorial combination of nine elevation angles with four sound sources. In effect, each of the four CVs (/hɑ/, /hae/, /hi/, and /hu/) was transformed by each of nine HRTFs measured at elevation angles spaced at 10-degree intervals on the frontal plane (-40, -30, -20, -10, 0, 10, 20, 30, and 40 degrees, all at 90 degrees azimuth). This processing resulted in a set of 36 virtual sound stimuli to be presented over STAX SRλ earspeakers to each listener, with the response of the earspeakers equalised at the HRTF measurement position on KEMAR (as described in (11)).

2.3 Procedure

Five listeners reported the apparent elevation of 36 headphone-presented virtual sound source stimuli using a graphical user interface (GUI). Listeners were seated in a darkened, sound-treated room, approximately one metre from a computer monitor. Listeners were asked to report elevation angles through mouse-based interaction with a GUI consisting of a pointer superimposed upon a circle intended to represent a projection of the outline of the head. The pointer was a continuously adjustable line segment that was fixed at one end at the centre of the circle and extended radially to indicate the apparent elevation of the auditory image relative to the listener's ear level (for which a horizontal line was the reference). The listener could replay each stimulus indefinitely until satisfied that the position of the pointer accurately reflected the apparent elevation of the auditory image, at which time the response was recorded and the next stimulus was presented after a one second silent interval.

3. RESULTS

The mean elevation angles reported by all listeners are plotted in Figure 2 as a function of the measured elevation angles of the transformed HRTFs. If the listener had heard the CVs to be located at elevation angles that precisely matched the elevation angle of the measured HRTF, then all the symbols would lie on the black dashed diagonal line. Given the individual differences observed in the pattern of results for the five subjects tested, it would be of value to collect data from a larger group of subjects. Although such individual differences are commonly observed in elevation reports given the use of generic HRTFs due to typical variation in pinna size, such differences can be corrected by scaling of the HRTF to better customise the deployed HRTFs (see, for example, (12)). In lieu of scaling the HRTFs, the mean reported elevation was subtracted from each subject's data to bring mean reports into alignment across subjects, while maintaining the distribution of reports for each. The reported elevation curves obtained in all four CV conditions generally increase as a function of the measured HRTF elevation angle. By examining the four plots of Figure 2 it can be noted that data obtained for the CV containing the /hi/ vowel sound generally lie higher on the graph than data obtained for the other three CVs. This finding suggests that there may be some support for the assumption regarding the influence of source spectral features on the perception of virtual sound source elevation.

Furthermore, data obtained for the darkest CV containing the /hu/ sound generally lie lower on the graph than data obtained for the other CVs. Although further statistical analysis revealed that the majority of listeners were biased towards giving elevation reports for sources containing different vowel sounds, an ANOVA run on the raw reported elevations did not show this effect to be significant. A univariate ANOVA was performed on the elevation reported data from all five listeners. The effect of HRTF (measurement elevation) on reported elevation was significant ($F = 139.97$, $p<.001$). The effect of Source (CV) on reported elevation did not attain significance ($F = 0.35$, $p = .790$). Reported elevation also was not modulated across the repeated measures ($F = 0.13$, $p = .971$), and finally, the HRTF by Source interaction was not significant ($F = 0.17$, $p = 1.00$). Though there seemed to be a bias for the majority of listeners that might have shown up as a Source effect, the ANOVA yielded a very small F value, and the correspondingly high probability of incorrectly rejecting the null hypothesis. The primary reason the null cannot be rejected here is that there were substantial offsets between the mean reported elevation by each listener, and therefore the standard deviation about the mean for each of the mean elevation was typically large enough to include all the other means at a given elevation level. Though the linear trend is quite strong in each individual curve, the variation between listeners is too

