

# EVALUATION OF A TECHNIQUE FOR IMPROVING THE MAPPING OF MULTIPLE SPEAKERS' VOWEL SPACES IN THE $F_1 \sim F_2$ PLANE<sup>1</sup>

Dominic Watt & Anne Fabricius

## Abstract

We evaluate a vowel formant normalisation technique that allows direct visual and statistical comparison of vowel triangles for multiple speakers of different sexes, by calculating for each speaker a 'centre of gravity'  $S$  in the  $F_1 \sim F_2$  plane.  $S$  is calculated on the basis of formant frequency measurements taken for the so-called 'point' vowel [i], the average  $F_1$  and  $F_2$  for the vowel category with the highest average  $F_1$  (for English, usually the vowel of the TRAP or START lexical sets), and hypothetical minimal  $F_1$  and  $F_2$  values (coordinates we label [u<sup>l</sup>]) extrapolated from the other two points. Expression of individual  $F_1$  and  $F_2$  measurements as ratios of the value of  $S$  for that formant permits direct mapping of different speakers' vowel triangles onto one another, resulting in marked improvements in agreement in vowel triangle (a) area and (b) overlap, as compared to similar mappings attempted using linear Hz scales and the  $z$  (Bark) scale.

## 1. Introduction

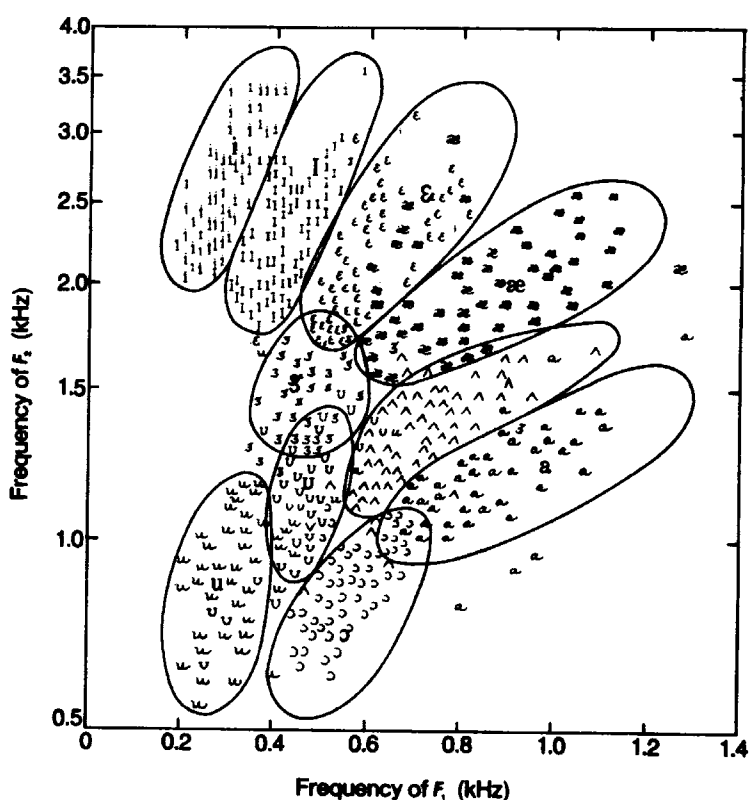
For some considerable time it has been commonplace in phonetic and sociolinguistic research to represent spoken vowels by means of the frequencies of their two lowest formants,  $F_1$  and  $F_2$ . The method has been adopted in order, among other things, to allow greater objectivity and replicability when classifying individual vowels than is possible using impressionistic auditory analysis alone.  $F_1$  has been shown to correlate inversely with the position of the highest part of the tongue body in the height dimension (open vowels have higher  $F_1$  values than close vowels do), while  $F_2$  is correlated with tongue frontness (front vowels have higher  $F_2$  values than back vowels do, especially if back vowels are also rounded). Vowels are frequently represented using straightforward measurements in linear Hz, or by expressing the relationship between the two parameters in some way (e.g. by plotting  $F_1$  against  $F_2 - F_1$  for a given vowel, as per Ladefoged & Maddieson 1990, Iivonen 1994), or by using some transform or 'warping' of the Hz scale so as to reflect the non-linear mapping of the acoustic parameter Hz to its perceptual correlates (e.g. through use of  $\log(\text{Hz})$  transforms, or the Mel, Koenig, Bark, or Equivalent Rectangular Bandwidth (ERB) scales). Some models also take account of higher formants such as  $F_3$ , or of the fundamental frequency ( $F_0$ ; see e.g. Hindle 1978, Disner 1980, Lobanov 1980, Moore & Glasberg 1983, Deterding 1990, Rosner & Pickering 1994, Labov 2001, or Adank *et al.* 2001 for evaluations of competing algorithms). In the case of the use of non-linear transforms, the intention is to minimise as far as possible the influence of non-linguistic factors on those properties in the acoustic signal which the researcher

---

<sup>1</sup> We are grateful to the following people for their input, comments and other feedback: Patti Adank, Paul Carter, Bernhard Fabricius, Paul Foulkes, Rob Hagiwara, Ghada Khattab, John Local, Richard Ogden, Peter Patrick, Jane Stuart-Smith, and an anonymous reviewer.

perceives to be important. Listeners appear capable of automatically factoring out certain aspects of the acoustic signal, such that they can, for example, understand natural speech produced by men, women and children with more or less equal proficiency, despite large differences in the acoustic signatures of 'equivalent' sounds produced by each type of speaker chiefly as a consequence of vocal tract length (VTL; e.g. Stevens 1998). A central concern in the acoustic analysis of vowels has therefore been to attempt to eliminate the effect of VTL on the relative frequencies of the lower formants for multiple speakers. By performing such 'normalisation' on speech signals, the researcher is permitted to make more direct comparison of formant frequencies of vowels spoken by speakers of different sexes and ages, and is also able to approximate more closely the way in which listeners may perceive spoken vowels.

