Establishing typicality: A closer look at individual formants

Vincent Hughes

Citation: Proc. Mtgs. Acoust. **19**, 060042 (2013); doi: 10.1121/1.4798775

View online: http://dx.doi.org/10.1121/1.4798775

View Table of Contents: http://asa.scitation.org/toc/pma/19/1

Published by the Acoustical Society of America

Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 1pSCc: Distinguishing Between Science and Pseudoscience in Forensic Acoustics II

1pSCc4. Establishing typicality: A closer look at individual formants

Vincent Hughes*

***Corresponding author's address: Language and Linguistic Science, University of York, Heslington, York, YO10 5DD, Yorkshire, United Kingdom, vh503@york.ac.uk**

Research into the forensic performance of individual formants has offered considerable evidence to support the traditional acoustic-phonetic view that whilst F1 and F2 encode broad phonetic contrast, higher formants may offer greater speaker-discriminatory potential (Peterson 1959, Ladefoged 2006, Clermont and Mokhtari 1998, Rose 2002). However, the comparative performance of formants has largely been assessed using posterior assessments of speaker-specificity (McDougall 2004, 2006; Clermont et al 2008). Using quadratic polynomial fittings of F1 to F3 from spontaneous tokens of /ai/ extracted from all 100 speakers in the DyVis database (Nolan et al 2009), this paper discusses issues relating to p(H|E)-based voice comparison analysis (particularly the use of discriminant analysis, DA). Further, DA performance is compared with an analysis based on likelihood ratios (LRs). LRs based on F3 are found to only marginally outperform F1 and F2 with regard to the magnitude of same-speaker and different-speaker strength of evidence, as well system performance metrics (EER and Cllr). The poorer than expected F3 LRs are assessed with regard to the distributions of values within- and between-speakers for the best performing F3 coefficient: the constant. The data go some way to establishing F3 population statistics which may potentially be applied to voice comparison casework.

Published by the Acoustical Society of America through the American Institute of Physics

Received 19 Jan 2013; published 2 Jun 2013

ESTABLISHING TYPICALITY: A CLOSER LOOK AT INDIVIDUAL FORMANTS

Introduction

Many previous studies have considered the comparative value of individual formants as speaker discriminants for forensic voice comparison (FVC). The results of those which have assessed discriminatory potential using Bayesian posterior probability metrics (particularly discriminant analysis, DA) (McDougall 2004, Simpson 2008, Hughes et al 2009) offer considerable support for the traditional view that higher formants, in particular F3, carry speaker-specific information since it is the lowest two formants which are primarily responsible for carrying social-indexical information, and maintaining phonetic contrast (Peterson 1959, Ladefoged 2006). However, such positive results have not been replicated when assessing strength of evidence using likelihood ratios (LRs). Based on the absolute magnitude of LRs computed using Japanese mid-point vowel data, Kinoshita (2001) found considerable variability in individual formant performance by phoneme. Similarly, Alderman (2004) found that whilst there was some improvement in discrimination using F2 and F3 compared with F1, systems displayed considerable variability in the percentage of same-speaker pairs achieving LRs > unity.

These results highlight two significant issues for FVC research and casework. Firstly, there is a practical concern about the extent to which optimistic posterior-based discriminatory performance using F3 reflects limitations with the analysis procedure itself. Specifically, since DA is a form of closed-set analysis (whereby the speaker space (Nolan 1991) is inhabited only by the speakers included in the model) it is predicted that the results are, in part, a consequence of the use of relatively small samples (usually up to 25 speakers). The use of larger samples is therefore likely to have a negative effect on performance since the range of F3 variation between speakers is not expected to increase sufficiently as the speaker space becomes more densely populated.

Secondly, the results of LR-based studies raise questions about the inherent speaker-discriminatory value of F3. Despite Rose's (2002) claim that higher formants are more closely related to resonances in smaller cavities in an individual's vocal tract, there are a number of extraneous factors which may introduce systematic variation in F3. Such factors include lip rounding (Stevens 2001), rhotacisation (Delattre and Freeman 1968, Lindau 1978, Ladefoged 2006) and accent-specific vocal settings (Laver 1994:§13.5.2.3, Esling and Dickson 1985), all of which may diminish speaker-discriminatory power.

