Cantonese forensic voice comparison with higher-level features: likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone

Phil Rose [1], *Xiao Wang*

[1] Australian National University Emeritus Faculty

philjohn.rose@gmail.com, hollamigo@gmail.com

Abstract

A pilot experiment relating to estimation of strength of evidence in forensic voice comparison is described which explores the use of higher-level features extracted over a disyllabic word as a whole, rather than over individual monosyllables as conventionally practiced. The trajectories of the first three formants and tonal F0 of the hexaphonic disyllabic Cantonese word *daihyat* 'first' from controlled but natural non-contemporaneous recordings of 23 male speakers are modeled with polynomials, and multivariate likelihood ratios estimated from their coefficients. Evaluation with the log likelihood ratio cost validity metric C*llr* shows an optimum performance is obtained, surprisingly, with lower order polynomials, with F2 requiring a cubic fit, and F1 and F3 quadratic. Fusion of F-pattern and tonal F0 results in considerable improvement over the individual features, reducing the C*llr* to ca. 0.1. The forensic potential of the *daihyat* data is demonstrated by fusion with two other higher-level features: the F-pattern of Cantonese /i/ and short-term F0, which reduces the C*llr* still further to 0.03. Important pros and cons of higher-level features and likelihood ratios are discussed, the latter illustrated with data from Japanese, and two varieties of English in real forensic casework.

Index Terms: Forensic voice comparison, likelihood ratio, Cantonese, higher-level features, F-pattern trajectories, tonal F0 trajectory.

1. Introduction

1.1. Forensic Speaker Recognition and Likelihood Ratios

The *Criminal Procedure Rules* relating to expert evidence in the United Kingdom [1] stipulate front and centre the expert's first duty is to the court, and that duty is to help it by giving objective and unbiased opinion within their area of expertise. The UK is obviously not the world's only jurisdiction, but these requirements are sufficiently reasonable to express their application to Forensic Speaker Recognition thus: the expert's primary aim is to compare suspect and offender speech samples to help the trier-of-fact decide whether the suspect said the questioned, often incriminating, speech. The crucial word here is *help*, and it is crucial in two ways (I defer discussion of the second to the end of the paper). For there is currently some disagreement as to how that help is construed – specifically whether or not the expert should help by furnishing the interested parties with an estimate of the strength of evidence by way of a ratio of conditional probabilities of the speech evidence under competing defense and prosecution hypotheses. In other words, whether or not they should estimate a likelihood ratio (LR) [2].

This so-called *logical approach* to Forensic Speaker Recognition (*rational approach* would be better) emerged in the late 1990's and is now often called *likelihood ratio-based forensic voice comparison* (LR-FVC), It became, after DNA, part of the new paradigm for the evaluation of forensic evidence [3]. It is interesting, in retrospect, to note how well the emergence of LR-FVC has conformed to the two Kuhnian stages for new paradigms of out-of-hand rejection and ridicule [4].

Like any good paradigm should, it has guided a lot of research over the past decade or so. During this time, emphasis has shifted from (successfully) trying to demonstrate that LR-FVC can indeed emulate the DNA benchmark [5], with both automatic and acoustic-phonetic approaches [6] to solving problems within the new paradigm that will enable improvements in the use of the LR in real case-work in different languages. For when experts do case-work they want to know what is the best method to adopt under the circumstances of the case – it is well to remember that there is no more compelling reason for FSR research than this. Typical LR-FVC research questions address the inevitable problems with reference sample uncertainty, choice and mismatch [7,8,9,10]; the vexed question of likelihood ratio precision [10]; and the suitability of different kinds of models and frontends [11,12,13,14,15]. This paper is a very modest example of the latter, using a higher-level, acoustic-phonetic frontend – formant and fundamental frequency trajectories – processed with a multivariate likelihood ratio model.

The paper's specific objectives are to see what kind of performance can be obtained if one processes the trajectories of formants and F0 of a disyllabic word as a whole, rather than their realization over the two separate constituent syllables. The contribution to an ASR system of formant trajectories in diphones was shown, rather convincingly, in [16]. The particular word used in this paper has many more segments. It contains two diphthongs, and two tones, and is therefore effectively hexaphonic. A subsidiary part of the research question is whether different formants require different degrees of trajectory modeling. Since the language under investigation is tonal, a separate question concerns to what extent tonal F0 trajectories over two syllables are of forensic use.

1.2. Validation

The main function of a forensic LR is to tell the trier-of-fact, or investigating parties, how strong the evidence before it

is. Its other function, as in this paper, is the essential one of so-called *validation* [17], i.e. testing the validity of the system, especially the features used, to compare samples. One merit of the LR approach is that it allows a system to be validated in a forensically realistic manner, and the discriminability of various forensic media have now been tested in this way, e.g. DNA [18], fingerprints [19,20], handwriting [21], SMS texts [22] and, especially, speech [23]. To perform validation, LRs are estimated for sets of known same-subject and different-subject comparisons. To the extent that the features used to compare the samples are valid, same-subject comparisons will have $\log_{10}$LRs greater than 0, and different-subject comparisons $\log_{10}$LRs lesser than 0. To illustrate this second function, figure 1 shows the cumulative distribution of LRs from 33 same-speaker and 528 different-speaker comparisons on non-contemporaneous telephone recordings of male speakers of General Australian English. These data were part of a real-world forensic voice comparison case that went to trial in 2007 involving a $150 million telephone fraud [2]. LRs from different-speaker comparisons increase towards the left; same-speaker LRs towards the right. A system is under validation comprising two acoustic-phonetic features. One is the F-pattern in the word *yes* as quantified by point measurements of the first three formants at onset, mid-point and offset – LRs from this feature are shown by the blue dotted lines. The other, shown by green dashed lines, is the intonational F0, realizing [H/L.L.LH] pitch, on the phrase *not too bad*. The LRs for the fused system are shown with thicker, solid red lines.

