SPATIAL EXPLANATIONS OF SPEECH PERCEPTION: A STUDY OF FRICATIVES

Won Choo & Mark Huckvale
Department of Phonetics and Linguistics, University College London, U.K.

ABSTRACT

This paper addresses issues of perceptual constancy in speech perception through the use of a spatial metaphor for speech sound identity as opposed to a more conventional characterisation with multiple interacting acoustic cues. This spatial representation leads to a correlation between phonetic, acoustic and auditory analyses of speech sounds which can serve as the basis for a model of speech perception based on the general auditory characteristics of sounds. The correlations between the phonetic, perceptual and auditory spaces of the set of English voiceless fricatives /f θ s ʃ h/ are investigated. The results show that the perception of fricative segments may be explained in terms of 2-dimensional auditory space in which each segment occupies a region. The dimensions of the space were found to be the frequency of the main spectral peak and the 'peakiness' of spectra. These results support the view that perception of a segment is based on its occupancy of a multi-dimensional parameter space. In this way, final perceptual decisions on segments can be postponed until higher level constraints can also be met.

INTRODUCTION

One of the major issues in the study of speech perception process is the listener's capacity to achieve *perceptual constancy* in spite of the vast amount of acoustic variability that exists in the speech signal (Nygaard & Pisoni, 1995). The task of identifying "segments" (or any other phonological units) in stretches of transient acoustic waveform has proven extremely difficult.

The traditional approach to this problem has been extraction of perceptually salient acoustic cues through comparisons of minimal pairs. These studies have made significant contributions in understanding detailed structure of acoustic signals, but at the same time, have left the legacy of a large number of interacting cues associated with each unit of meaningful contrast.

An alternative approach is to define a segment by the *space* which it occupies at each level of the perceptual system, from signal to percept. Rather than focusing on individual acoustic characteristics of segments which identify them, we view the perception of segments in the perspective of a multi-dimensional parameter space in which different segments occupy different regions. This leads to a more 'holistic' approach to speech perception where identity of each segment can be defined in relation to other segments. The axes of this space are simply the measurements we make of a segment in each level of perceptual description and a point in this space is a combination of a set of these measurements over a given time interval. Realisations of segments cluster in this space as points or trajectories, and different segments cluster at different locations.

The advantage of defining segments in the perspective of parameter spaces is that interactions between different levels of speech processing can be addressed/quantified in terms of the correlations between different spatial representations. A simple correspondence between spatial representations at acoustic, auditory, perceptual, and phonetic levels, suggests a speech perception model based on

general characteristics of speech signal.

This approach also leads to a model of speech perception which is more readily combined with top-down constraints (Huckvale, 1996). That is, no definite decision about a segment needs to be made in the signal level; only a distinct region in the listener's auditory space is alerted, and the final perceptual decisions on meaning can be postponed until higher level constraints are also satisfied. After all, the task of speech perception is not to identify each individual segment, but to decode the message contained in the acoustic waveform.

Early examples of studies based on such an approach are by Pols et al. (1969) and Klein et al. (1970). They showed that the Principal Component Analysis (**PCA**; Harman, 1967) on outputs of 1/3 octave bandpass filtering of vowels leads to a three-dimensional physical space. These physical dimensions were not arbitrary and were closely matched to phonetic/articulatory dimensions of frontness and height as well as the acoustic dimensions of vowel formant values. They were also closely related to the perceptual dimensions revealed from Multidimensional Scaling analysis (**MDS**; Kruskal, 1964) of similarity judgments.

This paper is the first to extend this specific approach to consonant studies. The aim is to illustrate the possibility of a spatial approach to explanations of speech processing in consonant perception. Fricatives are used as convenient materials in which the knowledge of vowel studies can be transferred to the analysis of consonant sounds.

AUDITORY SPACE

Materials

Five male speakers of English in the 20-40 age group recorded the fricatives /f θ s ʃ h/ twice, followed by the vowel [ɑ]. These were digitized with 20 kHz sampling rate and 16-bit quantization. In addition, the overall amplitudes were normalised in terms of the mean RMS (root mean squared) values of the fricative section.

Analyses

The fricative spectra were analysed using 32-channel, 1/3 octave bandpass filters, to resemble the ear's critical bandpass analysis. The outputs in decibels in the 32 frequency channels from a spectral slice of each fricative produce a 5×32 data matrix. A spatial model of 5 fricative points in a 32-dimensional space will produce a perfect fit of the spectral data. The goal of auditory analysis, however, is to determine the minimal number of dimensions required to model the data with maximal variance in the data accounted. For this purpose, PCA could be used to reduce the data points to a few principal auditory components (dimensions). However, it may not be accurate to sample a particular spectral section since articulation of fricatives also change in time. A non-linear time alignment technique (Sakoe & Chiba, 1978) was used to account for the dynamic fluctuation of the fricative signal, and differences in the length between the different fricatives and speakers. The Euclidean distances between the aligned auditory spectra are calculated for each production of each speaker and used for 2-way MDS analysis. The object of this technique is to obtain an optimal spatial representation of the scaled objects on the basis of analysed distances.