Figure 1. Frequency of second formant versus frequency of first formant for ten American English vowels produced by 76 men, women and children (adapted from Peterson & Barney 1952 by Lieberman & Blumstein 1988).



An especially frequently used technique of visually assessing the similarities and differences between  $F_1$  and  $F_2$  frequencies for vowels produced by different speakers is one involving plotting unnormalised  $F_1$  and  $F_2$  against each other on  $x$ - $y$  scatter graphs (e.g. Peterson & Barney 1952, Hagiwara 1997, Watt & Tillotson 2001).<sup>2</sup> This method allows the researcher to superimpose one speaker's vowel sample onto another's, and thereby to estimate whether or not, for example, Speaker A has on the

---

<sup>2</sup> Hagiwara (1997) presents scatter plots in which the units are in Hz plotted on a Bark scale such that higher frequencies are compressed relative to lower ones, but this is a matter of adjusting the scaling on the axes of the plots rather than transforming the data themselves.

whole a higher  $F_2$  for a given vowel category than does Speaker B; such an observation might confirm a hypothesised process of vowel fronting. Data in this form also permit straightforward statistical comparison of samples, but only if it is assumed that VTL, and therefore the potential ranges of values for both  $F_1$  and  $F_2$ , are effectively constant across all the speakers sampled.

So as to minimise the potentially problematic influence of VTL-related variation among speakers of different ages and sexes, some researchers have used only post-pubertal male speakers as informants for investigations of vowel variation (e.g. Eremeeva & Stuart-Smith 2003). Serious problems are encountered if samples more representative of the population as a whole are used, because the  $F_1 \sim F_2$  frequencies for women tend to be significantly higher for adult females than for adult male speakers, with children having formant frequencies which are still higher than those of women. It is obviously not possible directly to compare (linear Hz)  $F_1 \sim F_2$  scatter plots for adult males and females, or for adults and children, because the  $F_1 \sim F_2$  planes for women – and particularly young children – are considerably stretched in both dimensions relative to those of male speakers (hence the elongation of the envelopes drawn around tokens of the peripheral monophthongs in Figure 1).

As mentioned above, numerous techniques have been devised in an attempt to reduce the discrepancies between the speech of men, women and children in this respect. Some are designed to compress the higher frequency ranges used by women and children relative to the lower ones; others work by expressing individual values in terms of distance from a mean derived from the formant frequency measurements themselves. An example of the first sort of transform is the Bark transform, which involves conversion of Hz measurements into perceptual units based on the critical bandwidth response of the ear (Zwicker & Feldtkeller 1967). We make no criticism of the use of Bark-transformed data, nor the validity of the scale itself, except to say that it does not in fact fully permit direct comparison of one speaker's vowel sample with another speaker's vowel sample in the way we would wish. This is because the influence of VTL is not actually wholly eliminated, since within the frequency range in which  $F_1$  typically falls – between c. 200Hz and 1 kHz – the mapping between Hz and Barks is effectively linear (see Traunmüller 1990; Adank *et al.* 2001). Within this frequency range, higher Hz values correspond very closely to proportionately higher Bark values, and it is only at frequencies well above those in which  $F_1$  is found that there is significant divergence between the scales. Therefore the problem of cross-speaker mapping persists, although the 'compression' of higher frequency ranges, such as those in which  $F_2$  is commonly found for adult speakers, corrects this problem to some degree. However, if our aim is to map one speaker's vowel space onto another's for the purposes of comparing their vowel systems, in a way which removes absolute differences in formant frequency further than Bark-transforming the data will allow, we must follow another approach.

We evaluate in this paper a method for allowing direct visual and statistical comparison of vowel spaces for different speakers which derives from measurements in Hz of  $F_1$  and  $F_2$  at the midpoints of stressed spoken vowels. Our focus will be on an assessment of the extent of reduction of speaker sex-related differences in samples of vowel formant frequencies for two RP British English speakers (one male, one female) where the frequency values are expressed on the following scales: (a) linear Hz; (b) critical band rate  $z$  (in Barks) and (c) a so-called 'S transform'. The last of these is calibrated from the  $F_1 \sim F_2$  plane's 'centre of gravity'  $S$  by taking the grand mean of the mean  $F_1$  and  $F_2$  frequencies for points at the apices of a triangular plane

which are assumed to represent  $F_1$  and  $F_2$  maxima and minima for the speaker in question (these being [i], [a] and [u<sup>1</sup>]; see below). The procedures for calculating  $z$  and  $S$  values for individual speakers are outlined in detail in the next section. Our estimate of the improvement in comparability between speaker samples is based on the increase in mapping between one speaker's vowel triangle and another's along two continuous parameters: (a) **the ratio of the area** of the female speaker's vowel triangle to that of the male speaker's triangle and (b) **the degree of overlap** between the two triangles, expressed in terms of that percentage of the male speaker's triangle which overlaps with the female speaker's triangle, and vice versa. It is demonstrated that on both counts the  $S$  transform performs much better than Bark-transformed representations of the two speakers' vowel triangles.

## **2. Methods**

### **2.1 Procedure for calculating critical band rate $z$ (in Barks)**

The transform used here is that from Traunmüller (1990):

$$z = \left[ \frac{26.81 f}{1960 + f} \right] - 0.53$$

where  $f$  is frequency in Hz. According to Traunmüller, the values obtained using this equation agree with the values tabulated by Zwicker (1961) to within  $\pm 0.05$  Bark in the frequency range 0.2 – 6.7 kHz.

For our present purposes, one advantage of converting all Hz measurements using the above equation is that one can apply the same transform to all the formant frequency measurements made for any number of speakers. The disadvantage, as noted above (and as demonstrated below), is that one only marginally reduces the effect of VTL, rather than eliminating it as far as possible. So while it is considerably more time-consuming to convert Hz measurements into the  $S$ -transformed values used for the comparison discussed in Section 3 below (because  $S$  values for each individual speaker must be calculated for  $F_1$  and  $F_2$ ), the latter technique, as we shall see, allows a much higher degree of mapping between samples for speakers whose VTLs are very different from each other.