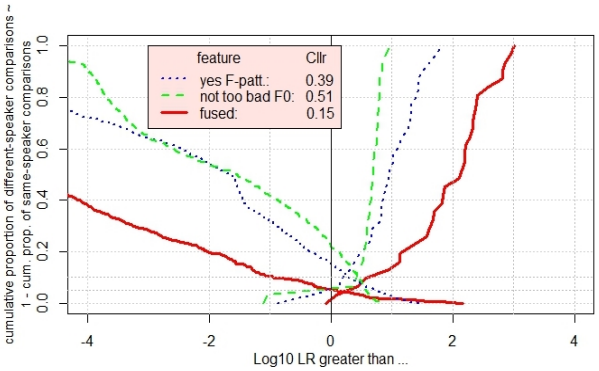

Figure 1: *Tippett plot for Validation LRs derived from comparisons based on intonational F0 in* not too bad *and F-pattern in* yes*. Legend shows* Cllr *values for individual and fused features. cum. prop. = cumulative proportion.*

The separation around $\log_{10}0$ in this so-called *type I Tippett* plot conveys visually that LRs based on the fusion of *yes* F-pattern and *not too bad* F0 improve on the individual features and can discriminate fairly well between same-speaker and different-speaker speech samples: the error rates for same-speaker and different-speaker comparisons based on the fused LRs are ca. 2% and 5% respectively.

Strictly speaking, however, the use of error rates with LRs is incorrect: by Bayes' theorem a prior probability is still required to decide whether the suspect said the incriminating speech. (There is therefore no speaker *recognition* involved in likelihood ratio-based forensic voice *comparison* – a misunderstanding surely encouraged by its superordinate term *Forensic Speaker Recognition*). However, assuming flat priors for convenience, error rates still remain useful as indicators of discriminative power, and *Daubert* requires them. The performance of a system like this, equivalently its validity, is currently assessed by the information-theoretic log likelihood ratio cost C*llr* [24] which is now one metric for LR-based detection systems. Figure 1 shows all features have C*llr*s below unity and so contribute information. The *not too bad* F0 LRs are the weakest, with a C*llr* of 0.51, the *yes* F-pattern LRs are a little stronger (C*llr* = 0.39), and the C*llr* for the fused system is quite low, at 0.15.

Tippett plots can also give interested parties useful indication of the kind of strength of evidence to be expected from a set of features. Figure 1 suggests that the average strength of evidence in favour of same-speaker hypotheses from F-pattern in *yes*, combined with intonational F0 in *not too bad*, will not be very big (ca. $\log_{10}2$), so the prior from other evidence will have to be more advantageous than ca. 1 in 12 to result in posteriors above 90%.

1.3. Polyphthongs

The strength of evidence from a forensic voice comparison system reflects, among other things, the amount of speaker-dependent information in the features used to compare the speech samples. The superiority of quantification of vocalic F-pattern in terms of formant trajectories, parametrised by either DCT or polynomial coefficients, rather than traditional acoustic-phonetic point measurements [25,26,27,28] is an example. This shows, presumably, that speakers can differ not just in their realization of the vocalic targets involved, but in how they get from one target to another. As part of this, previous research in the forensic comparison of vowel acoustics, e.g. [28, 29], has suggested that strength of evidence, as quantified by C*llr*, increases with complexity of F-pattern, where complexity is defined in terms of number of vocalic targets. Thus monophthongal F-pattern, with a single vocalic target in a syllable, does not, in general, yield such strong evidence as diphthongal F-pattern with two targets [30] and triphthongal, with three. It is likely that triphthongs are also superior to diphthongs, but the appropriate comparisons, e.g. between [ia] and [iau], have not yet been done.

Up to now, LR-based comparisons with vowel acoustics have been performed on the F-pattern of monosyllables. But offender and suspect speech samples often contain the same polysyllabic words and expressions. The telephone fraud case mentioned above, for example, made use of the phrase *not too bad* said by both suspect and offender. What if polysyllabic words were treated as polyphthongs, and their vocalic formant trajectories quantified globally over the whole word rather than separately, over its constituent syllables? What strength of evidence would that yield? That is the specific question asked in this paper, using the word for *first* in Cantonese: *daihyat* /tai.jat L.H/. The main reason for asking this question is that more vowel targets might offer more possibilities for speakers to differ in their trajectories between them. There might also be cases – *daihyat* is one – where there is no clear acoustic demarcation between the syllables to facilitate separate measurement. Measuring one F-pattern rather than two would also save time in real case-work.

Cantonese, of course, is a tone language. Thus *daihyat* also has tonological structure, and its F0 will reflect the pitch targets of its constituent tones. Experiments with monosyllabic tonal F0 in running speech have shown very little forensic potential [28,29]. It is of interest, therefore, to see whether the tonal F0 trajectory over two syllables is any different, and this is another aim of the paper.

The paper further aims to demonstrate the general forensic potential of higher-level features by fusing the *daihyat* results with two others. Since the paper is about LRs from higher-level features it then concludes with a discussion of some of their important pros and cons.

2. Procedure

2.1. Test word

Daihyat was chosen because it represented a useful compromise between F-pattern complexity and practicality of elicitation. Both its syllables have diphthongs: a closing /ai/ in the first, a rising /ja/ in the second, so its phonological structure involves four F-pattern targets - /a/ /i/ /j/ and /a/. To realize this structure in a carefully spoken word, the speaker's supralaryngeal articulatory mechanism goes from half-open central unrounded [ɐ] to close high front unrounded [j], and back to [ɐ], and thus traverses about two-thirds of its height dimension, and about half its backness. As far as *daihyat*'s tones are concerned, conservative Cantonese varieties have six tonemes [31] and *daihyat* has two of them, one on each syllable. Its first syllable has a low /L/ toneme, realized by a long low level pitch, and the second syllable a high /H/ toneme, realized with a short high-pitched allotone, its short duration conditioned by the syllable-final obstruent [ˀtˀ].

