THE RELATIONSHIP BETWEEN PERCEPTUAL AND PHYSICAL SPACE OF FRICATIVES

Won Choo
Seijo University, Tokyo 157-8511, JAPAN

ABSTRACT

This study investigates the correlations between perceptual and physical spaces of voiceless English fricatives /f θ s ʃ h/, using Multidimensional scaling (MDS) technique. The spatial representations were constructed from perceptual similarity judgments and various spectral distance measures [1]. The results show that physical configuration based on Euclidean distance metric produces best prediction of the perceptual data. In both spaces, 2-dimensional solutions adequately accounted for the data and fricatives in the map were clearly organised in terms of their 'place' and 'sibilance' properties. The two dimensions were also closely correlated to the frequency of the main spectral peak and the 'peakiness' of spectra in their acoustic characterisation. The close correlation between linguistic characterisations of fricatives and their perceptual and physical configurations supports a model of speech perception based mainly on the general physical characteristics of speech.

1. INTRODUCTION

This study attempts to investigate the relationships between the perceptual and physical properties of fricatives and their linguistic classification.

Traditional approaches to such investigations have been based on the perceptual testing of acoustic cues hypothesised to be responsible for each phonetic contrast. Such studies have made significant progress in understanding the detailed structure of acoustic signals, but at the same time produce perceptual models which show great complexity in cue interaction associated with any contrast.

Instead, this study concentrates on similarity data, which can be used to build spatial models of each level of perceptual processing - acoustic signals, perceptual judgements, and phonetic contrasts. At each level of representation, a spatial map is used to indicate the relative location of units on measurement dimensions where distance is inversely related to the similarity. The phonetic units are recognised in terms of regions they occupy in such spaces with respect to the other units. This approach is consistent with models of speech perception such as the prototype theory, according to which the correct identification of speech segments depend on the perceived distances between speech stimuli and a prototype/region in perceptual space [2].

These similarity-based approaches are justified if there is a close correspondence between spatial representations at physical, perceptual, and phonetic levels (after appropriate acoustic metric analysis), and this provides evidence that the perceptual system could be operating in a very simple way; matching inputs to prototypes/regions with a similarity measure based on general characteristics of the speech signal. Conversely, if there is no match or poor match between these spaces, this implies that some other process is involved, for example; the distance metric may be inadequate, the primary auditory analysis (input) may be inappropriate or there may be some top-down processing involved.

Early examples of studies based on such an approach are by Pols et al. [3] and Klein et al. [4] using vowels. They showed that Principal Component Analysis applied to outputs of 1/3 octave bandpass filtering of Dutch 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 perceptual dimensions revealed from MDS analysis of similarity judgements.

In this paper, the same principle is applied to illustrate the possibility of a spatial explanation of consonant perception. Fricatives are used since they are known to be relatively steady-state and spectral characteristics are more important than the transitions. The next section establishes the spatial configurations of the perceptual similarity data. Section 3 is about multiple speaker production tests where the physical spaces of fricatives are estimated using various auditory distance metrics, and a non-linear time alignment technique [5]. In section 4, the acoustic correlates of the physical dimensions are identified. A general discussion of the results and the theoretical implications are given in section 5.

2. PERCEPTION TEST

2.1. Materials and procedures

The stimuli were fricatives /f θ s ʃ h/, read by a *female* native speaker of R.P., on a falling tone, followed by /ɑ/. The materials were recorded in an anechoic room onto a Sony DTC-1000ES digital audio tape recorder. They were digitised with a 20 kHz sampling frequency and 16-bit quantization, and transferred onto computer disk. Fricatives were excised and normalised in their intensity with respect to the RMS levels.

First, each stimulus was paired with two other stimuli, to make up a triad, ABC, which was in turn paired to AB AC to help the short term memory of the listeners. 30 (= 5x4x3/2) triads were constructed. Half of the stimuli were presented in the order AB AC while the other half was presented in the order AC AB. These were recorded back to DAT tapes. There was 0.1 second of inter-stimulus and inter-stimulus-pair gap and 2 seconds pause after two pairs were presented. There was a pause of ten seconds after each block of 5 similarity judgement pairs. After that pause the listeners were prompted by a tone for the next block.

20 students were paid to listen to the stimuli. All were native speakers of English, and reported normal speech and hearing.

Similarity data were accumulated such that the pairs selected to be more similar were assigned scores of 1, and the pairs which

were not selected scores of 0. In this way, a matrix of data indexing the perceived relationships among the five stimuli was obtained for each subject. An example of a subject's similarity matrix is given below:

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

Table 1. Similarity matrix for a subject who rated the fricative pairs.