### **2.2 Procedure for calculating $S$**

Our procedure for determining the  $F_1$  and  $F_2$  values of  $S$  for an individual speaker is discussed in this section. For clarity, we follow Wells (1982) in assigning the keywords FLEECE and TRAP to the lexical sets containing the vowels labelled /i:/ and /a/ in other descriptions of British English phonology, since we believe the use of phonetic symbols to represent vowel categories which are highly variable in British English (to the extent that, for example, the TRAP vowel can be realised as anything from [æ] to [ɔ], depending upon accent) to be potentially confusing.

#### **2.2.1 Step One**

- Assume that for a given speaker's sample the average  $F_1$  and the average  $F_2$  for the vowels of words of the FLEECE set represent that speaker's minimum

$F_1$  and maximum  $F_2$ . This seems a reasonable assumption, if no observations are made to the contrary (but see below).

- Assume that for a given speaker's sample the average  $F_1$  for the vowels of words of the TRAP set represents that speaker's maximum  $F_1$ . Depending upon the accent, one might wish to select words of the START set instead, since in certain accents of British English the TRAP vowel is generally produced with a somewhat raised quality. The influence of post-vocalic rhoticity in certain accents might present problems if START is used, however, because of the influence of a following rhotic on the formants of vowels in words like *start*, *car*, *farm*, etc. The point is to obtain an estimate of the region in which a speaker's maximum  $F_1$  is located, but clearly it is sensible to be consistent within a given sample (i.e. choosing either TRAP or START for all the informants concerned).

By definition, there will be individual formant frequency values higher and lower than the average  $F_1$  and  $F_2$  values we take to be maxima and minima for these formants. It might therefore be said that because  $F_1$  and  $F_2$  values for a given vowel category are generally somewhat - indeed often highly - variable, taking the mean values for  $F_1$  and  $F_2$  runs the risk of giving a false picture of the extremes of a speaker's vowel plane. However, averaging the  $F_1$  and  $F_2$  values for a given vowel category eliminates (or at least reduces) the potential of inaccurate individual formant frequency measurements to distort the geometry of an individual speaker's vowel triangle.

It might also be objected that this routine assumes that each speaker's FLEECE and TRAP vowels are more or less invariant, when it is clear from many previous studies that they are not, even in highly controlled speech elicited using artificial means. We must assume for the time being that FLEECE is rather less variable in accents of British English than other vowels, and that TRAP (or START) is likely to be the most open vowel speakers of British varieties will use. Again, it should be stressed that if the researcher is satisfied that FLEECE is relatively stable across a sample of speakers, that if he/she is circumspect about the choice of open vowel to use as the  $F_1$  maximum, and that if formant measurement is done as consistently as possible, we should be able to arrive at optimally comparable samples for speakers of different sexes and ages.<sup>3</sup>

### 2.2.2 Step Two

The next step is to arrive at an estimate of the  $F_1$  and  $F_2$  *minima* for a given speaker. In a very large number of studies of vowel variation in English, this limit is taken to be represented by the average  $F_1$  and  $F_2$  values for the vowel /u/, which we label here GOOSE. We take the view, however, that in many accents of English GOOSE is only rarely fully back, fully close, and fully rounded (see e.g. Hagiwara 1997; Watson *et al.* 1998; Labov 2001: 475ff), and that the average formant frequencies for this vowel produced by the average British English speaker are not a good reflection of the minimum possible  $F_1$  and  $F_2$  frequencies that such a speaker could achieve.

---

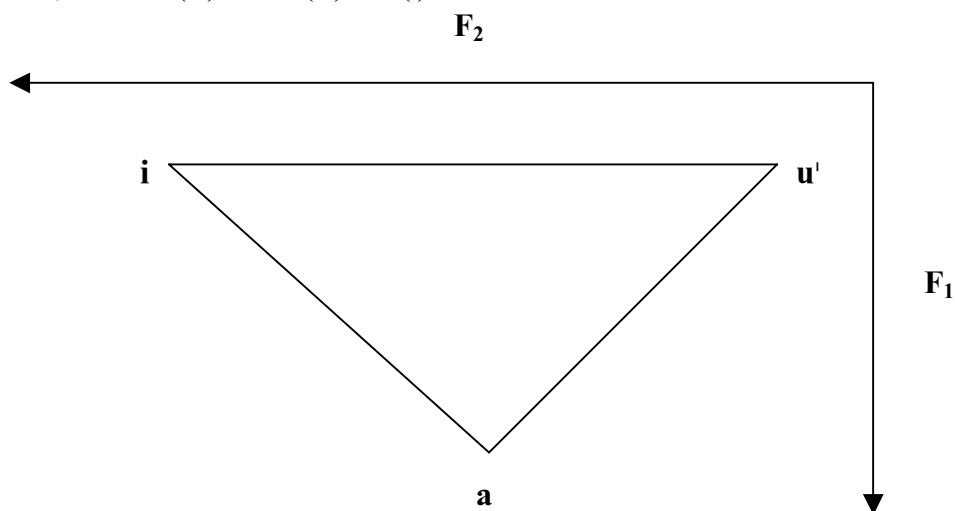
<sup>3</sup> There is no reason why other vowel categories such as KIT and/or FACE could not be used to represent  $F_1$  minima and  $F_2$  maxima, should it be anticipated or observed that the average formant values for FLEECE do not in fact provide a reliable estimate of these limits in the accent(s) under scrutiny.

Instead, we advocate the use of hypothetical lower limits on  $F_1$  and  $F_2$  which, though almost certainly not attested in a sample of informant's speech, are nonetheless arrived at in a principled way. These minimal values (or rather coordinates on the  $F_1 \sim F_2$  plane) we label [u']. They are arrived at as follows:

- It will be recalled from Section 2.2.1 that the average  $F_1$  for FLEECE was assumed to represent the minimum  $F_1$  for a given speaker. Therefore, we may assume that the  $F_1$  of [u'] is equivalent to that for FLEECE, since we have no evidence to suggest that it is any lower.
- Since - by definition -  $F_2$  cannot have a lower frequency than  $F_1$ , but often has a frequency so close to it that the spectral peaks cannot reliably be distinguished from one another using instrumental analysis, we can justifiably assume for present purposes that the speaker's closest, backest possible vowel has an  $F_2$  exactly equivalent to its  $F_1$  frequency. Thus,  $F_1$  and  $F_2$  of [u'] are (a) equal to the average  $F_1$  value for FLEECE for a given speaker, and therefore (b) exactly equal to one another.