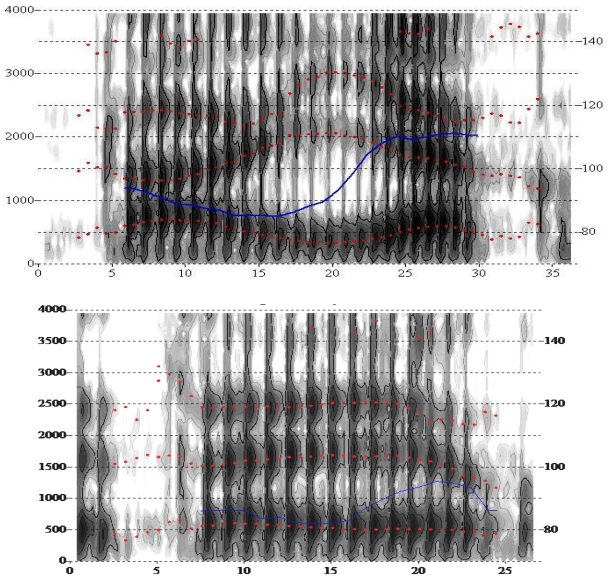

Figure 2: *Wideband spectrograms of two* daihyat *tokens from the running speech of the same speaker. Red dots = superimposed formant centre-frequencies, blue line = F0. X-axis = duration (csec.), left & right vertical axes = spectrographic frequency & F0 (Hz).*

The top panel of figure 2 shows a spectrogram of a clearly spoken *daihyat* from the experiment. In this token the phonological targets for /ai/ and /a/ are minimally and intrinsically perturbed: [tɐɪ jɛ̌tˀ]. Its time-varying F-pattern is clear and relates straightforwardly to the articulation, with F1 and F2 inversely proportional to vowel height and backness respectively. The vocalic portion of their F-pattern changes sign only once, corresponding to the highest and frontest articulatory position of the /j/ target. In this example peak F3 occurs slightly after peak F2, indicating a shortening of front cavity, presumably by lengthening of the constriction between tongue and palate for [j]. The F3 trajectory also changes sign at ca. csec. 14. This is the point where the affiliation changes between front and back cavities and F2 and F3 [32]. (Although the F2 and F3 trajectories give the visual impression of crossing over, this cannot happen physically because the cavities with which they are associated are acoustically coupled.) Typical consonantal perturbations to the F-pattern from the alveolar obstruents are also seen.

The token's F0, shown by the blue line, nicely reflects the [low level] – [short high level] tonal pitch targets, with the transition between high and low targets occurring mostly in the portion corresponding to the second syllable onset [j]. The slightly falling F0 on the first syllable is not perceivable as pitch and is probably an intrinsic perturbatory effect from the initial voiceless unaspirated obstruent [t].

2.2. Corpus, speakers & Hong Kong MTR database

The forensic potential of *daihyat* disyllabic F-pattern and tonal F0 was tested with 23 young Cantonese male speakers from the *Hong Kong Mass Transit Railway* (HKMTR) database. This is a small (ca.50 speakers), quasi map-task database primarily designed as a resource for controlled but natural speech for testing likelihood ratio-based approaches to forensic voice comparison in Cantonese. It was collected in 2012 as part of a go-to-whoa Cantonese FVC experiment run as a one-semester postgraduate course at the Hong Kong University of Science and Technology [33]. In addition to its obvious educational aim (learning how to properly evaluate the strength of evidence supporting competing hypothesis should surely be part of every student's education), the course had another objective. The continuing poor understanding of LRs, especially on the part of the legal profession [2], suggests that the LR framework actually might be much more of a hindrance to the court than the help mandated by the UK *Criminal Procedure Rules*. I wanted to see, therefore, whether it was indeed too difficult to be learnt. The six subsequently published papers from the course participants e.g. [28,29], and the second author's thesis [34] (upon which this paper builds), rather suggested otherwise.

Subjects were given a map of the HKMTR and asked various questions about it, among which was whether a particular station was the first (*daihyat*) or second after another. For this paper we used responses to the question where the two stations were adjacent and therefore the answer was *daihyat*. Subjects were instructed to answer in a whole sentence, so a typical exchange between experimenter and subject was:

Q: Jimsajéui haih Jódan jihauh daihyatgo dihnghaih daihyihgo jaahm a?

尖沙嘴係佐敦之後第一個定係第二個站阿

Is Tsimshatsui the first or second station after Jordan?

A: Jimsajéui haih Jódan jihauh **daihyat** go jaahm.

尖沙嘴係佐敦之後**第一**個站

Tsimshatsui is the first station after Jordan.

It is, of course, an essential component of a forensic voice comparison database to include non-contemporaneous recordings [35]. For this experiment speakers were recorded on two occasions separated by about a month. Following a protocol for collection of forensic speech data [36], participants communicated by phone while high quality recordings were made from lapel microphones. An average of eight *daihyat* replicates was measured per speaker per session.

To exemplify the kind of within- (and between-) speaker variation observed in the HKMTR corpus, the bottom panel of figure 2 shows a spectrogram of a *daihyat* token from the same speaker, and the same session, as the first. Said more quickly, it displays much less differentiation in F-pattern and F0, and sounds monosyllabic: [tɛɪ].

2.3. Processing

Speakers' *daihyat* tokens were identified aurally, and wideband spectrograms of them generated in *Praat* as in figure 2, with superimposed formant and F0 traces. The sampling base for the F-pattern and F0 was determined by eye from the wide-band spectrogram with its good time-domain resolution. Onset was taken to be at the first strong glottal pulse of /ai/ in *daih-*. In the top panel of figure 2 this pulse occurs at ca. csec. 6. Offset was adjudged at the last strong glottal pulse of /ja/ in *yat* (at ca. csec. 29 of the top panel of figure 2). Usually at most the first three formants will be of use forensically and these were extracted with *Praat*'s Burg option. It was found that a setting of four formants below four kHz usually gave acceptable results, in the conventional acoustic-phonetic sense of the extracted formant centre-frequencies appearing to track through the middle of the spectrographic formants – as can be seen in figure 2 – and this was adopted as default. When this was not the case, the settings were changed to get a better visual resolution. This usually involved increasing the number of formants to extract to five.