Results

The results of MDS analyses for each speaker show that 2-dimensional solutions adequately account for the data. The variance accounted for, averaged over 10 productions, were .958 and .035 for dimensions 1 and 2 respectively. The resulting auditory spaces of each production of each speaker were plotted on the same axes as shown in Figure 1.

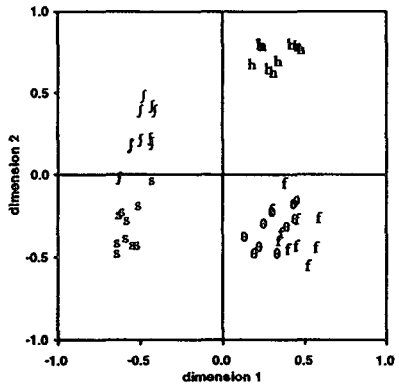

Figure 1. General auditory map of English fricatives based on 10 productions by 5 speakers.

According to this auditory map, two parameters explain most of the variations in the acoustic signal. Each fricative category occupies a separate region in the auditory space. There is an overlapping space between the fricative regions of /f/ and /θ/. This overlap may be explained by the similarity in the auditory spectra of the two fricative types or the close proximity between the fricatives in the perceptual space. This leads to investigations of the acoustic correlates and perceptual responses to each auditory region.

ACOUSTIC CORRELATES

In an attempt to identify acoustic correlates of each auditory dimension, the average spectral shape of each fricative type used in the last section is placed in the corresponding region of the fricative on the general auditory map in Figure 1. This is shown in Figure 2.

The average spectral shape was obtained in three separate stages. First, the output energy levels of each auditory filter were averaged across the whole length of each fricative segment. In order to accommodate the differences in the overall level of the fricative segments, the output levels of the 32 bands were reduced by the mean level of that particular production for each production of each fricative. These spectra were averaged over the ten productions spoken by the five different speakers.

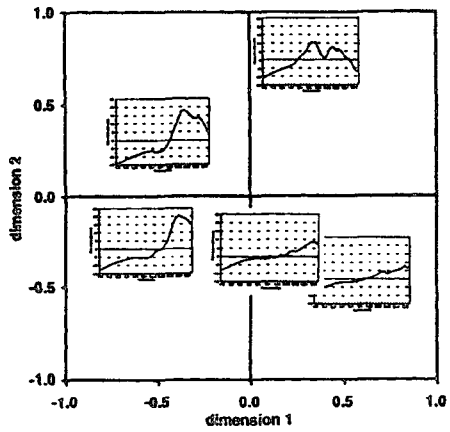

Figure 2. The average spectrum of each fricative is placed on the corresponding region of each fricative on the auditory axes.

It is noticeable that the spectral characteristics of /f/ and /θ/ are very similar; in both cases, the spectra are mainly flat. /s/ and /ʃ/ can be characterised by a single broad-band peak; however, the low cut-off frequency occurs a little higher for /s/ at around 3600 Hz, than /ʃ/, at 2000 Hz. For /h/, the spectral peaks occur at around 770 Hz and 2000 Hz, which correspond to the formant frequencies of the following [ɑ] vowel.

Overall, the auditory dimension 1, in Figure 2, may be related to the 'peakiness' of spectra — the maximum distance to mean amplitude — while dimension 2 may be related to the frequency of the main peak — the centre of gravity of the spectra.

PERCEPTUAL SPACE

Materials

The same set of fricatives was recorded by a speaker and digitised in the same way as above. Each fricative part was cut-out and paired with two other fricatives to make up a triad, ABC, which was in turn paired to AB AC to help the short term memory of the listeners. 30 (=5×4×3/2) triads were constructed and randomised. There were 0.1 sec of inter-stimulus and inter-stimulus-pair gap and 2 sec pause after two pairs were presented. The stimuli were generated and recorded onto a DAT tape with a pause and a tone every block of 5 stimuli for presentation to the listeners.