The result of these calculations is a triangular area on the  $F_1 \sim F_2$  plane, as shown in Figure 2. Note that the axes are reversed, as is conventional in  $x$ - $y$  plots representing vowel systems.

**Figure 2. Schematised representation of the 'vowel triangle' used for the calculation of  $S$ .  $i$  = min.  $F_1$ , max.  $F_2$  (average  $F_1 \sim F_2$  for FLEECE);  $a$  = max.  $F_1$  (average  $F_1 \sim F_2$  for TRAP);  $u'$  = min.  $F_1$ , min.  $F_2$ , where  $F_1(u')$  and  $F_2(u') = F_1(i)$ .**



### 2.2.3 Step Three

The next step is to calculate for the individual speaker in question the  $F_n$  frequencies of the centre of gravity or 'centroid'  $S$  (following Koopmans-van Beinum 1980), which is quite simply the grand mean of  $F_n$  for  $i$ ,  $a$  and  $u'$  (a worked example is provided in Appendix 2). We then divide all the observed measurements of  $F_n$  by the  $S$  value for that formant, and express all resulting figures as values on scales  $F_n/S(F_n)$ , i.e. as ratios of  $S$ . Because  $S(F_n)$  divided by  $S(F_n)$  is always equal to 1 (with the coordinates of  $S$  therefore always being (1,1) in any speaker's vowel triangle), vowel

tokens with low  $F_n$  values on the Hz scale will have  $F_n/S(F_n)$  values between 0 and 1, while vowels with  $F_n$  values greater than the  $S$  value for that formant will have  $F_n/S(F_n)$  values higher than 1. Since all speakers' vowel triangles will be defined relative to  $S$ , we can compare samples for different speakers, both statistically and visually, directly with another. Plotting average or individual  $F_1 \sim F_2$  measurements for phonologically 'back' vowels such as GOOSE and GOAT on the  $F_1/S(F_1) \sim F_2/S(F_2)$  plane is thus straightforward, regardless of how phonetically back or front these vowels are.

### 2.3 Materials

We turn now to compare vowel samples for the two British English RP speakers referred to in Section 1. The data are drawn from formant frequency measurements made by Deterding (1997) from recordings of BBC broadcasts held in the MARSEC (Machine Readable Spoken English Corpus) database (Roach *et al.* 1993).<sup>4</sup> The programmes in question were broadcast in the 1980s, and according to Roach *et al.*, 'the accent of all the speakers is RP or close to it' (Roach *et al.* 1993:48). From the ten speakers (5 male, 5 female), we selected a male speaker and a female speaker at random. The speakers in question are A (female) and C (male); speaker A's sample is drawn from a religious affairs programme, while C's is based on a radio lecture on economics (see Deterding 1997:48). Because our intention here is to assess the relative effectiveness of  $z$ -transforming and  $S$ -transforming the linear Hz data in terms of mapping one speaker's FLEECE  $\sim$  TRAP  $\sim$  GOOSE triangle onto another's, it is sufficient to use two speakers whose formant frequencies in the Hz domain for 'equivalent' vowels are markedly mismatched, though of course any number of speakers could be compared using this technique.

Details of how the original formant measurements themselves were made can be found in Deterding (1997:48-50); the source figures can be downloaded directly from the Internet.<sup>5</sup>

## 3. Results

### 3.1 Triangles plotted using Hz scale

The relative shapes, sizes and degree of overlap between the triangles generated from the raw Hz data for speakers A and C are shown in Figure 3.

Agreement of the areas of the two triangles is poor: that for the female speaker A ( $\Delta A$ ) is almost four times larger than that for the male speaker C ( $\Delta C$ ) at a  $\Delta C : \Delta A$  ratio of 1 : 3.93 (see Table 1 below for the full results in tabular form). The degree of overlap is also low: the proportion of  $\Delta C$  overlapping  $\Delta A$  is just 46.1%. That is, more than half of  $\Delta C$  lies in an area of the vowel plane which is unoccupied by  $\Delta A$ , as we would expect given the significantly lower average  $F_1 \sim F_2$  frequencies for adult male speakers. The proportion of the vowel plane occupied by  $\Delta A$  which lies outside  $\Delta C$  approaches 90% (86.3%). We can say, therefore, that the mapping of the samples for these two speakers is overall very poor.