2.4. Parametrization

The raw F-pattern trajectories were modeled by permuting polynomials of degree from one to cubic separately on each formant, and their coefficients extracted for LR processing. It was confirmed that, as previously e.g [25], better results are obtained when polynomials are calculated from normalised as opposed to raw duration. Panel A of figure 3 shows the extracted raw F-pattern for a speaker's 12 *daihyat* tokens in a single recording. A certain amount of F-pattern target undershoot as a function of raw duration is visible. Panel B shows the corresponding equalized (not normalized!) duration cubic polynomial trajectories for the three formants, with the mean polynomial overlaid in a thick black line. Panel C compares the mean polynomial F-pattern of the speaker's two non-contemporaneous recordings. It can be seen that, plotted as a function of equalized duration, the first recording (black line) shows a little undershoot relative to the second, but otherwise they agree fairly well. Panels D – F show the same thing for the speaker's tonal F0, where it can be seen that the longer tokens tend to have a higher second-syllable tonal F0 target, and that the mean F0 of the second recording lies higher than that of the first.

Likelihood ratios were estimated using the multivariate kernel-density likelihood ratio (MVKD) formula developed at the *Joseph Bell Centre for Forensic Statistics and Legal Reasoning* [37], which has been used in many previous studies as well as real-world case-work [2]. As explained in [13], this formula compares the *Mahalanobis* distance between the suspect and offender mean vectors against measures derived from the same- and different-speaker (co)variances (i.e. sampling errors of the mean) to determine the LR. In this way it takes into account any correlation between variables. The ratio of between- to within-speaker (co)variances acts as a major scaling factor. Although it is called a LR formula, it is best to assume it outputs a *score* quantifying the ratio of the similarity of the difference between suspect and offender samples to their typicality given a suitable reference sample. It is usual then to calibrate these multivariate scores with logistic regression to convert them to true LRs [20]. Scores for the tonal F0 and each formant were obtained separately, and then fused/calibrated using the first author's *R* implementation of the *focal* toolkit [38].

Each speaker's mean F-pattern and F0 of their first recording was compared with their second recording (as in figure 3 panels C,F) to get 23 known same-speaker LRs, and with the F-pattern and F0 of the other speakers' first recordings to get 253 known different-speaker LRs. Leave-one-out cross-validation was used, whereby all data for the particular pair being tested are removed from the reference sample used to estimate the typicality of the comparison.

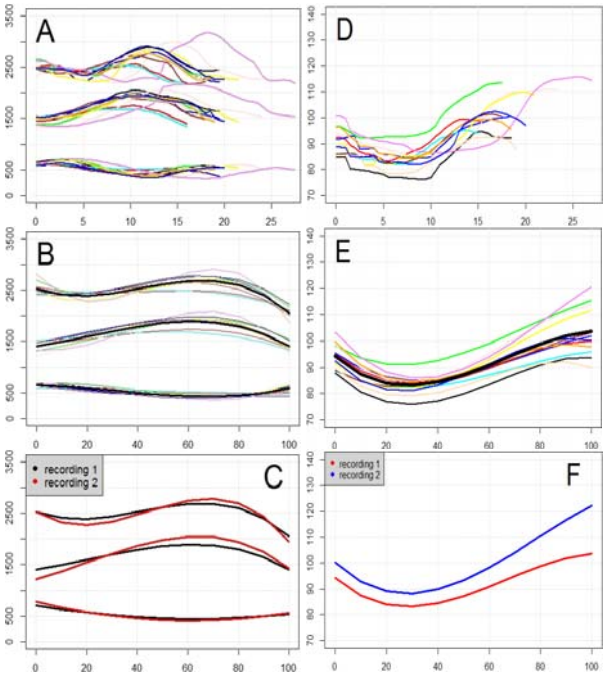

Figure 3: *Stages in* daihyat *acoustic data extraction and comparison. A,B,C = F-pattern, D,E,F = tonal F0. Y-axis = frequency (Hz). X-axes: (A,D) = duration (csec.), (B,C,E,F) = equalized duration (%).*

3. Results

Optimum C*llr* values for the various features and their fusions are shown in table 1. The order of the polynomial with the best C*llr* is also given. Figure 4 has the corresponding Tippetts.

Table 1: *Optimum Cllr values for daihyat MVLR discrimination.* Fused1 = *fused F-pattern,* fused2 = *fused F-pattern and F0.* Quad., cub. = *polynomial degree with best Cllr.*

F1	F2	F3	fused1	Tonal F0	fused2
0.43 (quad.)	0.43 (cub.)	0.63 (quad.)	**0.16**	0.53 quad.=cub.	**0.10**

The first thing to note is that, perhaps surprisingly, optimum C*llr* was not obtained with the highest degree polynomials tested, the best combination being quadratic for F1 and F3 and cubic for F2. This resulted in a C*llr* for the

fused F-pattern of 0.16. (Modeling the F-pattern trajectories with all cubics resulted in a considerably higher C*llr* of 0.24, presumably by overfitting; and an all-quadratic fit gave a C*llr* of 0.22, presumably by underfitting. Single degree polynomials had a C*llr* of 0.3, again by underfitting.) There was therefore more speaker-dependent information in the trajectory of F2 than the other two formants. Looking at the spectrogram in the top panel of figure 2, it can be appreciated that the token's complex F3 trajectory could only be accurately modeled by a quintic; but from the point of view of forensic voice comparison that would constitute a considerable overfit. Evidently, there appear to be limits to the usefulness of phonetic detail.

As far as tonal F0 was concerned, quadratic and cubic modeling gave essentially the same results with a C*llr* of 0.53, on a par with the individual formants. Therefore the disyllabic tonal F0 in *daihyat* performs much better than monosyllabic tonal F0. Fusion of tonal F0 with best-performing F-pattern improves the C*llr* again to 0.1, showing that LRs from tonal F0 and F-pattern are not totally correlated.