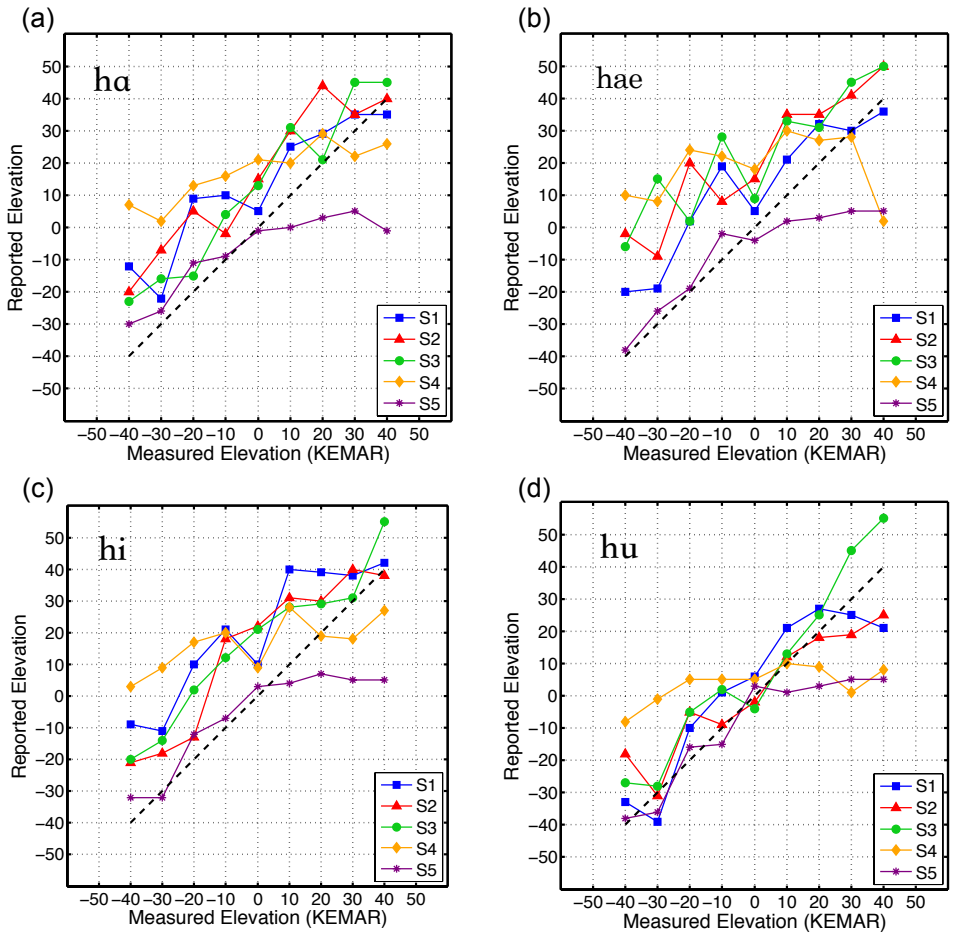

Figure 2 – Mean reported elevation angles for all CVs plotted using colour-coded symbols representing data for each listener (S1 to S5 in the legend). Panels a, b, c, and d show respectively data for CVs /hɑ/, /hae/, /hi/, and /hu/, where the black dashed diagonal line represents the veridical mapping of elevation angle from the measured HRTF angles to the reported angles.

great to allow a conclusion that source spectral variation that is introduced by the various CVs are given higher raw elevation ratings (the power of the test for a Condition effect was only .10). Nonetheless, if the elevation reports obtained from each listener are centred in order to remove these individual offsets, which offsets are naturally associated with differences in pinna size (12), the significant main effect of Source is revealed.