2.2. Analyses

The similarity matrices obtained from each of the listeners were typically not symmetrical as in the example above. Thus, the square matrix option was used in the MDS analysis (pro-ALSCAL program, SAS Windows version 6). Since more than one similarity matrix was involved, weighted MDS (WMDS) analysis was carried out. Results of 3-way, square matrix, interval level analysis are presented below.

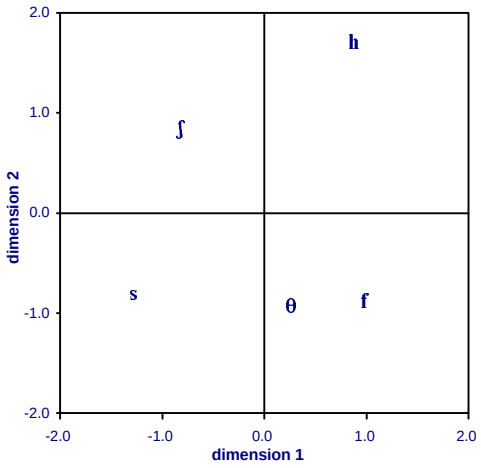

Figure 1. The 2-dimensional interval level solution obtained from perceptual similarity judgments of the isolated fricatives.

The badness-of-fit curve and the interpretability of spatial arrangements suggest that a 2-dimensional solution is most appropriate to model the data. Subject spaces in WMDS showed that subjects behaved consistently with no obvious outliers. As a further indication of stability of the stimulus configurations, stimulus spaces from two split-halves of subject data were also compared. The results showed that the MDS solution of one-half of the sample is similar to that of the other half, which means that the solution as whole is reliable.

Dimension 1 clearly separates the sibilants /s ʃ/, from nonsibilants, /f θ h/, while dimension 2 places the fricatives according to their place of constriction.

3. PHYSICAL ANALYSES

3.1. Materials

In contrast to the perceptual test, this time, five *male* native speakers of English in the 20-40 age group recorded the fricatives, /f θ s ʃ h/, followed by the vowel [ɑ]. They were asked to utter the syllables twice, clearly and in a falling tone. The recordings were made in an anechoic room onto a Sony DTC-1000ES digital audio tape recorder. They were digitised as before.

3.2. Analyses

The physical space is obtained in four main stages. Firstly, the spectra were processed by a simple 1/3-octave bandpass filtering to model filter bank analyses in the auditory periphery. The intensity axis is also transformed into a logarithmic scale, to reflect the non-linear loudness density pattern in the auditory periphery. The outcome is an auditory excitation pattern. 32-channel filters are used for the Euclidean metric analysis, while 64-channel filters are used for the slope and N2D metrics.

Next, spectral distances between these auditory excitation patterns are calculated with three different metrics - Euclidean, slope, and N2D metrics. The Euclidean metric takes the square root of the squared differences in the outputs of each filter between any two compared spectra. Thus, the physical distance between two spectra S_1 and S_2, can be expressed as:

$$EDM = (\sum(S_1(i) - S_2(i))^2)^{1/2}$$

This means that the Euclidean metric gives equal weight to peaks and troughs although spectral peaks are known to have more perceptual weight than troughs. For a comparison of two excitation patterns which have the same peak locations but varying slopes of shoulders around the peaks, the Euclidean metric has been considered to be unsuitable [6]. As the difference between the slopes increases, the distance calculated from Euclidean metric would increase, whereas the perceptual distance would remain unchanged. This was the result of the perceptual analysis by Klatt [6], who suggested the slope metric which emphasises the formant frequency values but is insensitive to relative formant amplitudes, or to spectral tilt changes. This effect is achieved by taking the square root of the squared differences of the first differential in the outputs of each filter.

The slope distance between two spectra, S_1 and S_2, with N channel filters is given by:

$$SDM = (\sum (S'_1(i) - S'_2(i))^2)^{1/2}$$

where S'_1 and S'_2 are the spectral slopes given by the first difference:

$$S'_1(i) = S_1(i+1) - S_1(i), \quad \text{for channel number, } i = 1, ..., N-1$$

The negative second differential metric (N2D) of Assmann & Summerfield [1] takes this idea further by comparing only the absolute value of the negative portions in the output spectra. In this case, spectral properties other than the formants are set to zero. Thus,

$$N2DM = (\sum (S''_1(i) - S''_2(i))^2)^{1/2}$$

where

$S''_1(i) = \max \{ - [S_1(i-1) - 2\,S_1(i) + S_1(i+1)\,], 0\})$, for $i = 1, ..., N-1$

So far, the distance analyses compare a particular spectral section of each auditory excitation pattern. However this may not be accurate since articulation of fricatives also change in time. To account for the dynamic fluctuation of the fricative signal, and differences in the length between the different fricatives and speakers, a non-linear time alignment technique was used [5]. This technique is based on a simple principle of optimisation; it relies on finding the shortest path between two compared segments aligned on a graph.