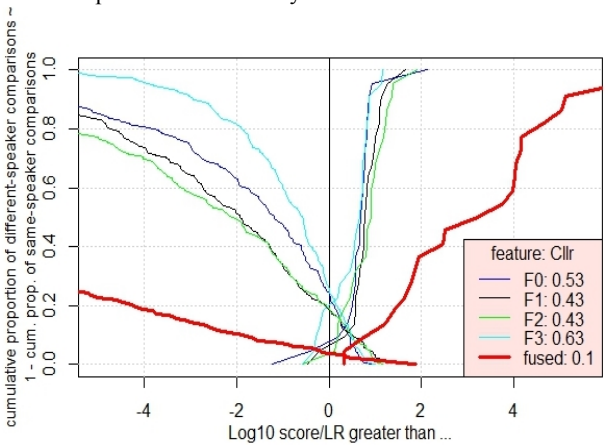

Figure 4: *Tippett plots of* daihyat *F-pattern and tonal F0. Legend shows* Cllr *values for individual and fused features.*

The Tippett plot in figure 4 shows that the word-based acoustic trajectories of hexaphone *daihyat* modeled in this way function quite well in distinguishing same-speaker pairs from different. The error rate for different-speaker comparisons is about 4%, and the 23 same-speaker comparisons are all correctly evaluated. It is worth noting that, the mean duration of a *daihyat* token being just under 20 csec., this performance is obtained on an average of about 1.5 seconds of net speech in each offender and suspect sample.

3.1. Spline analysis

Since the optimum C*llr* was obtained with relatively low order polynomials, it is unlikely that an approach quantifying the formant trajectories separately for each syllable, where one would expect a lower order to give good results, would be any better. This was confirmed by a spline analysis performed on the *daihyat* F-pattern data, whereby the vocalic portion of the word was divided into two sections corresponding (as best as possible with a continuous F-pattern) to its two syllables, scores derived for the F-pattern of each section separately, and then fused to obtain the overall LR. (In replicates with F2 and F3 maxima in mid-word, the boundary was located at the duration point mid-way between them; otherwise at mid-duration of the vocalic portion.) The best spline C*llr* obtained with this method was 0.21, and thus somewhat worse than the word-trajectory value of 0.16. This suggests that it may be advantageous, if the F-pattern of disyllabic words is to be compared in case-work, to model the trajectories over the whole word rather than separately on the constituent syllables.

3.2. Fusion with other higher-level features

Real-world case work can seldom rely on a comparison of a single feature, like, say, tonal F0, because the prior is unlikely to be usefully offset by the low magnitude of the expected LR. (Even with the most advantageous prior of 1 in 2, an average LR for a same-speaker segmental comparison of about 5 will mean a posterior of ca. 83%, which is far short of convincing.) A greater number of features is likely to force the LR further away from $\text{Log}_{10}0$ and thus provide more useful information. In order to demonstrate this, the *daihyat* data were logistic-regressively fused with scores from two more higher-level features from the MTR database. One feature was the F-pattern trajectory from the 23 speakers' /i:/ vowels in the word *yih* 'two' [29]. The other feature was the 23 speakers' short-term F0. (The short-term F0 scores were obtained from six parameters from the normal distribution of F0 over stretches of speech of about 40 – 90 seconds duration [39].)

Figure 5 shows the Tippett plots for the individual *daihyat*, /i/ F-pattern and short-term F0 features, with C*llr* values of 0.1, 0.64, 0.46 respectively. It can also be seen that fusion of the three features (thick red line) reduces the C*llr* to 0.03, with error rates of ca. 0.5% and 0% for different- and same-speaker comparisons respectively.

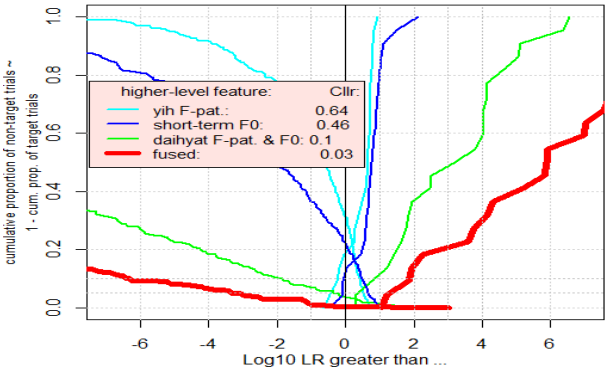

Figure 5: *Tippett plots for fusion of* daihyat, yih *and short-term F0 data. Legend shows* Cllr *values for individual and fused features.*

4. Discussion

4.1. Higher-level features

This paper has provided a further intimation of the forensic potential of higher-level features: they too contain speaker-dependent information. Higher-level features, characterised as involving either linguistic information or information from stretches of speech longer than the frame size used in cepstral-based systems [40], have been shown, for a relatively small amount of speech data, to contribute an improvement over ASR baseline performance e.g. [41,42,16]. The excellent performance of ASR baseline systems has shown that considerable speaker-dependent information resides in a long-term spectrum that has averaged-out linguistic information.

The fact that addition of higher-level features improves ASR baseline performance must mean that the two types of information are complementary. Complementary, not dichotomous: if they can be quantified, all types of information in the speech wave are grist for the LR mill and forensic voice comparison systems ideally should exploit them where practical.

In LR-FVC, the clear relationship of higher-level features to linguistic or phonological structure, but especially to articulatory phonetics, brings several benefits in addition to their acknowledged robustness and interpretability (which the descriptions in section 2.1 were meant to illustrate). Firstly, it allows the expert to better interpret the variability of their data and its effect on the performance of their system. Secondly, because higher-level features often reflect articulatory gestures deliberately implementing speech sounds, their variance is constrained by two main factors: the limitations on the size of the human vocal apparatus, and the limitations imposed by the sound being produced. The acoustic theory of speech production specifies, for example, that a normally-produced adult male [i] can be expected to have F1 somewhere in the 200 – 400 Hz range: you will not find, even with Helium, a normal adult male [i] F1 at, say, 1000 Hz. These limitations should also impose limits on the number of speakers necessary for a reference sample (which is also a function of the feature dimensionality involved). Although no-one I think has yet done so, it should also be possible to use such general phonetic information to specify quite strongly informative priors for reference distributions within a so-called *Fully Bayesian* approach to LR-FVC [10].

The downside of this double constraining of higher-level features is that their variance ratio will be poor. Since the variance ratio is the major term in both uni- and multivariate LR formulae, that means that one cannot normally expect great strength of evidence from single features.