Using quadratic polynomial estimations of F1, F2 and F3 of spontaneous /aɪ/ tokens produced by a homogeneous population of speakers, the present study firstly assesses the effects of different probabilistic approaches on the estimation of speaker-discriminatory value and discusses issues with the use of small samples in posterior-based DA. LR-based testing is then used to investigate the extent to which speaker-specificity is encoded in the individual formants of /aɪ/. The intrinsic speaker-discriminatory value of F3 is considered with reference to the distribution of values within- and between-speakers. The implications of these findings with regard to the informed choice of speaker discriminants for FVC are also explored.

Method

Data

Testing was performed using F1, F2 and F3 trajectories of /aɪ/ in spontaneous speech from the DyVis database (Nolan et al 2009) of young, male, Standard Southern British English (SSBE) speakers. Data was extracted from DyVis Task 1 which involved a mock police interview, aiming to "elicit spontaneous speech in a situation of 'cognitive conflict,' where speakers are made to lie" (Nolan et al 2009:41). Only target tokens occurring in /aɪp/, /aɪt/ and /aɪk/ contexts were included for ease of segmentation. A Praat script extracted nine time-normalised Hz values from each of the first three formant trajectories tracking maximally between 5.0 and 6.0 formants, and errors were hand-corrected. Of the original 100 speakers, three were removed due to small numbers of available tokens. The resulting data set contained 97 speakers with between 11 and 19 tokens per speaker (mean = 14.5). Formant trajectories were fitted with quadratic polynomial curves of the form $y = ax^2 + bx + c$ reducing the nine raw Hz values to three coefficients per formant.

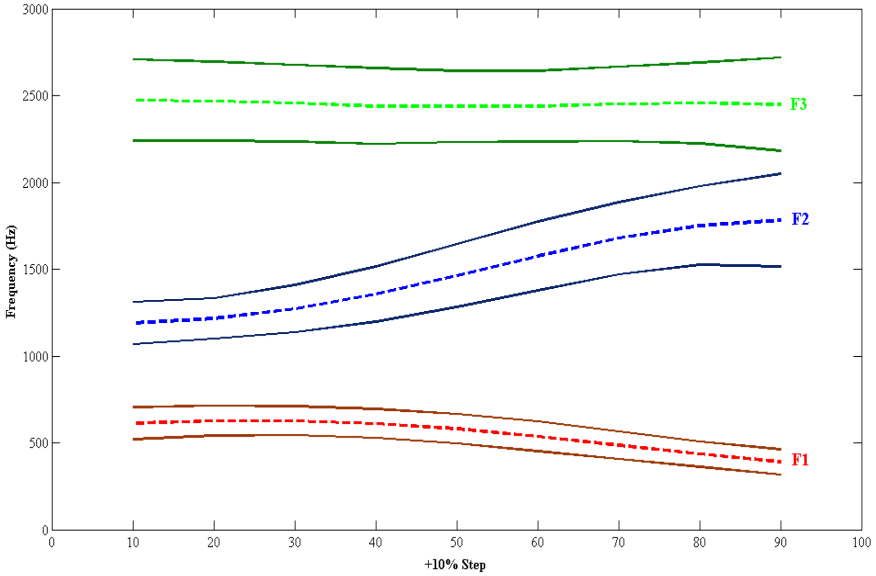


Figure 1. Mean F1, F2 and F3 trajectories (dashed lines) ± one standard deviation (solid lines) for 97 DyVis speakers based on between 11 and 19 tokens per speaker

Discriminant analysis

DA is a closed-set form of Bayesian posterior analysis which generates a classification rate of the proportion of cases correctly assigned to a given group based on a series of input predictors (Tabachnick and Fidell 2007:23-24). As an expression of $p(\mathrm{H}|\mathrm{E})$, the DA classification rate is logically and legally at odds with the *paradigm shift* (Saks and Koehler 2005) towards a Bayesian LR framework for forensic comparison evidence (see Morrison 2008:261-264). Despite this, DA is still commonly used in FVC research as an exploratory tool. Previous studies often used relatively small numbers of speakers, so DA testing was performed here on a much larger population in order to address how the speaker space is affected by population size. DA was performed using the 'leave one out' method, starting with five speakers and increasing in blocks of five up to a maximum of 89 (13 tokens per speaker). A cross-validated classification rate was generated at each stage.