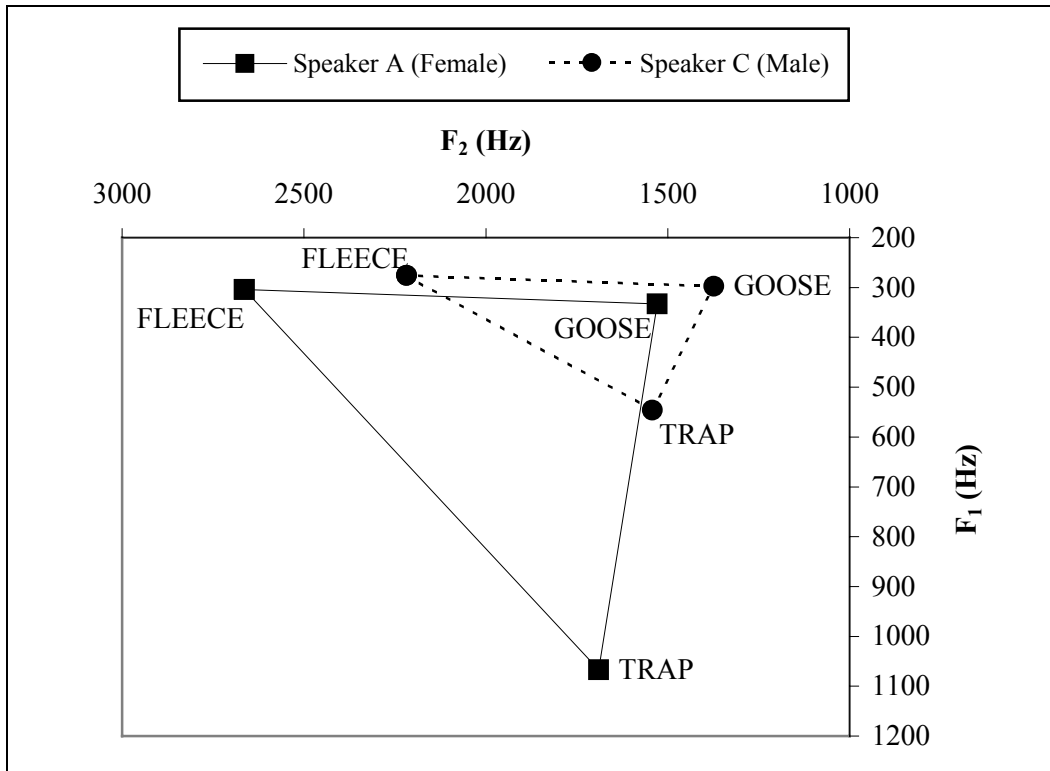
<sup>4</sup> See <http://www.rdg.ac.uk/AcaDepts/ll/speechlab/marsec/>.

<sup>5</sup> <http://www.arts.nie.edu.sg/ell/davidd/data/jipa-vowels/index.htm>. Note that the URL provided in the appendix of Deterding (1997) is no longer active.

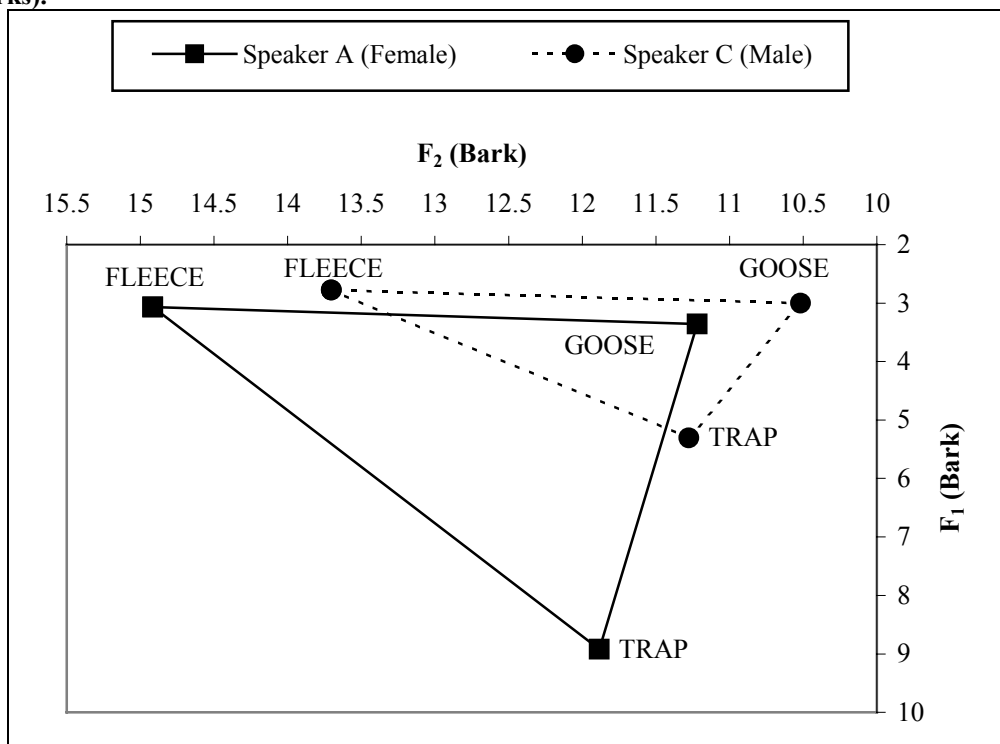
### 3.2 Triangles plotted using $z$ (Bark) scale

Figure 4 shows the same data  $z$ -transformed using Trautmüller's equation discussed in Section 2.1.