Further analysis that might explain the variation in listeners' reported elevation angles for each CV examines the spectral moments that characterise the CVs stimuli. Spectral moments consist of the spectral centroid, bandwidth, skewness, and kurtosis associated with a signal's frequency spectrum. The CVs processed with HRTFs to simulate sound sources at particular elevation angles were submitted to spectral moments analysis. Figure 3 consists of two plots showing the spectral centroid (panel a) and bandwidth (panel b) as a function of elevation angles of the measured HRTFs, where CVs are symbol coded. Spectral centroid was calculated for each processed HRTF of the four CVs using the following equation:

$$Centroid = \sum_{n=0}^{N-1} \frac{|X|(n)}{|X|} * n \tag{1}$$

Where X is the magnitude response of the signal n at N frequencies. Then, spectral bandwidth, also known as variance, as implemented in the MATLAB routine *spectralFeatures*, was calculated for each processed HRFT of the four CVs, based on the calculated spectral centroid:

$$Bandwidth = \sum_{n=0}^{N-1} \sqrt{\frac{|X|(n)}{|X|} * (n - Centroid)^2} \tag{2}$$

The third moment, skewness, was calculated based on the cubed spectral centroid and the cubed bandwidth:

$$Skewness = \frac{\sum_{n=0}^{N-1} \frac{|X|(n)}{|X|} * (n - Centroid)^3}{Bandwidth^3} \quad (3)$$

Similarly, the forth moment, kurtosis, was calculated based on the biquadratic spectral centroid and the biquadratic bandwidth:

$$Kurtosis = \frac{\sum_{n=0}^{N-1} \frac{|X|(n)}{|X|} * (n - Centroid)^4}{Bandwidth^4} \quad (4)$$

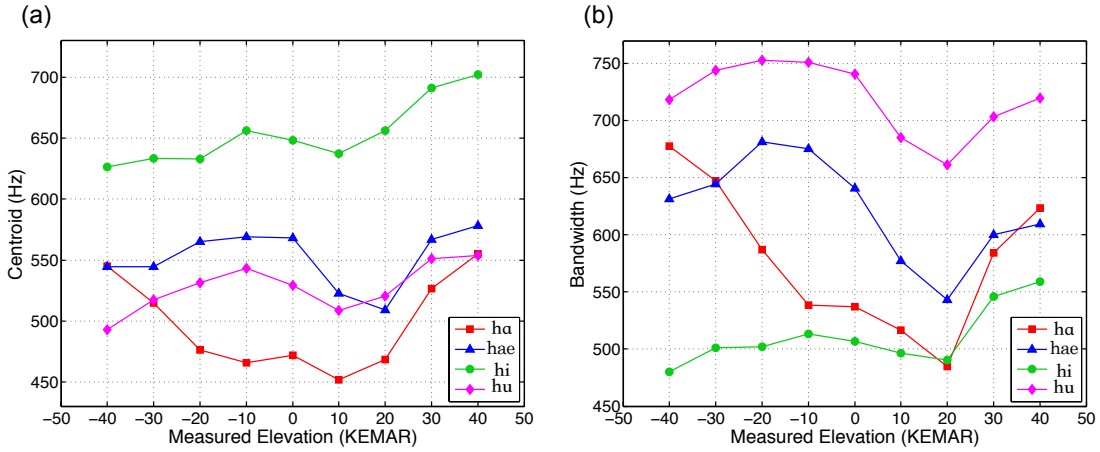

Figure 3 – Spectral moments analyses showing the spectral centroid (panel a) and bandwidth (panel b) as a function of elevation angles of the measured HRTFs. The data points resulting for each CV are colour and symbol coded, as presented in the legend. Note that the CV containing the /hi/ sound exhibited the highest centroid values but the lowest bandwidth values.

Interestingly, the CV /hi/ yield significantly higher frequency centroid and narrower bandwidth rate of all other CVs. Examining this finding in relation to the tendency of listeners to report higher elevation angles for this CV, it is clear that bright and narrow bandwidth stimuli are more difficult to be localised accurately by listeners. Contrasting connection can be drowned by observing the low centroid and wide bandwidth values obtained for the CV /hu/, meaning this type of stimulus is more easily localised by listeners accurately. Furthermore, a comparison between the curves resulting for each of the CVs from the centroid analysis and those curves resulting from the bandwidth analysis show similar trends.