LR-based testing

From the 97 available speakers, 20 were chosen at random to act as LR test data. The typicality of within- and between-speaker variability was assessed against models generated using the remaining 77 speakers. A MatLab implementation (Morrison 2007) of Aitken and Lucy's (2004) Multivariate Kernel Density (MVKD) formula was used to compute LRs, modelling within-speaker values using an assumption of normality and between-speaker values with a multivariate Gaussian kernel. The first ten tokens per speaker were divided in half and contemporaneous SS and DS LR comparisons were performed, outputting raw, $\log_{10}$ and natural log LRs for each pair. The magnitudes of LRs are assessed with reference to Champod and Evett's verbal scale (2000:240). Performance is assessed using equal error rate (EER) and log-LR cost function (C_{llr}) (Brümmer and du Preez 2006).

This study is of course limited in its forensic realism given the optimal testing conditions. Therefore the magnitude of the LRs achieved may be somewhat optimistic relative to those in real FVC casework. For the purposes of the present study comparative performance is of primary concern, rather than absolute strength of evidence.

Results

Discriminant Analysis

Figure 2. reveals that despite small numbers of predictors (three per formant), very positive DA classification rates are achieved for each formant when including small numbers of speakers. With five speakers F3 achieves the highest classification rate (60%) and F2 the lowest (52.3%). However, increasing the number of speakers has two important effects on system performance. First, there is a marked decrease in the classification rates achieved between the minimum and maximum number of speakers, although performance is consistently better than chance. The largest decrease is found in F3 (51.7%), such that the F3 classification rate using 89 speakers is just 8.3%. This confirms the prediction that DA classification is highly dependent on the number of speakers included and with only a small sample can provide an overly optimistic estimation of absolute speaker-discriminatory potential.

Secondly, the number of speakers affects comparative performance of individual formants. F3 predictors consistently outperform F1 and F2 when between five and 45 speakers are used. With greater than 45 speakers, classification rates for F1 are generally higher than those for F2 and F3. Further, with 10 and 15 speakers F2 marginally outperforms F1. Therefore, DA classification rates can misrepresent the comparative value of different formants depending on the number of speakers included.

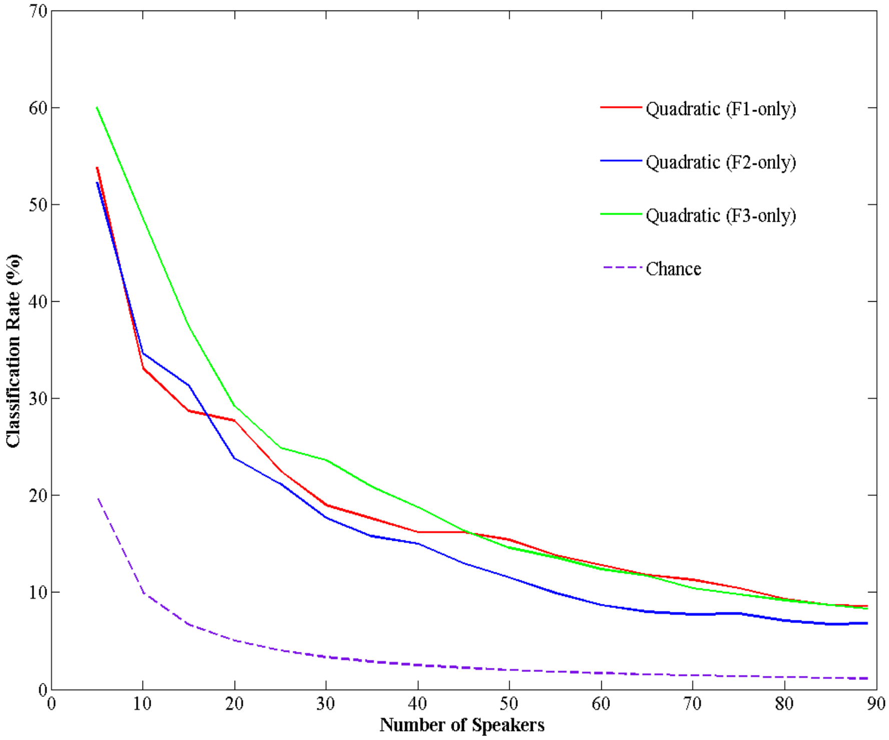


FIGURE 2. Classification rates using predictors from each of the first three formants individually according to the number of speakers included, with chance classification rate plotted (dashed line)