Finally, the distances between the aligned auditory spectra were calculated for each production of each speaker and used for 3-way, triangular matrices, interval level MDS analysis. The object of this technique is to obtain an optimal spatial representation of the scaled objects on the basis of analysed distances. In this way, we determine the minimal number of physical dimensions required to model the production data with maximal variance in the data accounted for.

3.3. Results and discussion

2-dimensional solutions were most appropriate for each distance metric analysis. Canonical coefficients between perceptual and physical spaces are first reported as an indication of the perceptual/physical relationship.

Dimensions	Canonical coefficients		
	Euclidean metric	Slope metric	N2D metric
1	0.995	0.596	0.935
2	0.933	0.377	0.175

Table 2. The canonical correlation values for each of the dimensions between the perceptual and physical data, compared for the different distance metrics.

It is clear from the above table that only the Euclidean metric gives high and consistent correlation values. This is contrary to the expectations that the slope and N2D metrics would give more accurate predictions of the perceptual data. This result may be attributed to the specific materials used in designing these metrics in the previous studies [1,6]. Indeed, Klatt's [6] study used 66 variations of the vowel /a/ each differing subtly in terms of acoustic properties. Klatt found that pairs of synthesised /a/ vowels which differed in terms of their formant frequencies were given the highest distance scores on a '10-point scale' of 'phonetic distance' judgements (as opposed to the 'psychoacoustic distance'). This means that he needed to devise a distance metric which would emphasize the formant frequencies, whilst ignoring other spectral variations. However, when the data involve different vowels with clearly different formant positions in the spectra, it is possible that the metric may be over-emphasising the differences. Therefore, the metric may be rather specific to the particular stimulus type used. Also in the Assmann & Summerfield study [1], there was concern over how well the pattern-matching procedure based on different distance metrics predicted the vowel identifications in the presence of competing voice (simultaneous double vowels). This means that they may have also needed to give extra emphasis to the spectral peaks, in order to allow each vowel to stand out from the other in the double vowels. Furthermore, the outputs of the slope and N2D metrics have not been transformed to MDS dimensions in previous studies. Thus, the results cannot be fully compared.

In respect to above findings, only the graphic representations of Euclidean physical spaces are presented in Figure 2. The Euclidean space of each speaker has been rotated for optimal congruence, and the new sets of coordinates are plotted on the same axes. The variance accounted for, averaged over 10 productions, was .958 and .035 for dimensions 1 and 2 respectively.

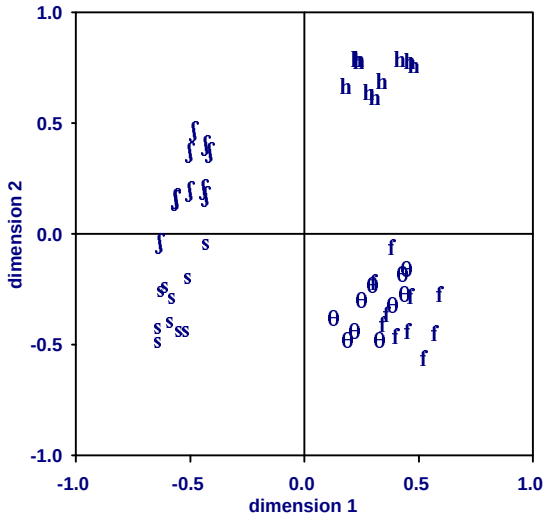

Figure 2. The physical space of English fricatives based on 10 productions by 5 male speakers.

Overall, it is remarkable how closely related the physical organisation is to their perceptual organisation, This physical map of fricatives shows that all the fricative regions are distinguished from one another, and are clearly organised in terms of their 'place' and 'sibilance' properties as in their perceptual map. There is an overlapping space between the fricative regions of /f/ and /θ/, which is, to some extent expected, given the proximity of their perceptual and phonetic properties. In comparison, /f/ and /θ/ in the perceptual space were much more distinct from each other, and the 'place' property was not clear. If we consider the point corresponding to the centre of gravity of each fricative region as its physical prototype, the stimuli in the perception tests may be regarded as acceptable variants of each prototype.