Figure 4 shows two boxplots resulting from the skewness (panel a) and kurtosis (panel b) of each of the CVs for all elevation angles of the measured HRTFs. Each boxplot represents the values of skewness and kurtosis for all measured elevation angles, where the red line inside the box shows the median inter-quartile range and the edge of the box shows the range within which the middle 50% of the distribution could be found. By examining panel a of Figure 4 it can be concluded that the bandwidth distributions of CVs containing /hɑ/ and /hae/ are asymmetric and shifted to the right, where the distribution of the /hu/ CV is asymmetric and shifted to the left. The distribution of the /hi/ CV is has a skewness value like that of the normal distribution. The width of the boxplots indicates the shape of overall distribution, where a wide box means a wide bandwidth and a narrow box means a narrow bandwidth. Using the kurtosis analysis, the width of peaks, also known as "peakedness", is measured in relation to the bandwidth distribution. By examining panel b of Figure 4 it can be seen that the CV containing /hɑ/ and /hu/ exhibit narrower "peakedness" than that of the /hi/ CV, and the /hae/ CV exhibits the widest peaks in its bandwidth distribution. These spectral moments measurements characterise well the LPC magnitude spectra of the CVs shown in Figure 1.

Initially, the raw reported elevation data was tested against the measured elevation angles of the processed HRTFs using a multi-linear regression analysis. The results of this analysis can be seen in panel a of Figure 5, where the raw reported elevation angles are plotted as a function of the measured elevation using blue dots along with the red dashed line indicating the veridical match. The coefficient of determination, denoted R^2, shows that the fit accounted for 55% of the variance. An F-test was conducted to examine the significance of the relationship between the response (i.e., raw reported elevation) and the predictor (i.e., measured elevation) .

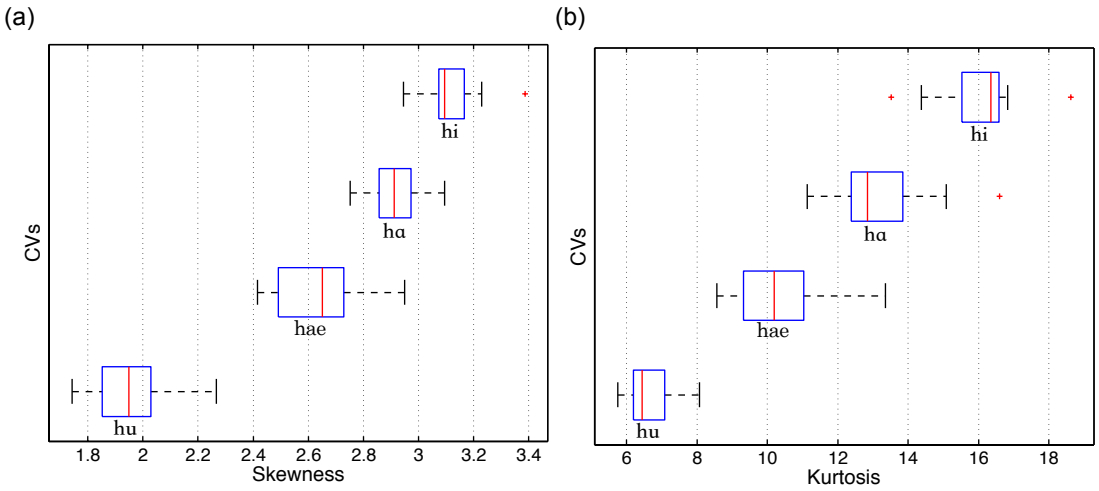

Figure 4 – Spectral moments analyses showing the spectral skewness (panel a) and kurtosis (panel b) in boxplots for all measured elevation angles. The red line inside the box shows the median and the edge of the box shows the inter-quartile range (within which the middle 50% of the distribution could be found). Each boxplot is labelled with its associated CV, and is positioned on the y-axis from largest to smallest mean bandwidth.