LR-based testing

Figure 3. reveals that the strongest same-speaker (SS) and different-speaker (DS) LRs are achieved when using combined F1, F2 and F3 input, suggesting that all three formants contribute towards speaker discrimination. The best-performing formant, in terms of the magnitude of the LRs, is F3. However, the differences in the magnitudes of SS LRs are marginal across formant conditions, with the F3 mean $\log_{10}$ LR (0.686) just 0.061 greater than that for F1 and 0.135 greater than F2. All three formants on average achieve SS LRs equivalent to ‘limited’ support for H_p. Despite a marginally higher mean SS $\log_{10}$ LR for F3, the largest individual SS LR is achieved using F1 (max $\log_{10}$ LR = 1.82). SS LRs for both F1 and F2 were found to be spread over a wider range than those for F3.

For DS pairs, the mean $\log_{10}$ LR for F3 is -2.046 (‘moderately strong’) which is a single verbal category higher compared with mean values for F1 and F2 (‘moderate’). In terms of the absolute $\log_{10}$ values, however, F3 is on average only 1.23 times greater than F1. As with SS pairs, F1 DS LRs offer higher strength of evidence than those for F2. According to the comparative magnitude of LRs, F3 appears to offer the most towards the combined strength of evidence achieved using all three formants, followed by F1 and F2. However, given the marginal differences between the three systems, the role of F3 in speaker-discrimination compared with those of lower formants is not considerably greater.

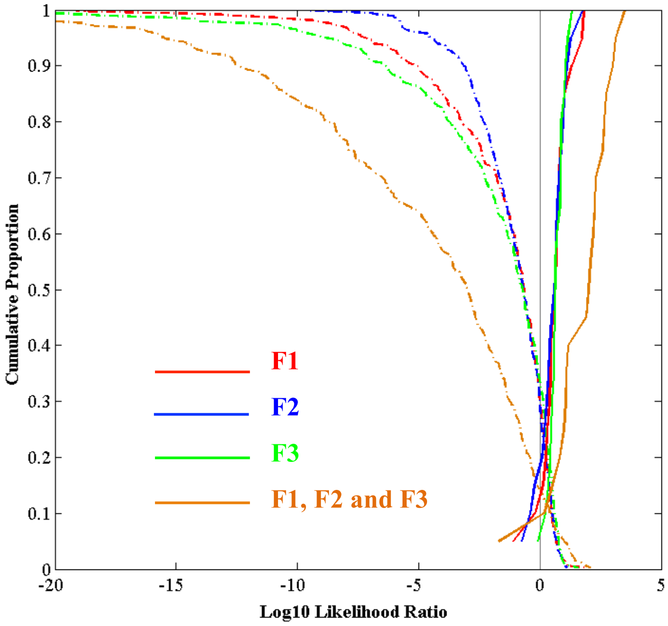


FIGURE 3. Tippett plot of SS (bold) and DS (dashed) LR comparisons using quadratic coefficients of all formants combined (orange), F1-only (red), F2-only (blue) and F3-only (light green) as input data

Discriminatory performance is also consistent with the error metrics displayed in Figure 4. The combination of F1, F2 and F3 achieves the lowest EER and C_{llr}, indicating that both the percentage of pairs offering contrary-to-fact evidence and the magnitude of 'errors' were lower for all three formants combined than any one formant individually. F3 is the best-performing individual formant, displaying lower EER and C_{llr} values than F1 or F2, with F1 marginally outperforming F2. Although the LR-based results offer some support for the value of F3 as a speaker discriminant, the magnitudes of the differences between formants, particularly in terms of strength of evidence, are small. Indeed, the finding that strength of evidence is almost as strong for F1 and F2 as it is for F3 raises the issue of how much community- and speaker-specific information is encoded in each of the formants. The speaker-discriminatory potential of individual elements of formant trajectories and potential explanations for the LR results are explored below.