**Figure 3. Comparison of FLEECE ~ TRAP ~ GOOSE vowel triangles for Speakers A and C (linear Hz).**



**Figure 4. Comparison of FLEECE ~ TRAP ~ GOOSE vowel triangles for Speakers A and C (Barks).**



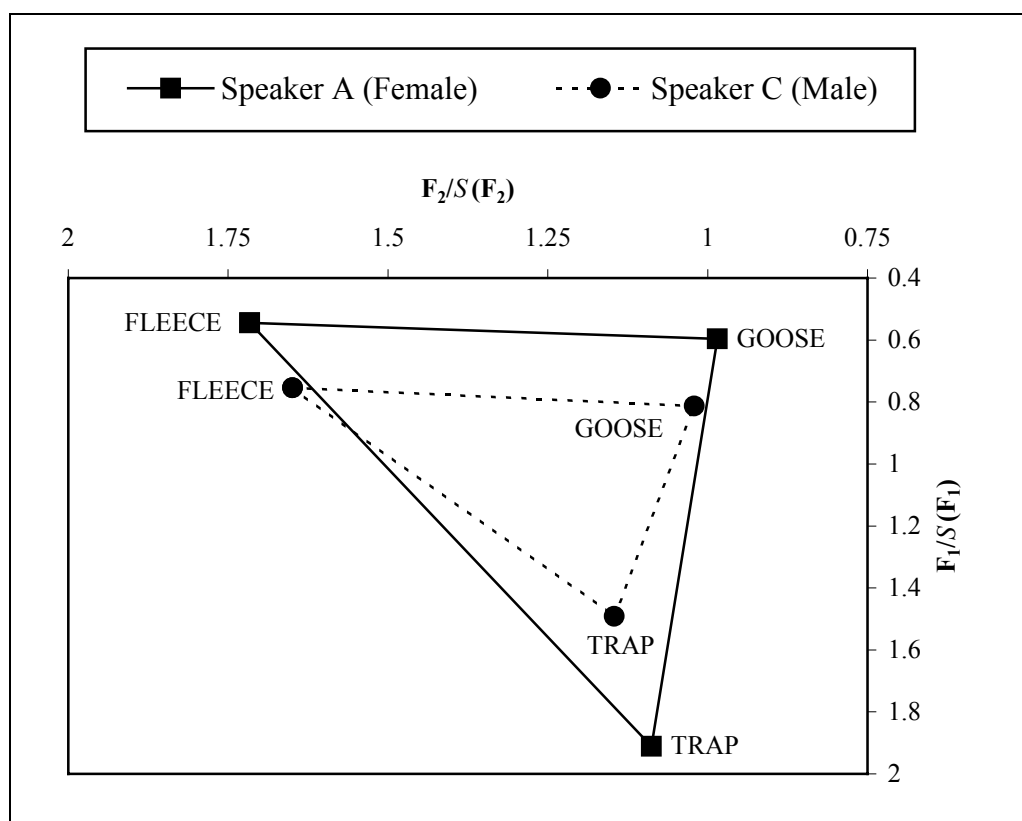


There is a noticeable improvement here in terms of area ratio, the ratio of  $\Delta C$  to  $\Delta A$  now being 1 : 2.76. This means that there is an improvement in agreement in area ratio of 29.8% over the equivalent triangles on an  $F_1$  (Hz) ~  $F_2$  (Hz) plane if we transform the Hz measurements into Bark units. However, the extent to which the two triangles overlap is not greatly improved: the portion of  $\Delta C$  which overlaps  $\Delta A$  still accounts for just under half (49.9%) of the total area of  $\Delta C$ , while the overlapping area occupies a mere 18.1% of  $\Delta A$ .

### 3.3 Triangles plotted using $S$ units

If the Hz figures are transformed using the  $S$ -transform described in Section 2.2 above, however, we see dramatic improvements in both area ratio and degree of overlap. Figure 5 shows that all but a tiny fraction of  $\Delta C$  overlaps with  $\Delta A$ , and that there is a substantial improvement in the match between the areas for the two triangles.

**Figure 5. Comparison of FLEECE ~ TRAP ~ GOOSE vowel triangles for Speakers A and C ( $F_n/S(F_n)$ ).**



Although there is still clearly a fair degree of mismatch between the areas of the two triangles – particularly in terms of  $F_1$  differences for each of the three vowel categories – at a  $\Delta C$  :  $\Delta A$  ratio of 1 : 2.16 the agreement in area is nonetheless improved relative both to Hz (45% improvement) and to the Bark-transformed data (21.7% improvement). Degree of overlap expressed in terms of the proportion of  $\Delta C$  overlapping with  $\Delta A$  approaches complete overlap, at 99.2%. That portion of  $\Delta A$  overlapping  $\Delta C$  is 45.8% of the overall area of  $\Delta A$ .

### 3.4 Summary

To summarise the marked improvements in area and overlap agreement resulting from *S*-transforming the original Hz data, the figures discussed in the preceding paragraphs above are shown in tabular form in Table 1.

**Table 1. Improvements in area ratio and degree of overlap between FLEECE ~ TRAP ~ GOOSE triangles for Speakers A (female) and C (male).**

	<b>Hz</b>	<b>Bark</b>	<b>S</b>
<b>area ratio (<math>\Delta C : \Delta A</math>)</b>	1 : 3.93	1 : 2.76	1 : 2.16
<b>% improvement over Hz</b>	-	29.8	45
<b>% improvement over Bark</b>	-	-	21.7
<b>% overlap (<math>\Delta C : \Delta A</math>)</b>	46.1	49.9	99.2
<b>% improvement over Hz</b>	-	8.2	115.2
<b>% improvement over Bark</b>	-	-	98.8
<b>% overlap (<math>\Delta A : \Delta C</math>)</b>	13.7	18.1	45.8
<b>% improvement over Hz</b>	-	32.1	234.3
<b>% improvement over Bark</b>	-	-	153

It may be noted from Figure 5, incidentally, that the  $F_n/S(F_n)$  values for Speaker C's GOOSE vowel approach (1,1). That is, his GOOSE vowel is on average very close to the centre of gravity calculated for his vowel space on the basis of the actual and extrapolated  $F_1 \sim F_2$  values in his sample. This can be seen as a demonstration of the advantages of not using the average  $F_1$  and  $F_2$  values for GOOSE in the calculation of *S*, since if C's GOOSE vowel has average  $F_1$  and  $F_2$  values in the central region of his vowel space it would be unwise to treat it as a 'back' vowel from a phonetic point of view. If we were to use it to represent the  $F_1 \sim F_2$  minima for this speaker because we assume it to be the closest and backest vowel that speaker could produce, we run the risk of distorting the overall shape and underestimating the extent of Speaker C's maximal triangle on the  $F_1 \sim F_2$  plane. Furthermore, by plotting a speaker's actual average  $F_1 \sim F_2$  values for GOOSE and other phonologically back vowels within the triangle whose rearward boundary is defined by the extrapolated coordinates [u<sup>1</sup>], we gain an impression of the location of these back vowels relative to this rearward limit. For example, we can assess whether one English speaker is in the habit of using on average a fronter pronunciation of the GOOSE or GOAT vowels than another, and we can, moreover, be confident that if differences of this sort are in evidence when the relevant formant frequency values are expressed in terms of  $F_n/S(F_n)$ , they will also be found in the original Hz measurements (i.e., that they are not artefacts of the *S*-transform algorithm but reflect real inter-speaker differences which are not attributable simply to difference in VTL).