The F-test showed a significant relationship between the reported and measured elevation (F = 221.54, p<.001), meaning that measured elevation should be included as a predictor in the stepwise regression. The raw reported elevation data was centred as an attempt to increase the goodness of fit between the response (i.e., the centred elevation reports) and the predictor (i.e., measured elevation). As expected, the centred reported elevation (CRE), when fit to the measured elevation, significantly increased the coefficient of determination (R^2 = 0.67), as well as providing a higher F-value of 358.16 with p<.001.

As an attempt to find more predictors of the CRE, the four spectral moments calculated for all stimuli were submitted to stepwise regression analysis, along with the measured elevation. A stepwise regression analysis provides statistical prediction information by including or excluding predictor variables in single steps. A total of five predictors were submitted to the stepwise regression analysis, including the measured elevation, centroid, bandwidth, skewness, and kurtosis. The first two predictors submitted to the stepwise regression are the measured elevation angles and the spectral centroid measured for each of the CVs. The coefficient of determination result for this regression is very similar to when only the measured elevation is compared against the CRE (R^2 = 0.67), and the relationship between the response and the predictors is significantly decreased (F = 181.35, p<.001). This finding proves that spectral centroid should be excluded as a predictor in the next step of the regression analysis, in which the measured elevation (as proven to contribute to the prediction model) and bandwidth are included. This analysis yield what at that point was assumed as the prediction model producing the best regression fit (R^2 = 0.71, F = 213.92, p<.001) and can be observed in panel b of Figure 5. However, to verify this assumption, it is necessary to complete the stepwise regression. In the next two steps, skewness and kurtosis were included in the model, resulting in a slightly higher coefficient of determination (R^2 = 0.72) and a lower statistical significant effect (F = 112.24, p<.001). The final step tested whether a triple-model predictor of measured elevation, bandwidth, and kurtosis might result in a better regression result, however this model does not provide a better prediction (R^2 = 0.71, F = 143.49, p<.001) than that based upon just the measured elevation and bandwidth as predictors. The prediction equation, with regression coefficients (β values) can be defined as follows:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_3 x_3 \tag{5}$$

Where β_0 is y-intercept, x_1 is the measured elevation data, and x_3 is the calculated bandwidth.

Panel a of Figure 6 shows the predicted elevation angles for each of four CV sources separately as a function of the measured HRTF elevation angles, based upon the two-term regression model that is plotted in Figure 5. The black dashed line represents the relation expected for veridical perception, in which the reports match to the original elevation angles. Panel b of Figure 6 shows the mean reported elevation angles for each of the four CV sources separately as a function of the predicted elevation angles, where the x-axis uses the $\hat{y}$