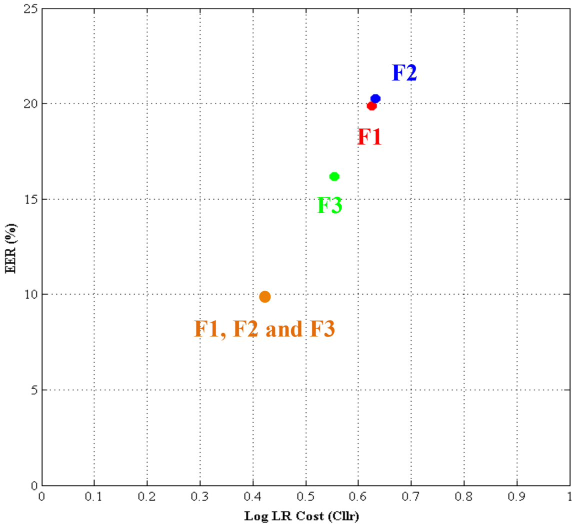


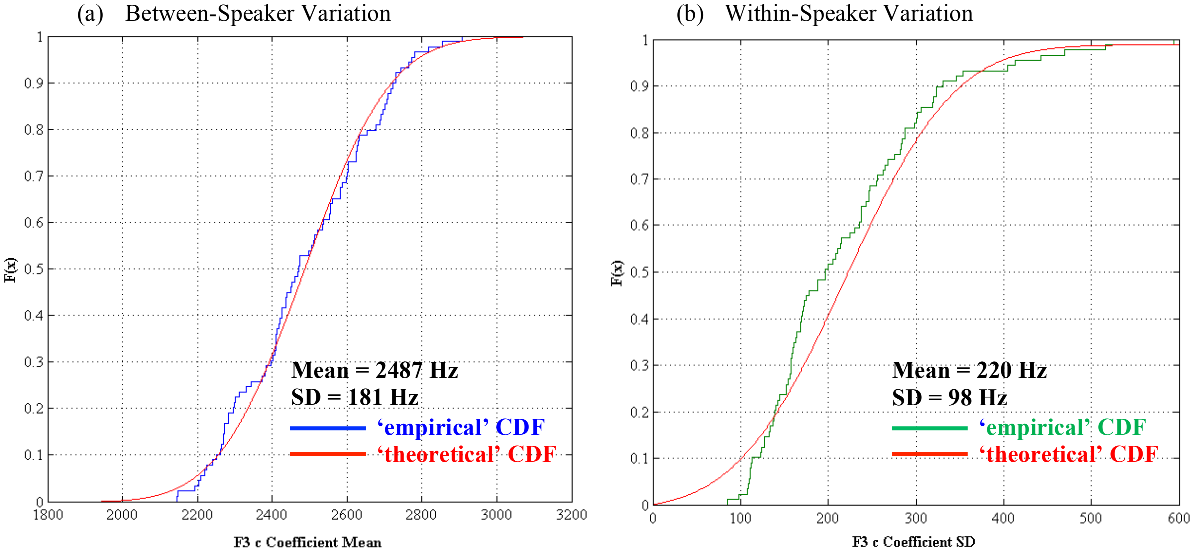
FIGURE 4. Comparative performance of quadratic coefficients of F1~F3 (orange), F1-only (red), F2-only (blue) and F3-only (light green) systems based on C_{llr} plotted against EER

Discussion

To address why F3 does not perform markedly better than F1 and F2 in LR-based testing, the levels of within- and between-speaker variation in the F3 constant (F3 c) are assessed. The constant term is related to absolute frequency across the formant trajectory and F3 c was chosen on the basis of the highest mean F-ratio generated from univariate ANOVAs when performing DA. Following Rose et al (2006) between-speaker variation is quantified using the variance (SD) in the distribution of mean F3 c values by-speaker and within-speaker variation is quantified using the central tendency of SD values. The variance ratio (VR) is defined as between-speaker variation divided by within-speaker variation where values > 1 = between-speaker > within-speaker, and < 1 = within-speaker > between-speaker.

