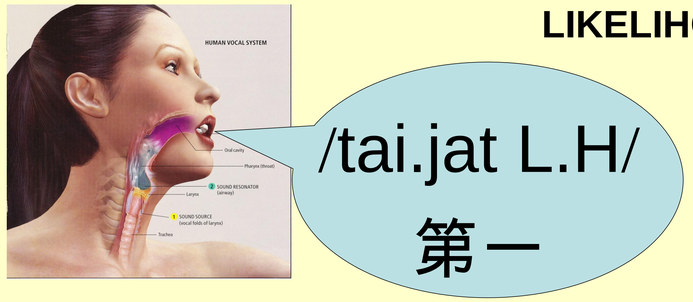
CANTONESE FORENSIC VOICE COMPARISON with HIGHER-LEVEL FEATURES:

LIKELIHOOD RATIO-BASED VALIDATION USING F-PATTERN AND TONAL F0 TRAJECTORIES OVER A DISYLLABIC HEXAPHONE

Phil Rose[1] & Wang Xiao

[1]*Australian Academy of Forensic Sciences & Australian National University Emeritus Faculty*



1. Introduction

- **Duty of Expert to Court**
  to **help** it by giving objective and unbiased opinion within their area of expertise (UK *Criminal Procedure Rules*)
- **Forensic Speaker Recognition**
  expert typically compares suspect and offender speech samples to **help** the trier-of-fact decide whether the suspect said the incriminating speech
- **One way: LR-based Forensic Voice Comparison**
  - furnishes interested parties with a ***Likelihood Ratio***
  - ratio of conditional probabilities of speech evidence E*sp* under competing prosecution and defense hypotheses H*p,* H*d :*

$$P(E_{sp} \mid H_p) \,/\, P(E_{sp} \mid H_d)$$

  - quantifies the strength of the speech evidence E*sp*
  - logically correct by Bayes' Theorem (posterior odds = prior odds * likelihood ratio): $P(H|E_{sp})$ *not possible* absent prior odds $P(H_p)/P(H_d)$ to which expert not privy
  - legally correct by avoiding *ultimate issue* considerations

2. Likelihood ratio

validation/discrimination function demo

- Figure 1:
- cumulative distribution of likelihood ratios from 33 target trials (same-speaker comparisons) and 528 non-target trials (different-speaker comparisons)
- non-contemporaneous phone recordings of male Australians answering *yes* and *not too bad*.
- LRs from different-speaker comparisons increase towards the left; same-speaker LRs towards the right
- features being validated: F-pattern in *yes*; intonational F0 (H.L.LH) in *not too bad*.
-curves show individual feature and log-reg fused performance.

This *Tippett Plot* conveys visually that LRs based on a fusion of F-pattern in *yes* and intonational F0 in *not too bad* can discriminate reasonably well between non-contemporaneous same-speaker (ER = 2%) and different speaker (ER = 5%) telephone speech samples.

3. LR-based system performance

(E)ERs etc.?

-use of error rates with LRs is incorrect, tho' intuitively appealing: by Bayes' Theorem a prior probability still required to decide whether the suspect said the incriminating speech.
- assuming flat priors for convenience, EERs useful as indicators of discriminative power.
- performance of LR-based detection systems (system *validity*) properly assessed by the simple information-theoretic hypothesis-dependent logarithmic cost function **C*llr:***

$$C_{llr} = \frac{1}{2}\left(\left[\frac{1}{N_{Hp}}\sum_{i}\log_2\left(1+\frac{1}{LR_i}\right)\right]+\left[\frac{1}{N_{Hd}}\sum_{j}\log_2\left(1+LR_j\right)\right]\right)$$

C*llr* < 1 = system is delivering information;
C*llr* << 1 = system delivers good strength of evidence

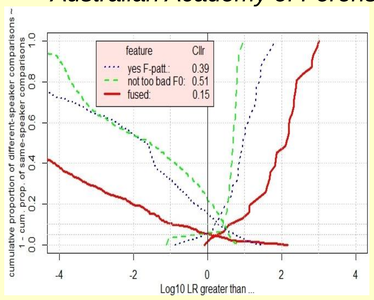


Figure 1: Tippett plot for validation LRs derived from F-pattern in *yes* and intonational F0 in *not too bad* in a $150 million telephone fraud

4. Research Questions

Previous research > **trajectories of acoustic phonetic features** give better strength of evidence parametrised by either DCT or polynomial coefficients, rather than traditional acoustic-phonetic point measurements.