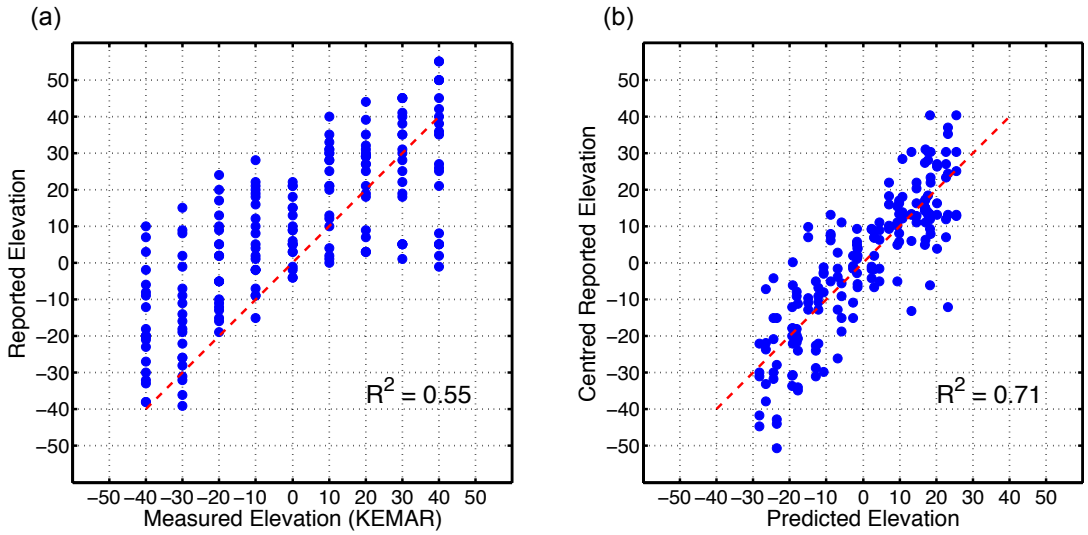

Figure 5 – Panel a shows the mean of the reported elevation angles for each of four CV sources for all five listeners as a function of the measured elevation angles of the HRTF-processed CVs. The red dashed line represents the relation expected for veridical perception (i.e., reports matching original elevation angles). Panel b shows the same reported elevation angles, but centred by subtracting the mean elevation reported at each measured elevation for each listener. Note that the x-axis in panel b is the result for the two-term regression analysis (the ŷ values from the model that included bandwidth as a predictor, in addition to the measured angles). In both cases, the coefficient of determination (denoted R^2) is indicated in the lower right corner.

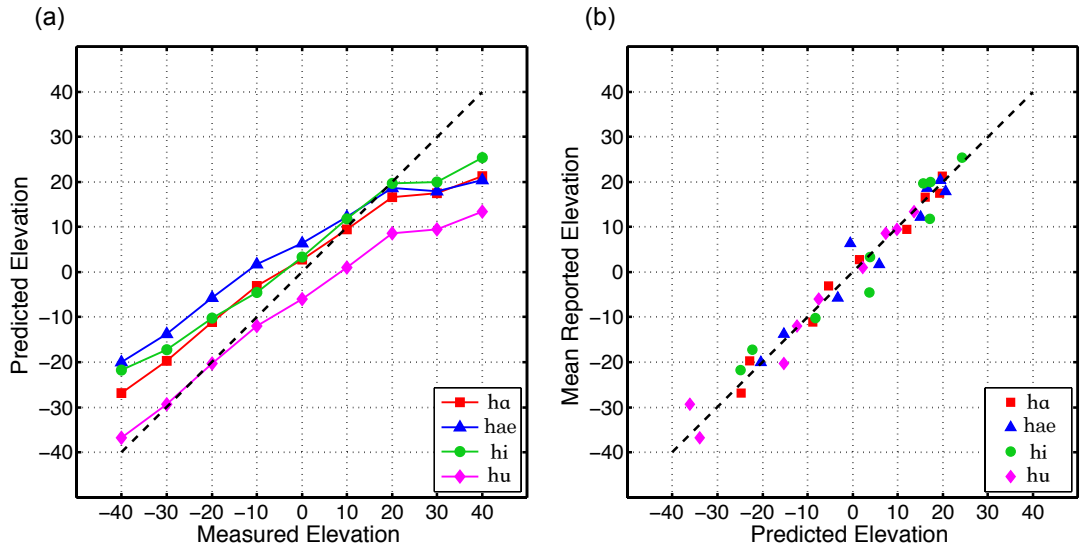

Figure 6 – Panel a shows the predicted elevation angles for each of four CV sources separately as a function of the measured HRTF elevation angles, based upon the two-term regression model. The symbols and colours associated with each CV are the same as in Figure 3, and are specified in the legend. The black dashed line represents the relation expected for veridical perception (i.e., reports matching original elevation angles). Panel b shows the mean reported elevation angles, again for each of four CV sources separately, but as a function of the predicted elevation angles for each of four CV sources; thus the x-axis in panel b uses the values resulting from the two-term regression analysis (the ŷ values from the model that included bandwidth as a predictor, in addition to the measured angles).

values from the model that included both the measured elevation angles and bandwidth as a predictors of the two-term regression analysis.