Figure 5.(a) displays the cumulative distribution function (CDF) of mean F3 c values for 89 speakers based on 10 tokens per speaker. The CDF was plotted first using the raw data ('empirical' CDF) and then fitted with a normal distribution to make the results more generalisable ('theoretical' CDF). There is a wide distribution of mean F3 c values over a range of 762Hz. This is considerably greater than the between-speaker distribution of mean constant terms for F1 (277Hz) and F2 (391Hz). The between-speaker variation in F3 c (SD = 181Hz) is therefore promising for speaker-discrimination.

However, Figure 5.(b) reveals that the distribution of within-speaker variation also has a large range, with SDs spread maximally between 85Hz and 594Hz, but the distribution is positively skewed. As such, the mean of the SDs, as a measure of within-speaker variation (220Hz), provides an overestimation of the central tendency. Assessing the histogram in Figure 6., it appears that the modal bin provides a more appropriate approximation of within-speaker variation (between 153Hz and 161Hz). Using the mid-point value of the modal bin (157Hz) to act as a single-point estimate to calculate the VR, between-speaker variation is only marginally higher than that within speakers (VR = 1.153). These findings are consistent with the results of LR-based testing, in that even when using the best-performing speaker-discriminatory coefficient and a conservative estimate of within-speaker variability, inter-speaker variation is only marginally higher than intra-speaker variability.


(a) Between-Speaker Variation

(b) Within-Speaker Variation

FIGURE 5. 'Empirical' and 'theoretical' (normal distribution) CDF of mean (between-speaker variation) (a) and SD (within-speaker variation) (b) of F3 c coefficient values from 89 speakers

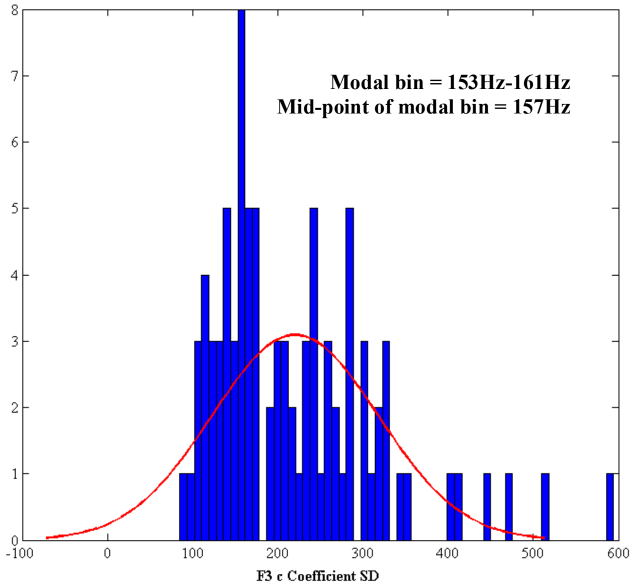


FIGURE 6. Histogram of F3 c SD values (using 60 bins of 8.49Hz each) fitted with a non-modified normal distribution

As such, F3 in these data does not appear to conform strongly to the requirement that a good speaker discriminant should display high between-speaker variation and low within-speaker variation (Nolan 1983). Further, the results are based on one style of speech from one recording with a single interlocutor recorded directly in studio quality. Therefore there is considerable scope for greater within-speaker variability according to different stylistic factors.

Correlations

Pearson correlations were calculated using mean coefficient values by-speaker in order to test whether elements of the F3 trajectory are correlated with F1 and F2. Since F1 and F2 carry considerable social-indexical information specific to the speech community, any correlation with F3 will necessarily impact on speaker-discriminatory potential. Table 1. displays all significant correlations involving F3 coefficients. Most significantly Table 1. reveals correlations between F3 c and both F1 and F2 c coefficients. The correlations between the c terms are somewhat predictable in that formant frequencies must remain acoustically distinct.