4.2. Likelihood ratios in forensic reality

Considerable support can now be found, outside the Law, for the LR framework [2]. For example, the *Board of the European Network of Forensic Science Institutes*, representing 58 laboratories in 33 countries, has now endorsed LRs on page 3 of their best practice guidelines on forensic automatic and semi-automatic speaker recognition [43]. The leading forensic speech science company in the UK, for some time a not particularly LR-friendly place, has also now moved to adopt the standards of the UK *Association of Forensic Science Providers* in reporting LRs either numerically or in verbal equivalents. The majority of forensic speech science research published between 2010 and 2013 used quantitative measurement and statistical models to calculate LRs and empirically test system performance [44]. LRs based on higher-level acoustic-phonetic features were used in the prosecution of a $150 million telephone case that went to trial in Australia in 2007 [2]. Much more recently, there have been two published implementations of the new paradigm under conditions reflecting those of real Australian cases [44,45]. One important message from these real-world studies, I think, is that in forensic speaker recognition there is not a one-size-fits-all solution: one needs to use material appropriate for the competing hypotheses and conditions of the specific case.

Not that LRs are without problems. One of the benefits of the LR approach is that it enables the estimating of evidential strength from a combination of different features. One can, for example, easily combine LRs for the acoustics of different vowels, or vowels and F0 as in this paper, or different evidential types as in the identification of Richard III's skeleton [46]. Combinability is also a significant problem, however. The mantra is that the trier-of-fact is to combine LRs for the separate types of evidence received, but how are they supposed to combine these LRs with other, non-quantifiable evidence [2]? Another problem is variability in LRs as a function of the system. Figure 6 shows LRs generated for the same set of data (99 Japanese speakers' voiceless fricative spectrum) using both GMM/UBM and MVLR models [12]. The discriminant performance and amount of information from both systems is roughly the same, but the GMM/UBM provides much stronger evidence for same-speaker comparisons, whereas MVLR is stronger for different-speaker comparisons. The MVLR is superior only in the rather important property of maximum magnitude of false alarms.

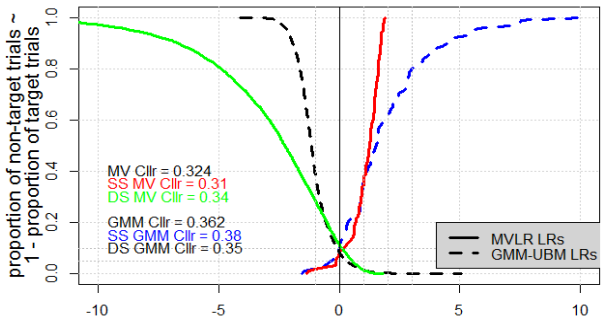

Figure 6: *Tippett plot for 99 speakers' multivariate (solid) and GMM/UBM (dashed) LRs derived from comparisons using cepstrally-mean-subtracted LPC CCs from Japanese [ɕ]. X axis = log10LR greater than … .*

There are also dissenting voices amongst ASR researchers. At the *Odyssey 14* session on *Speaker Recognition for Forensic Applications* the point was made that LRs, though fine in theory, have to be treated with great caution when output from black-boxy automatic systems that may not be using data comparable with the case at hand. This opinion was also heard at one *Interspeech 15* Special Event: *Speaker Comparison for Forensic and Investigative Applications*. "The LR framework was indeed logical, but, due to lack of data, could not be implemented properly now or in the foreseeable future." [I paraphrase here from OSAC email exchanges.] It was not clear whether these and similar comments were intended to apply only to fully automatic systems, or to *any* attempt to estimate LRs for forensic speech data in general.

Lack of data is certainly a problem, and this is where the second sense of the word *help* comes in mentioned at the beginning of the paper. Recall that the expert is supposed to help the trier-of-fact. What if you are a committed LR person but cannot estimate one because you do not have a reference sample? Do you give up? Here are some acoustic-phonetic data from the investigatory phase of two real cases, demonstrating that the expert can indeed provide help without estimating a numerical LR. Data from the first case are in figure 7, both panels of which show the trajectories for the first three formants in the diphthong /ei/ after /k/, as in the word *OK* for example. Thin lines are the individual replicates, thick lines are their mean cubic polynomial trajectories. The left panel shows the suspect data separately for two recordings: his police interview (thick red lines) and his arrest (thick black lines). The right panel shows the questioned data.

The suspect's data in figure 7 are unremarkable examples of an Australian /ei/ diphthong, the difference between the two recordings being plausibly conditioned by tempo: the suspect undershoots his first diphthongal target when speaking more quickly (an example of what I meant by the interpretability of higher-level feature variance by phonetic knowledge). The questioned data in figure 7 realize an auditory mid front monophthong [e] and as such are atypical for an Australian /ei/. Suspect and offender vowels differ enormously in their acoustic and auditory qualities. Other things being equal one would be more likely to observe this kind of difference if they had come from different speakers. How much more likely? You cannot say, because there is no reference sample. Neither is there ever likely to be one, because of the complex nature of the accents involved (West-African accented Australian). The large difference between the samples does not, of course, mean that they have come from different speakers; only that Prosecution will have a very hard time should they wish to so argue. This is useful information: It can help investigative authorities decide how best to handle the forensic voice comparison evidence in the light of the severity of the crime, the financial cost in prosecuting it, and the political cost in not.