4. CONCLUSION

The results of the current study shed some light upon the mystery of how listeners might resolve the dilemma of distinguishing between the trial-to-trial spectral variation between sound sources, and the trial-to-trial spectral variation due to HRTF-based processing which should supposedly allow them to identify virtual sound source elevation. Although the four input sound sources presented here were familiar sounds exhibiting spectral features that naturally occur in human speech, the reported virtual source elevation angles were not those of an optimal observer, whose reports should be unaffected by input source spectral variation. Rather, a systematic bias in elevation reports was observed which indicated that listeners failed to attend exclusively to the directionally dependent spectral variation imposed by the HRTF-based processing. Furthermore, the dependence of elevation reports upon a particular spectral feature of the presented stimuli was revealed through stepwise regression analysis, so that prediction of elevation reports could be improved significantly over predictions based only upon the elevation angle at which the employed HRTFs had been measured. The conclusion must be that listeners engaged in such localisation tasks rely upon spectral features of the sound signal presented to their ears without resolving whether trial-to-trial spectral variation is due to differences between input sound sources or differences between the HRTFs by which those sound sources are processed, at least for sources that originate from within the frontal plane.

With regard to the particular spectral features of the presented stimuli that influenced the obtained elevation reports, stepwise regression analysis revealed which of four spectral moments provided the best measure of the stimulus spectral variation for predicting variation in elevation reports. While systematic variation was observed between stimuli for all four spectral moments, which included the centroid, bandwidth, skewness and kurtosis of the spectra, it was the second moment, bandwidth, that provided the greatest improvement in a two-term regression equation: The first term was the elevation angle at which nine binaural HRTFs were measured, while the second term captured source spectral bandwidth variations for the four sound sources processed using those nine HRTFs.

Whereas it had been suggested previously (11) that spectral centroid might provide a good measure for explaining the observed dependence of elevation reports upon input source spectra, the current results support the conclusion that variation in source spectral bandwidth provides a better measure. These results have both theoretical and practical implications. First, the results make it clear that listeners are not able to distinguish perfectly well the spectral variation between sound sources versus that due to HRTF-based processing. Secondly, these results suggest that an adaptive procedure that includes source information regarding spectral variation might produce more accurate positioning of virtual sound sources in practical applications.

REFERENCES

1. Gardner MB, Gardner RS. Problem of localization in the median plane: effect of pinnae cavity occlusion. The Journal of the Acoustical Society of America. 1973;53(2):400–408.

2. Woodworth RS. Experimental Psychology. H. Holt; 1938.

3. Blauert J. Spatial Hearing: The psychoacoustics of Human Sound Localization. Cambridge, MA: MIT Press; 1983.

4. Butler RA. Does tonotopicity subserve the perceived elevation of a sound?. vol. 33. Federation proceedings; 1974.

5. Burkhard MD, Sachs RM. Anthropometric manikin for acoustic research. Journal Acoustical Society of America. 1975;58:214–222.

6. Blauert J. Sound Localization in the Median Plane. Acoustica. 1969/1970;22:957–962.

7. Hiranaka Y, Yamasaki H. Envelope representation of pinna impulse responses relating to three-dimensional localization of sound sources. Journal Acoustical Society of America. 1983;73:291–296.

8. Humanski RA, Butler RA. The contribution of the near and far ear toward localization of sound in the sagittal plane. The Journal of the Acoustical Society of America. 1988;83(6):2300–2310.

9. Fulop SA. Speech Spectrum Analysis. Springer; 2011.

10. McAdams S, Winsberg S, Donnadieu S, De Soete G, Krimphoff J. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. Psychological research. 1995;58(3):177–192.

11. Martens WL. Directional Hearing on the Frontal Plane: Necessary and Sufficient Spectral Cues. Northwestern University, Evanston, Illinois; 1991.

12. Middlebrooks JC. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. The Journal of the Acoustical Society of America. 1999;106(3):1493–1510.