TABLE 1. Significant correlations (p<0.05) involving F3 coefficients using mean values by-speaker

		p-value	rho
F1 c	F3 c	0.0015	0.3313
F2 bx	F3 c	0.0069	0.2846
F2 c	F3 bx	0.0015	0.3318
F2 c	F3 c	0.0029	-0.3126

The interaction between F3 c and coefficients of F1 and F2 can possibly be explained in two ways based on the finding that F3 c offers the most towards speaker-discrimination. Firstly, the initial prediction for LR-based testing was that F3 would markedly outperform the lower formants because F3 is more closely linked to speaker-specificity. Following this assumption the results may be interpreted as better than anticipated LR performance for F1 and F2, in part, due to the relationship between their coefficients and the best performing speaker-discriminatory predictor (F3 c). This offers some evidence that F1 and F2 trajectories encode 'extrinsic' speaker-specific information, by virtue of a correlation with a stronger F3 speaker-predictor.

According to the second interpretation, if F1 and F2 are primarily, if not categorically, responsible for defining community norms, then their correlation with F3 c suggests that F3 also encodes 'extrinsic' information about the community, rather than exclusively being associated with the speaker. Such extrinsic community information encoded in F3 may account for the low variance ratio and poorer than expected LRs..

Choosing Discriminant Parameters for FVC

These findings highlight a fundamental issue in FVC which relates to the choice of discriminant parameters. F3 discrimination may be somewhat poorer than expected here, in part, due to the specific phoneme under investigation. In particular the front, close offset in the second element of /aɪ/ offers greater potential for interaction between F2 and F3 since high F2 encroaches on the potential range of F3, thus forcing F3 higher. It is therefore likely that vowels with no front, close element will have F3 trajectories which are more independent of F2, leading to better F3 performance. Further, the better than expected performance of F1 and F2 may be due to the specific dialect used, with SSBE offering greater potential for between-speaker variation on the open-close/ front-back dimensions specifically at the onset of /aɪ/. Therefore in identifying the best general discriminant or parameters to investigate, it is essential to acknowledge that the speaker space is constrained by anatomical, articulatory, social-indexical and phonological factors all of which determine the speaker-discriminatory power of a parameter in a given speech community.

Moreover, Rose (2006) claims that "not all speakers differ from each other in the same way." Therefore, as with many traditional phonetic-acoustic parameters (such as fundamental frequency (f0) (Hudson et al 2007) and articulation rate (AR) (Gold 2012a, 2012b)), the variance ratio for /aɪ/ F3 in SSBE suggests that its contribution lies primarily at the tails of the distribution where values in the KS and DS are atypical with regard to the population models of within- and between-speaker variation. Given the LRs based on F1, F2 and F3 combined, inherently poor discriminants may still be able to play an important role as part of a componential FVC analysis.

Conclusion

The comparative DA performance of individual formants has revealed a number of practical problems with the application of posterior probability metrics to FVC tasks. This highlights that considerable care should be taken when using DA as a metric of speaker discrimination. Using the LR, F3 was found to outperform F1 and F2 only marginally in the magnitude of strength of evidence, C_{llr} and EER. Further, it was found that whilst the distribution of between-speaker mean values for F3 c is rather large, and considerably larger than those for the c terms of F1 or F2, the level of within-speaker variation was only marginally smaller resulting in good separation of SS and DS pairs but low LRs. The results of this study support a more conservative approach to the speaker-discriminatory value of F3 in FVC casework and a consideration of the biological, linguistic, social and stylistic factors which constrain the speaker space.

ACKNOWLEDGMENTS

This research is funded by a UK Economic and Social Research Council DTC scholarship. Thanks to Paul Foulkes, Dominic Watt and Peter French for comments and suggestions on earlier drafts of this paper. Data extraction and analysis was facilitated by scripts written by Caroline Williams and Toby Hudson, Ashley Brereton, Geoffrey Morrison, Philip Harrison, Anil Alexander and Niko Brümmer.

REFERENCES

Aitken, C. G. G. and Lucy, D. **(2004)**. "Evaluation of trace evidence in the form of multivariate data," App. Stat. **54**, 109-122.

Alderman, T. **(2004)**. "The Bernard data set as a reference distribution for Bayesian likelihood ratio-based forensic speaker identification using formants," Proceedings of the 10th Australasian Conference on Speech Science and Technology. Sydney, Australia. 8-10 December 2004. 510-515.

Brümmer, N. and du Preez, J. **(2006)**. "Application-independent evaluation of speaker detection," Computer Speech and Language **20**, 230-275.

Champod, C. and Evett, I. W. **(2000)**. "Commentary on A.P.A. Broeders (1999) 'Some observations on the use of probability scales in forensic identification," Forensic Linguistics **7**, 238-243.