It is perhaps trivial to point out that individual vowels can be plotted on the  $F_1/S(F_1) \sim F_2/S(F_2)$  space as easily as averaged  $F_n/S(F_n)$  values for vowel categories can. By way of illustration, Figure 6 in Appendix 1 shows Hz and  $F_n/S(F_n)$  plots for all the individual vowel tokens for Speaker A. We feel, however, that it is important to note that the absence of warping of the vowel space of the sort inherent in Bark-transformed data means that one can inspect vowel plots plotted on axes using the  $F_n/S(F_n)$  scale *as though* they were plotted using Hz scales, while simultaneously

being able to map multiple plots onto one another more fully than is possible using either the Hz or the Bark scale.

#### 4. Conclusion

We may see from Table 1 that the *S*-transform allows much closer mapping of samples for different speakers onto one another than do the original measurements in linear Hz, and their equivalent values on the Bark scale. It outperforms the *z*-transform on both criteria, and more particularly on the overlap criterion, in which improvements are on the order of 100 – 150%.

We do not intend the above evaluation as a criticism of the Bark scale in any other respect, however: we propose the *S*-transform *only* as a means of allowing enhanced visual and statistical comparisons between vowel formant data sets collected for different speakers, and do not claim it has any psychoperceptual validity (e.g. that it mimics the normalisation process assumed to exist for the auditory processing of speech signals, or such like). Instead, we see it solely as a useful tool for researchers wishing to reduce inter-speaker differences resulting from variations in VTL when performing analyses of vowel samples in, for instance, instrumental studies of vowel variation and change.

Although it has been demonstrated using only very limited amounts of data drawn from recordings of two English speakers, we do not expect that the effectiveness of the *S*-transform on the area ratio/overlap criteria will be diminished much, if at all, if applied to data from other languages, or from larger numbers of speakers. Although it is a relatively cumbersome algorithm to use on large samples of vowel formant data (especially compared to converting Hz values into Bark units) there are clear advantages - at least according to the criteria chosen for this evaluation - to the *S* transform over Hz measurements or their equivalents on the Bark scale. There are obviously great improvements that could be made, for example by finding some means of correcting the discrepancy between male and female speakers with respect to  $F_1/S(F_1)$ , or perhaps by running the *S*-transform on *z*-transformed data. There are also many other normalisation algorithms that the *S*-transform can be evaluated against on the area ratio and overlap parameters used as criteria in the present study; comparisons will be reported on in due course.

#### References

- Adank, P., van Hout, R. & Smits, R. (2001). A comparison between human vowel normalization strategies and acoustic vowel transformation techniques. *Proceedings of the 7<sup>th</sup> International Conference on Speech Communication and Technology (Eurospeech 2001) Aalborg, Vol. I*. pp. 481-4.
- Deterding, D. (1990). *Speaker normalization for automatic speech recognition*. Unpublished PhD thesis, University of Cambridge.
- Deterding, D. (1997). The formants of monophthong vowels in Standard Southern British English pronunciation. *Journal of the International Phonetic Association* **27**. 47-55.
- Disner, S.F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America* **67**(1). 253-61.
- Eremeeva, V. & Stuart-Smith, J. (2003, forthcoming). *A sociophonetic investigation of the vowels OUT and BIT in Glaswegian*. Proceedings of the 15<sup>th</sup> International Congress of Phonetic Sciences, Barcelona.

- Hagiwara, R. (1997). Dialect variation and formant frequency: the American English vowels revisited. *Journal of the Acoustical Society of America* **102**(1). 655-8.
- Hindle, D. (1978). Approaches to vowel normalization in the study of natural speech. In Sankoff, D. (ed.) *Linguistic Variation: Models and Methods*. New York: Academic Press. pp. 161-72.
- Iivonen, A. (1994). A psychoacoustical explanation for the number of major IPA vowels. *Journal of the International Phonetic Association* **24**(2). 73-90.
- Koopmans-van Beinum, F.J. (1980). *Vowel contrast reduction: an acoustical and perceptual study of Dutch vowels in various speech conditions*. PhD thesis, University of Amsterdam.
- Labov, W. (2001). *Principles of Linguistic Change, vol. II: Social Factors*. Oxford: Blackwell.
- Ladefoged, P. & Maddieson, I. (1990). Vowels of the world's languages. *Journal of Phonetics* **18**. 93-122.
- Lieberman, P. & Blumstein, S.E. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge: Cambridge University Press.
- Lobanov, B.M. (1980). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America* **67**. 253-61.
- Moore, B.C.J. & Glasberg, B.R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America* **74**. 750-3.
- Peterson, G.E. & Barney, H.L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* **32**. 693-702.
- Roach, P., Knowles, G., Varadi, T. & Arnfield, S. (1993). MARSEC: a machine-readable spoken English corpus. *Journal of the International Phonetic Association* **23**. 47-54.
- Rosner, B.S. & Pickering, J.B. (1994). *Vowel Perception and Production*. Oxford: Oxford University Press.
- Stevens, K.N. (1998). *Acoustic Phonetics*. Cambridge, Mass.: MIT Press.
- Trautmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America* **88**(1). 97-100.
- Watson, C.I., Harrington, J. & Evans, Z. (1998). An acoustic comparison between New Zealand and Australian English vowels. *Australian Journal of Linguistics* **18**(2). 185-207.
- Watt, D.J.L. & Tillotson, J. (2001). A spectrographic analysis of vowel fronting in Bradford English. *English World-Wide* **22**(2). 269-302.
- Wells, J.C. (1982). *Accents of English (3 vols)*. Cambridge: Cambridge University Press.
- Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *Journal of the Acoustical Society of America* **33**. 248.
- Zwicker, E. & Feldtkeller, R. (1967). *Das Ohr als Nachrichtenempfänger*. Stuttgart: S. Hirtzel Verlag.