4. ACOUSTIC CORRELATES

Although the physical dimensions were interpreted in terms of phonetic properties, the question of whether they may be related to any concrete acoustic properties of spectra was not investigated. In particular, there may be many acoustic parameters that correspond to each physical dimension, or there may be a one-to-one correspondence between physical and acoustic properties as in

vowels. For this purpose, the average spectral shape of each fricative type used in the last section is placed in the corresponding region of the fricative on the physical map in Figure 2. This is shown in Figure 3.

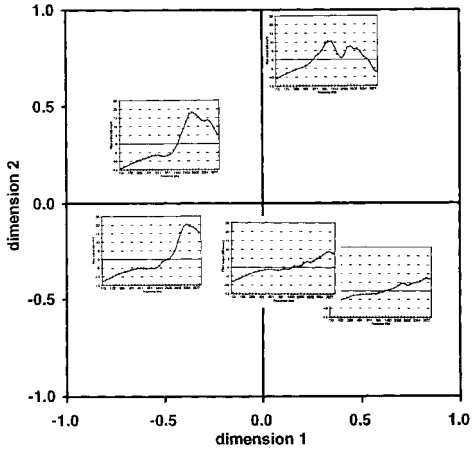

Figure 3. The average spectrum of each fricative is placed on the corresponding region of each fricative on the physical dimension.

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 this way, for every individual production, a series of 32 numbers was obtained, representing 32 filter bands. 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. This process was repeated for each production of each fricative. These spectra were averaged over the ten productions spoken by the five different speakers. The horizontal axis represents the centre frequencies of the 32 filters in Hz (from 100 to 9000 Hz). The vertical axis represents the energy levels of each filter in dB (-15 to 25).

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 physical dimension 1, in Figure 3, may be related to the 'peakiness' of spectra - the maximum distance to mean amplitude - while dimension 2 may be related to the centre of gravity of the spectra.

5. CONCLUSION

The principal achievement of this research is that studies of spatial representations on vowels have been successfully replicated in a set of consonants, and furthermore, that the findings are congruent with those obtained in vowel studies. The results have shown that the perception of fricative segments could be explained in 2-dimensional physical space in which each segment occupies a region. The dimensions of the space corresponded to 'sibilance' and 'place' in conventional phonetic terms, and to the frequency of the main spectral peak and the 'peakiness' of spectra in their acoustic characterisation. Therefore, the study of spatial representations, based on similarity data, 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.

Another important finding was that the essential peripheral auditory processing in the fricative data was adequately modelled by the auditory transformations used in the vowel data. A 1/3-octave bandpass filter bank analysis and non-linear intensity scaling were used. In addition, a non-linear time alignment procedure was employed in order to account for the time-varying spectral properties in fricatives. This was particularly important as this technique provides a basis for wider application of the same auditory analysis procedure in studying other consonants.

Physical distances between fricatives were most accurately modelled by the simple Euclidean distance metric in comparison to slope and N2D metrics (see section 3.3).

Note also that the materials used for perceptual and physical analyses were from different speaker groups; female and male speakers respectively. This means that the spatial representations, based on scaled distances between speech sounds, can automatically account for speaker normalisation effect in different processing domains.

Overall, these results suggest a unified experimental paradigm in which the development of speech perception models may be investigated in parallel for both vowels and consonant in terms of spatial representations based on similarity data.

ACKNOWLEDGEMENTS

Above all, I wish to thank Mark Huckvale who supervised the Ph.D. thesis [7] on which this paper is based. I am also grateful to Stuart Rosen and other members of Wolfson House, University College London, for their help and advice.

REFERENCES

[1] Assmann, P. F. & Summerfield, Q. (1989) Modelling the perception of concurrent vowels: Vowels with the same fundamental frequency. Journal of the Acoustical Society of America 85. 327-338.

[2] Kuhl, P. K. (1995) Mechanisms of developmental change in speech and language. Proc. ICPhS 95 Stockholm. Vol. 2, 132-139.

[3] 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.

[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] Sakoe, H. & Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE transaction. Acoustics, Speech, and Signal Processing 26. 43-49.

[6] Klatt, D. H. (1982) Prediction of perceived phonetic distance from critical-band spectra: a first step. Proc. ICASSP-82: IEEE transaction. Acoustics, Speech, and Signal Processing . 1278-1281.

[7] Choo, W. (1996) Relationships between phonetic perceptual and auditory spaces for fricatives. PhD thesis. University of London.