Delattre, P. and Freeman, D. C. **(1968)**. "A dialect study of American r's by x-ray motion picture," Linguistics, an international review" **44**, 29-68.

Esling, J. H. and Dickson, B. C. **(1985)**. “Acoustical procedures for articulatory setting analysis in accent,” in *Papers from the Firth International conference on Methods in Dialectology* edited by H. J. Warkentyne (University of Victoria, British Columbia), pp. 155-170.

Gold, E. **(2012a)**. “Articulation rate as a discriminant in forensic speaker comparisons,” UNSW Forensic Speech Science Conference. Sydney, Australia. 3 December 2012.

Gold, E. **(2012b)**. “The evidential value of articulation rate in forensic speaker comparison,” BBfor2 Short Summer School in Forensic Evidence Evaluation and Validation. Madrid, Spain. 18 June 2012.

Hudson, T., de Jong, G., McDougall, K., Harrison, P. and Nolan, F. **(2007)**. “F0 statistics for 100 young male speakers of Standard Southern British English,” Proceedings of the 16th International Congress of Phonetic Sciences. Saarbrücken, Germany. 6-10 August 2007. 1809-1812.

Hughes, V., McDougall, K. and Foulkes, P. **(2009)**. “Diphthong dynamics in unscripted speech,” Paper presented at International Association of Forensic Phonetics and Acoustics conference. Cambridge, UK. 2-5 August 2009.

Kinoshita, Y. **(2001)**. *Testing realistic forensic speaker identification in Japanese: a likelihood ratio-based approach using formants* (Ph.D. dissertation, Australian National University, 2001).

Ladefoged, P. **(2006)**. *A course in phonetics (5th edition)* (Wadsworth Cengage Learning, Boston).

Laver, J. **(1994)**. *Principles of phonetics* (Cambridge University Press, Cambridge).

Lindau, M. **(1978)**. “Vowel features,” Language **54**, 541–563.

McDougall, K. **(2004)**. “Speaker-specific formant dynamics: an experiment on Australian English /aɪ/,” International Journal of Speech, Language and the Law **11**, 103-130.

Morrison, G. S. **(2007)**. “MatLab implementation of Aitken and Lucy’s (2004) forensic likelihood ratio software using multivariate-kernel-density estimation,” Downloaded: 31st May 2011.

Morrison, G. S. **(2008)**. “Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/,” Journal of Speech, Language and the Law **15,** 249-266.

Nolan, F. **(1983)**. *The phonetic bases of speaker recognition*. (Cambridge University Press. Cambridge).

Nolan, F. **(1991)**. “Forensic phonetics,” Journal of Linguistics **27**, 483-493.

Nolan, F., McDougall, K., de Jong, G. and Hudson, T. **(2009)**. “The DyVis database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research,” International Journal of Speech, Language and the Law **16**, 31-57.

Peterson, G. E. **(1959)**. “The acoustics of speech – part II: acoustical properties of speech waves,” in *Handbook of Speech Pathology*, edited by L, E. (Peter Owen, London), pp. 137-173.

Rose, P. **(2002)**. *Forensic Speaker Identification* (Taylor and Francis, London).

Rose, P., Kinoshita, Y. and Alderman, T. **(2006)**. “Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/,” Proceedings of the 11th Australasian International Conference on Speech Science and Technology. University of Auckland, New Zealand. 6-8 December 2006. 329-334.

Rose, P. and Morrison, G. S. **(2009)**. “A response to the UK Position Statement on forensic speaker comparison,” Int. J. Sp. Lang. and the Law **16**, 139-163.

Saks, M. J. and Koehler, J. J. **(2005)**. “The coming paradigm shift in forensic identification science,” Science **309**, 892–895.

Simpson, S. **(2008)**. *Testing the speaker discrimination ability of formant measurements in forensic speaker comparison cases* (MSc dissertation, York, UK, 2008).

Stevens, K. N. **(2001)**. *Acoustic phonetics* (MIT Press, Cambridge, MA).

Tabachnick, B. G. and Fidell, L. S. **(2007)**. *Using multivariate statistics (5th edition)* (Pearson. Boston).