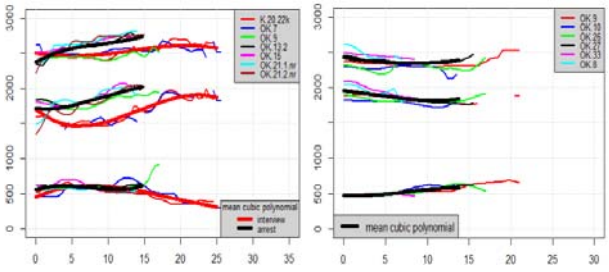

Figure 7: *Comparison of extracted F-pattern for /ei/ after /k/ in suspect and questioned voice samples. Explanation in text. X-axis = duration (csec.), y-axis = frequency (Hz).*

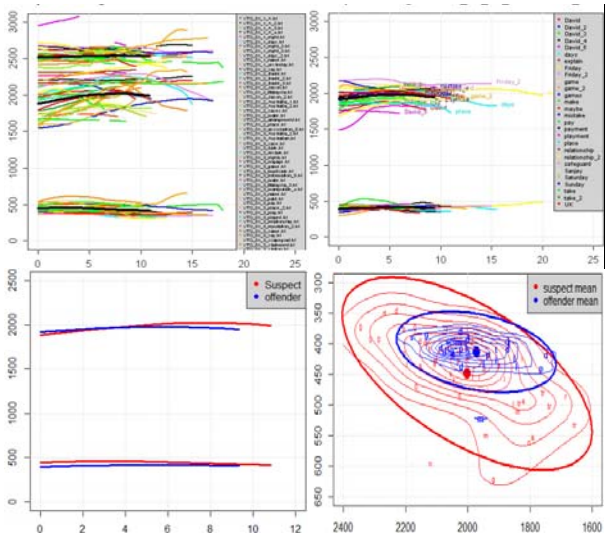

Figure 8: *Comparison of extracted F-pattern for /e:/ in suspect and questioned voice samples. Explanation in text.*

Data from the second case are in figure 8. Both suspect data, from an atypically superb quality police interview recording, and offender data, from mobile telephone intercepts, contained copious amounts of different vowels. The top row of figure 8 shows raw formant values and their mean polynomial trajectories extracted for mid front vowel /e:/ in suspect (left panel) and offender (right panel). The bottom row shows two types of comparison for the /e:/ F-pattern data. The left panel compares mean raw duration polynomial trajectories for suspect and offender F1 and F2. Assuming within-speaker variation is both typical and smaller than between-, this degree of similarity will result in a LR in favour of same-speaker provenance. How much more? Once again, you cannot say because no reference sample exists (a variety of Singaporean English was involved). The bottom right panel of figure 8 compares the joint distribution of suspect and offender values for F1 and F2 in /e:/ mid duration. Suspect values are shown in red, offender in blue. It could be pointed out to a trier-of-fact that the offender values all occur within the range of the suspect's values defined by his 95% confidence ellipsis, and also that, given the suspect density – modeled either bivariate-normally or with kernels – the probability of getting the difference between the suspect and offender means assuming the offender is the suspect, is quite high. It is of course possible in both these real-world cases to estimate the similarity term $p(E|H_p)$ of the LR. However, the expert should also make absolutely clear that a proper evaluation of the difference between suspect and offender mean values cannot be done absent the probability of the evidence under the alternative hypothesis. This is information that Defense should be expected to be capable of understanding and eliciting during cross. This, too, is useful information. It would be interesting to see information analogous to these two real-world examples provided by an automatic system.

5. Summary

In examining speaker-dependent information in that part of the audio biosignal concerned with speech sounds, this paper, an example of what some would now call *Forensic Semi-automatic SR* [43], has demonstrated that estimating LRs from the formant and F0 trajectories over a disyllabic word can indeed yield reasonable strength of evidence, and that it may be a viable approach in real case-work when such words exist. It has also shown that different formants may warrant different orders of polynomial for an optimum performance. Contrary to expectations, higher order polynomials did not achieve the best results, indicating that phonetic detail may constitute noise rather than signal; and so the idea of benefitting from between-speaker differences in trajectories linking complex vocalic targets has not been borne out. Finally, the comparison of acoustic-phonetic higher-level features was shown to still be of use even though absence of reference sample renders LR estimation problematic.

6. References

[1] UK Ministry of Justice Website, Rules and Practice Directions 19: Expert Evidence.http://www.justice.gov.uk/courts/procedure-rules/criminal/docs/2015/crim-proc-rules-2015-part-19.pdf

[2] P. Rose, "Where the science ends and the law begins – likelihood ratio-based forensic voice comparison in a $150 million telephone fraud", *Int'l J. Speech Language and the Law* 20/2, pp. 277-324, 2013.

[3] G.S. Morrison, "Forensic voice comparison and the paradigm shift", *Science & Justice* 49, pp. 298-308, 2009.

[4] G.S. Morrison, "Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework", *Australian J. Forensic Sci.*, 41, pp. 155–161, 2009.

[5] D.J. Balding, *Weight of Evidence for Forensic DNA Profiles*, Wiley, Chichester, 2005.

[6] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. Torre and J. Ortega-García, "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition", *IEEE Transactions on Audio Speech and Language Processing* 15/7, pp. 2104-2115, 2007.

[7] G.S. Morrison, F. Ochoa and T. Thiruvan, "Database selection for forensic voice comparison", *Odyssey* 2012, pp. 62-77.

[8] S. Ishihara, "Replicate mismatch between Test/Background and Development Databases: The impact on the Performance of Likelihood Ratio-based Forensic Voice Comparison", *Interspeech* 2014, pp.393-397.

[9] V. Hughes and P. Foulkes, "The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age", *Speech Comm.* 66, pp. 218-230, 2015.

[10] N. Brümmer and A. Swart, "Bayesian Calibration for Forensic Evidence Reporting", *Interspeech* 2014, pp.388-392.

[11] G.S. Morrison "Forensic Voice Comparison using likelihood ratios based on polynomial curves fitted to the formant trajectory of Australian English /aɪ/", *Int'l J. Speech Language and the Law* 15/2, pp.249-266, 2008.

[12] P. Rose, "Forensic voice comparison with secular shibboleths - a hybrid fused GMM-multivariate likelihood ratio-based approach using alveo-palatal fricative cepstral spectra", *ICASSP* 2011, pp. 5900-5903.

[13] P. Rose, "More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends", *Int'l J. Speech Language and the Law* 20/1, pp. 77-116, 2013.

[14] N. Balamurali, E. Alzqhoul, and B. Guillemin, "Determination of likelihood ratios for forensic voice comparison", *Int'l J. Speech Language and the Law* 21/1, pp. 83-112, 2014.