*Dominic Watt  
Department of English  
University of Aberdeen  
Taylor Building  
Old Aberdeen  
Aberdeen AB24 3UB  
Scotland*

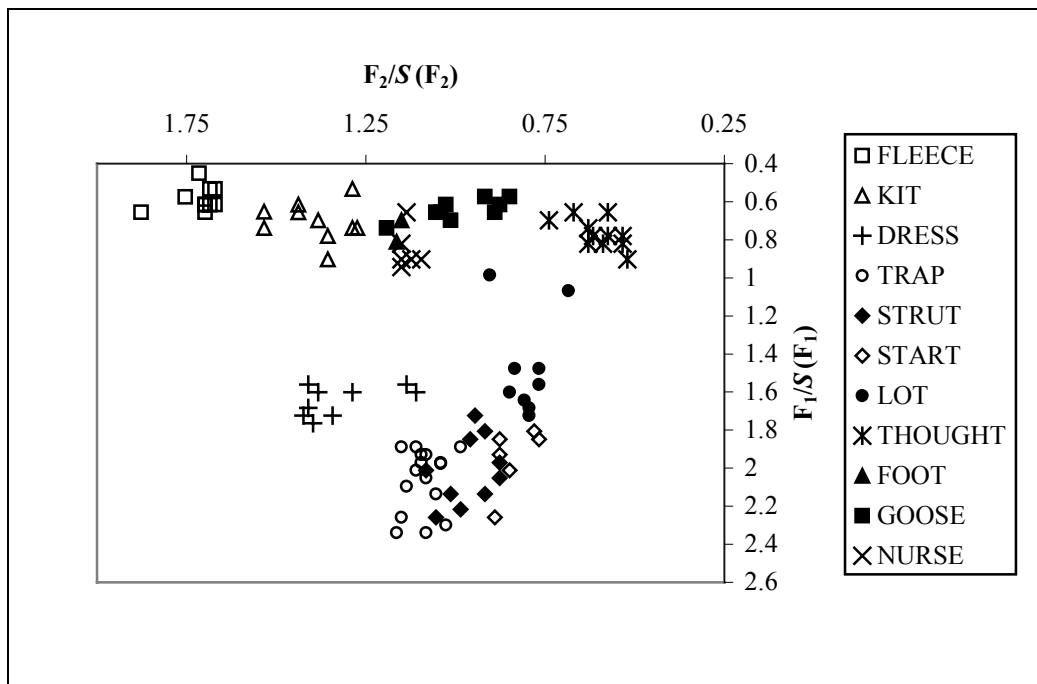
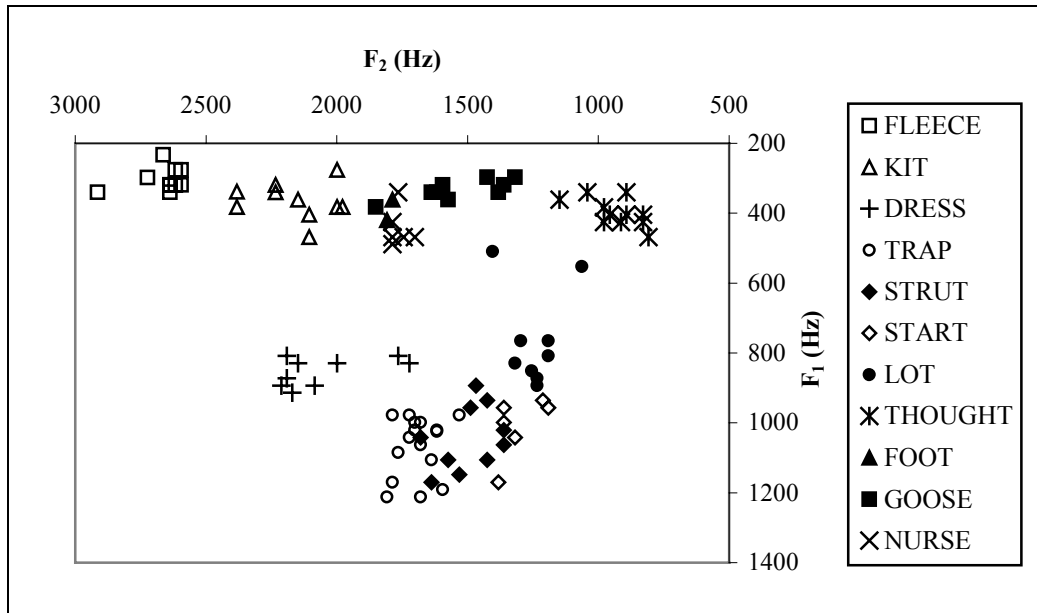
*d.j.l.watt@abdn.ac.uk*

*Anne Fabricius  
English Section  
Department of Language & Culture  
Roskilde University  
PO Box 260  
DK-4000 Roskilde  
Denmark*

*fabri@ruc.dk*

**Appendix 1**

**Figure 6. Vowel plots for Speaker A (data from Deterding 1997). Scales are in Hz (upper pane) and in  $F_n/S(F_n)$  units (lower pane).**



**Appendix 2**

Calculation of *S*: worked example (figures for Speaker A).

**Mean  $F_1$  and  $F_2$  for [i a u<sup>l</sup>], derived from Deterding's (1997) data**

Vowel	$F_1$ (Hz)	$F_2$ (Hz)
i	304	2664
a	1067	1690
u <sup>l</sup>	304	304 (i.e. both values equal to $F_1$ for [i])

⇓

**Mean  $F_1$  and  $F_2$  for *S***

$$S(F_1) = \frac{304 + 1067 + 304}{3} = \frac{1675}{3} = 558.3$$

$$S(F_2) = \frac{2664 + 1690 + 304}{3} = \frac{4658}{3} = 1552.7$$

⇓

**Speaker A's FLEECE, TRAP and GOOSE means (Hz) converted into *S* units**

$$\left. \begin{array}{l} 304 \\ 1067 \\ 333 \end{array} \right\} \div 558.3 \qquad \left. \begin{array}{l} 2664 \\ 1690 \\ 1529 \end{array} \right\} \div 1552.7$$

⇓

Vowel	$F_1/S(F_1)$	$F_2/S(F_2)$
FLEECE	0.545	1.716
TRAP	1.911	1.088
GOOSE	0.596	0.985