Subjects and procedure

Five native speakers of English listened to the stimuli and were asked to choose the more similar sounding pair out of the two pairs. Data were accumulated over trials by assigning 1 scores for the pairs selected as more similar, and 0 scores for the pairs which were not selected. In this way, a matrix of data indexing the perceived relationships among the five fricatives was obtained for each subject. An example of a subject's similarity matrix is given below:

	f	θ	s	ʃ	h
f	0	3	2	0	1
θ	3	0	1	2	0
s	1	3	0	2	0
ʃ	1	3	2	0	0
h	0	1	2	3	0

Analyses

The similarity matrices obtained from each of the listeners were typically not symmetrical as shown above. Thus, the square matrix option was used in MDS analysis (proc-ALSCAL program, SAS Windows version 6). Since more than one similarity matrices were involved, weighted MDS analysis was carried out. This technique, not only calculates the relative locations of the objects in a space, but also, calculates the relative weights that each subject places on a particular dimension in order to find an optimal orientation of the space. Results of 3-way, square matrix, ordinal level analysis (for details see Choo, 1996) are presented below.

Results

The badness-of-fit curve and the interpretability of spatial arrangements suggest that a 2-dimensional solution is most appropriate to model the data. This perceptual space is presented in Figure 3.

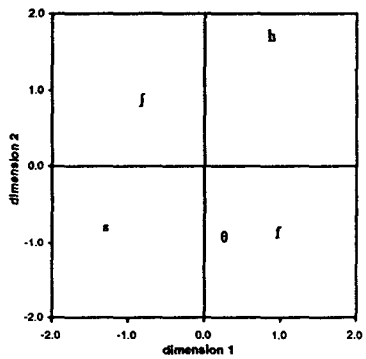

Figure 3. The 2-dimensional ordinal level solution obtained from perceptual similarity judgments of the cut-out fricatives.

We can observe a remarkable similarity between the perceptual and auditory maps of the fricatives. Furthermore, the perceptual dimensions can be readily related to the phonetic properties of fricatives; dimension 1, in Figure 3, corresponds to 'sibilance' and dimension 2 corresponds to 'place'.

Quantitative measurement of correlation was carried out by canonical correlation analysis. The average point of each fricative region on the auditory map was calculated and these average coordinates of auditory map were compared with those of perceptual map. The canonical

coefficients were .995 and .987 for each dimension respectively. This implies that the two spaces are highly correlated.

DISCUSSION

The main characteristic of the spatial representations discovered in the experiments is that two-dimensional solutions proved adequate in accounting for almost all the variations in the perceptual and auditory domains. This means that the parameters involved in perceptual similarity judgments and in spectral similarity of fricatives are reducible to two main components. Since there is a similar number of parameters involved in both domains, and also the correlation between perceptual and auditory spaces was high for fricatives, this leads to the strong conjecture that the two domains are closely related to each other.

As suggested in the introduction, the study of spatial representations enables us to identify key factors involved in each domain of the processing, and to demonstrate simple correlation across the different domains. This result stands in sharp contrast to the contemporary detailed cue studies in which many different spectral characteristics seem to be intricately interwoven and often interact in specifying the perception of any one fricative category.

The auditory and perceptual dimensions have clear phonetic interpretations, and are also related to concrete physical properties of fricative spectra. This is a novel finding, to the extent that spatial representations had never before been clearly established for consonants; as a consequence, the relationship between the spatial representations across the different domains had never been open to investigation.

The validity of spatial approach needs to be confirmed with respect to the phenomena of normalisation, coarticulation, and with other types of consonants, for example, the plosives. It also needs to be applied to other languages, with different phonological structures, in which the occupation of the space differs and the imputed criteria for identifying a segment may change.

REFERENCES

1. Choo, W. (1996) Relationships between phonetic perceptual and auditory spaces for fricatives. *PhD thesis*. University of London.
2. Harman, H. H. (1967) *Modern Factor Analysis*. The University of Chicago Press, Chicago.
3. Huckvale, M. (1996) Learning from the experience of building automatic speech recognition systems. *Speech, Hearing and Language; work in progress UCL* 9.
4. Klein, W., Plomp, R. & Pols, L. C. W. (1970) Vowel spectra, vowel spaces and vowel identification. *Journal of the Acoustical Society of America* 48. 999-1009.
5. Kruskal, J. B. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29. 1-27.
6. Nygaard, L. C. & Pisoni, D. B. (1995) Speech perception: New directions in research and theory. In Miller, J. L. & Eimas, P. D. (eds) *Speech, Language, and Communication*. Academic Press. San Diego.
7. Pols, L. C. W., van der Kamp, L. J. Th. & Plomp, R. (1969) Perceptual and physical space of vowel sounds. *Journal of the Acoustical Society of America* 46. 456-467.
8. Sakoe, H. & Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions. Acoustics, Speech, and Signal Processing*. 26. 43-49.