[15] G.S. Morrison, "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM)", *Speech Communication* 53, pp. 242-256, 2011.

[16] J Franco-Pedroso, F. Espinoza-Cuadros and J. Gonzalez-Rodriguez, "Formant Trajectories in Linguistic Units for Text-Independent Speaker Recognition", *Intl. Conf. Biometrics* 2013.

[17] D. Ramos and J. Gonzalez-Rodriguez, "Reliable Support: Measuring Calibration of Likelihood Ratios", *Forensic Science International* 230/1-3, pp. 156-169, 2013.

[18] I.W. Evett, J. Scrange, and R. Pinchin, "An illustration of the advantages of efficient statistical methods for RFLP analysis in forensic science", *Am. J. Human Genetics* 52, pp. 498-505, 1993.

[19] C. Neumann, I.W. Evett and J. Skerrett, "Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm", *J. Royal Statistical Society* 175, pp. 371-415, 2012.

[20] G.S. Morrison, "Tutorial on logistic regression calibration and fusion: converting a score to a likelihood ratio", *Australian J. Forensic Sci.*, pp. 1-25, 2012.

[21] A.B. Hepler, C.P. Saunders, L.J. Davis, and J. Buscaglia, "Score-based likelihood ratios for handwriting evidence", *Forensic Science International* 219/1-3, pp. 129-140, 2012

[22] S. Ishihara, "A Likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams", *Int'l J. Speech Language and the Law* 21/1, pp. 23-49, 2014.

[23] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia Gomar and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition", *Computer Speech and Language Special IEEE Odyssey 2004 Issue* 20/2-3, pp. 331-355, 2006.

[24] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection", *Computer Speech and Language IEEE Odyssey 2004 Issue* 20/2-3, pp. 230-275, 2006.

[25] G.S. Morrison, "Likelihood Ratio forensic voice comparison using parametric representation of the formant trajectories of diphthongs", *JASA* 125, pp. 2387-2397, 2009.

[26] G.S. Morrison, "Vowel inherent spectral change in forensic voice comparison", in Morrison and Assmann (eds.), *Vowel Inherent Spectral Change*, Springer, Heidelberg, pp. 263-283, 2013.

[27] P. Rose, "Forensic Voice Comparison with Monophthongal Formant trajectories – a Likelihood Ratio-based discrimination of "schwa" vowel acoustics in a close social group of young Australian Females", *ICASSP* 2015.

[28] J. Li and P. Rose, "Likelihood Ratio-based Forensic Voice Comparison with F-pattern and Tonal F0 from the Cantonese /ɔy/ Diphthong", *Australian Int'l Conf. on Speech Science & Technology* 2012, pp. 201–204.

[29] C. Wang and P. Rose, "Likelihood Ratio-based Forensic Voice Comparison with the Cantonese /i/ F-pattern and Tonal F0", *Australian Int'l Conf. on Speech Science & Technology* 2012, pp. 209–212.

[30] P. Rose, "Bernard's 18 – Vowel Inventory Size and Strength of Forensic Voice Comparison Evidence. *Australasian Int'l Conf. on Speech Science and Technology*, 2010, pp. 30–33.

[31] P. Rose, "Hong Kong Cantonese Citation Tone Acoustics: A Linguistic-Tonetic Study', *Australian Int'l Conf. on Speech Science and Technology* 2000, 198–203.

[32] K.N. Stevens, *Acoustic Phonetics,* MIT Press, Cambridge Mass., 1998.

[33] http://philjohnrose.net/HKUST_FVC/HUMA_6000_PG_FVC_syll.pdf

[34] X. Wang, *Likelihood ratio-based Forensic Voice Comparison in Cantonese – An examination in disyllable word /daihyat/ 'first'*, unpublished MSc. Sub-thesis, University of York, 2014.

[35] E. Enzinger, "The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems", *Australasian Int'l Conf. on Speech Science & Technology* 2012, pp. 137-140.

[36] G.S. Morrison, P. Rose and C. Zhang "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice." *Australian J. Forensic Sci.* 44/2, pp.155–167, 2012.

[37] C.G.G. Aitken and D. Lucy. "Evaluation of trace evidence in the form of multivariate data", *Applied Statistics* 53/4, pp. 109-122, 2004.

[38] N. Brümmer, "Focal Toolkit", http://www.dsp.sun.ac.za/nbrummer/focal

[39] R. Zheng and P. Rose, "Likelihood Ratio-based Forensic Voice Comparison with Cantonese Short-term Fundamental Frequency Distribution Parameters", *Australasian Int'l Conf. on Speech Science & Technology* 2012, pp. 153-156.

[40] E. Shriberg and A. Stolcke, "The Case for Automatic Higher-Level Features in Forensic Speaker Recognition." *Interspeech* 2008, pp. 1509-1512.

[41] D. Reynolds, W. Andrews, J. Campbell, J. Navrátil, B. Peskin, A. Adami, Q. Jin, D. Klusáček, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones and B. Xiang, "The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition." *ICASSP* 2003, pp 784-787.

[42] J. González-Rodríguez, "Speaker recognition using temporal contours in linguistic units: the case of formant and formant-bandwidth trajectories." *Interspeech* 2011, pp. 133-136.

[43] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen and T. Niemi, "Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition", European Network of Forensic Science Institutes, 2015.

[44] E. Enzinger, G.S. Morrison and F. Ochoa, "A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case", *Science and Justice* 56, pp. 42-57, 2016.

[45] E. Enzinger and G.S. Morrison, "Mismatched distances from speakers to telephone in a forensic-voice-comparison case", *Speech Comm*. 70, pp. 28-41, (2015).

[46] T.E. King, G.G.Fortes, P. Balaresque, M.G. Thomas, D. Balding, P.M. Delser, R. Neumann, W. Parson, M. Knapp, S. Walsh, L. Tonasso, J. Holt, M. Kayser, J. Appleby, P. Forster, D. Ekserdjian, M. Hofreiter and Kevin Schürer, "Identification of the Remains of King Richard III", *Nature Communications 5,* 2014.