Forensic voice comparison with word-based acoustics

a likelihood ratio-based discrimination using F-pattern and tonal F0 trajectories over a disyllabic Cantonese word

Bruce Xiao Wang, Phil Rose [1]

[1] *Australian National University Emeritus Faculty*

1. **Introduction**

- Forensic voice comparison
  -- comparing suspect and offender speech samples to assist the trier-of-fact decide whether the suspect said the incriminating speech
  -- furnishing the interested parties with Likelihood ratio.

- Likelihood ratio

--Estimating how much more likely one is to get the speech evidence (E_{sp}) – the observed differences between the known suspect and unknown offender speech samples – assuming the incriminating speech has come from the suspect (the prosecution hypothesis H_p) rather than someone else in the relevant population (the defence hypothesis H_d).

-- Ratio of conditional probabilities of speech evidence E_{sp} under competing hypotheses

→ $p(E_{sp} \mid H_p) / p(E_{sp} \mid H_d)$

- Disyllabic word

-- *Acoustic parameters were extracted from a disyllabic Cantonese word as a whole rather than over its individual monosyllables as conventionally practiced.*

2.Research questions

1. What strength of evidence would it yield if we treat this disyllabic word as a whole polyphthong, rather than as a sequence?

2. How good or bad will the tonal F0 work over a sequence?

3. **Procedure**

3.1 Test Word

Phonemic representation	Phonetic realisation	Tone	Meaning	Character
/daihyat/	[taihjat˧ 22.5]	low-levelled pitch + high-levelled pitch	First	第一

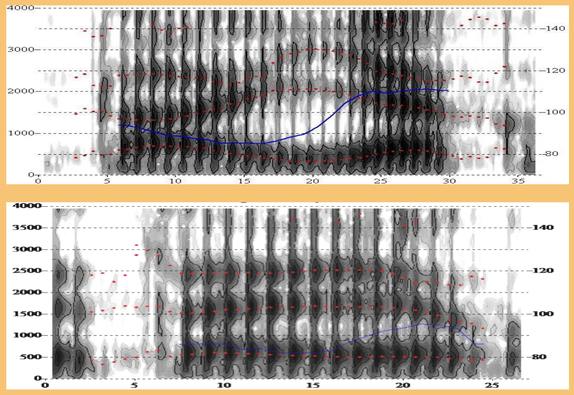


Figure.1 Wideband spectrograms of two daihyat *tokens from the same speaker. Red dots = superimposed formant centre-frequencies, blue line = F0. X-axis = duration (csec.), left & right vertical axes = spectrographic frequency (Hz) & F0 (Hz).*

3.2 Subjects & number of token

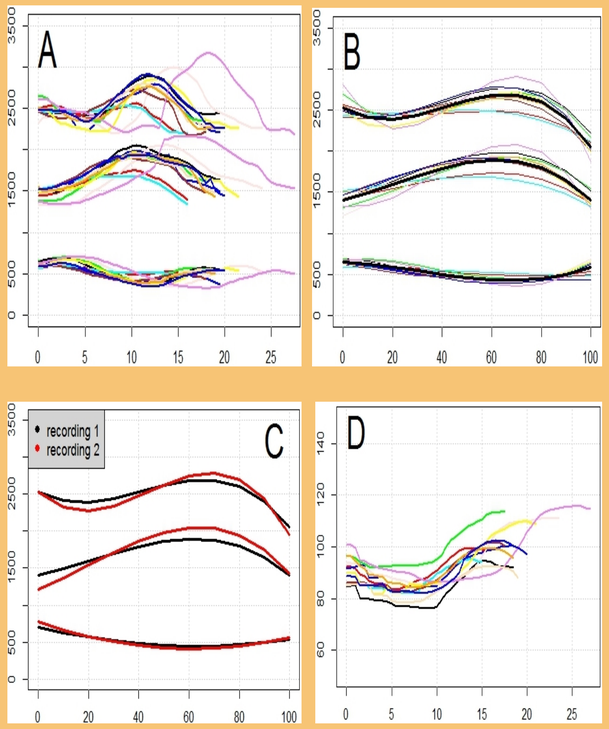
-- 23 young Cantonese speakers
-- Non-contemporaneous data
-- 6-10 tokens per subject per recording

3.3 Processing

-- speakers' *daihyat* tokens were identified aurally
--wideband spectrograms of them generated in *Praat* with superimposed formant and F0 traces
-- the first three formants and F0 were extracted
-- onset was taken to be at the first strong glottal pulse of /ai/ in *daih*.
-- offset was adjudged at the last strong glottal pulse of /ja/ in *yat.*

3.4 Parametrization

-- raw F-pattern trajectories were modeled by permuting polynomials of all degree from one to cubic separately on each formant
-- coefficients extracted for LR processing



A: 12 raw /daihyat/tokens from a single recording
B: Cubic fit of A, Equalized duration. Black thick line -> Mean.
C: Comparison of the mean of cubic fit for F1,F2 and F3 from one speaker's two non-contemporaneous recordings.
D. Raw tonal F0 for the same speaker's first recording.

4. Results

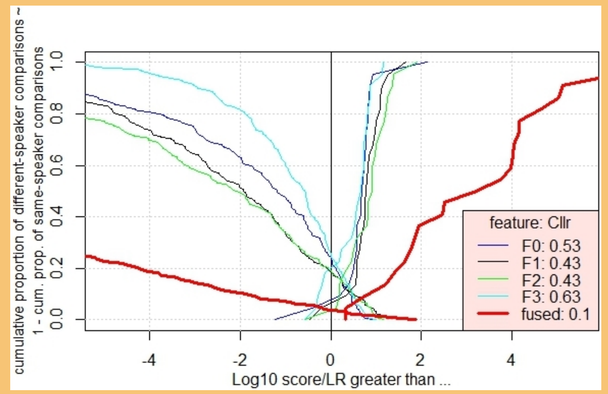


*Figure.2 Tippett plots of results (*daihyat *word F-pattern and tonal F0).*

-- optimum C_{llr} was obtatained with quadratic for F1 and F3 and cubic for F2.
-- C_{llr} of F123 --> 0.16
-- Quadratic and cubic modeling of F0 gives the same results
-- fusion of F-pattern and F0 improves C_{llr}

5. Summary

-- estimating LRs from the formant and F0 trajectories over a disyllabic word can yield good strength of evidence

-- different formants may warrant different orders of polynomial for an optimum performance

-- higher order polynomial fitting did not achieve the best results