- 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.
- LR-based comparisons with vowel acoustics usually performed on the F-pattern of monosyllables. But offender and suspect speech samples often contain the same polysyllabic words and expressions. Therefore:
- **Q1: do you get better evidence strength if you model formant trajectories *globally* (e.g over disyllabic word)?**
- **use Cantonese daihyat *first* F-pattern to find out (figure 2).**
- **Q2: daihyat has L.H tones, so**
- **what sort of evidence strength from tonal F0 trajectory over a disyllabic word?**

Figure 2 F-pattern and F0 in clearly spoken token of *daihyat*

5. Elicitation, Corpus, Speakers

- **map-task**, to elicit natural speech while still controlling the speech segments required for forensic voice comparison.
- **non-contemporaneous recordings** essential in FVC testing to preserve realistic within-speaker variation.
- **23 young Cantonese males** recorded on two occasions separated by about one month. Elicitation varied in the second recording to avoid learning effects.
- protocol for collection of forensic speech data : participants communicated by phone while high quality recordings were made from lapel microphones.
- ***Hong Kong Mass Transit Railway*** (HKMTR) database, e.g.

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.

- **replicate number** = ca. 8 per recording session
- = ca. **1.5 secs.** net speech per suspect/offender sample

6. Processing & Parameterization

- F1, F2, F3, F0 identified from wideband spectrograms, trajectories modelled with polynomials from one to cubic as *f*(equalised duration). Coefficients used as features.

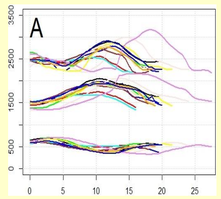
-each speaker's 1st recording mean compared with their 2nd gives 23 known same-speaker scores; compared with other speakers' 1st recording gives 253 different-speaker scores. (Refer figure 3)

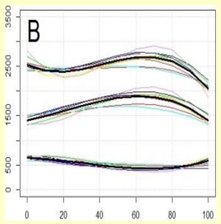
- leave-one-out cross-validation necessary

- scores (measure of similarity taking typicality into account) estimated with multivariate kernel-density likelihood ratio ex *Joseph Bell Centre for Forensic Statistics and Legal Reasoning*.

- scores calibrated with logistic regression (*Focal* toolkit) to get LRs.

-Log-reg fusion of formants and F0. C*llr* evaluation.





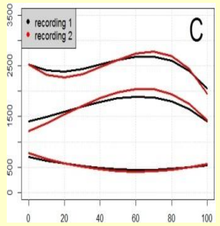


Figure 3 *Stages in* daihyat *F-pattern acoustic data extraction and comparison.*

7.Results

- refer Tippett Plots in figure 4.
- No improvement from complex e.g. quintic trajectories; and
- Different formants need different orders:
- Optimum C*llr* < quadratic (F1, F3), cubic (F2)
- Improvement from fusion of F0, F-pattern: disyllabic tonal F0 *does* contribute.
- Spline analysis confirms advantageous to model the trajectories over the whole word rather than separately on the constituent syllables.
- Fusion with two other higher-level features from HKMTR database (yih F-pattern, short term F0) improves C*llr* still more.

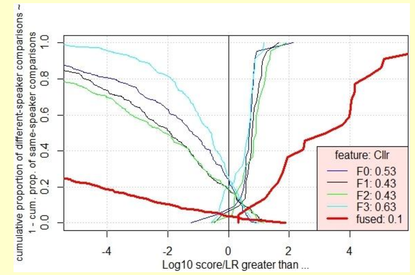


Figure 4 *Validation Tippetts for individual & fused* daihyat *components*

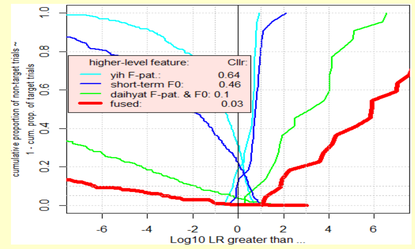


Figure 5 *Validation Tippetts for* daihyat, yih *and short term F0*

8. Sidebar: LRs in forensic reality

Considerable support, outside the Law, for LRs. Problems:

- difficult to understand for posterior-obsessed legals and lays.
- combination with non-quantifiable evidence.
- different systems (GMM, MVLR) produce different LRs for same data.

BUT expert can often **help** the court without LRs (figures 6 & 7)

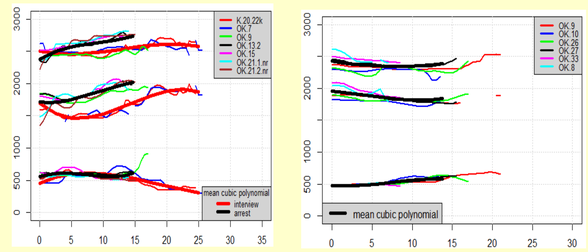


Figure 6 *F-pattern for /ei/ in OK in suspect (left) and questioned voice samples. (West-African accented Australian).*

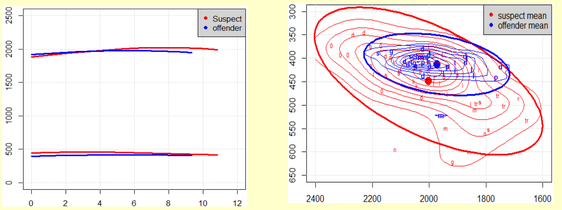


Figure 7 left: *F1 & F2 in /e:/ in suspect and questioned voice samples (Singapore English. Right: densities of corresponding F1 F2 targets.*