Two sides of the same coin: between-speaker F0 differences in linguistic-tonetic description and forensic voice comparison.

Division of Humanities, Hong Kong University of Science & Technology
School of Language Studies, Australian National University
Joseph Bell Centre for Forensic Statistics and Legal Reasoning, University of Edinburgh

phil rose
• Between-Speaker Differences in tonally-relevant acoustic output from two complementary perspectives:

• (1) BSD’s in Forensic Voice Comparison
• A case of “prosody in the real world”:
• A real world FVC case where intonational F0 played an important part

• (2) BSD’s in Linguistic Tonetics
• Tonal Normalisation and some of its uses for a quantifiable linguistic-tonetic representation of tonal and intonational pitch.
Main anatomical source of F0 BSDs

Cords vibrating like a string Titze 1994

- \( F_0 = \frac{1}{2L} \sqrt{\frac{\sigma}{\rho}} \)
- \( L \) = vocal cord \textit{length}
  - \( \sigma \) = longitudinal \textit{stress} in the cords
  - stress = the \textit{tension} in the cords divided by the cross-sectional area of vibrating tissue
  - (cover) tension is controlled by Cricothyroid contraction/relaxation
  - \( \rho \) = tissue density

Since F0 is inversely proportional to cord length, \textit{other things being equal}, if the speaker's cords are long, their F0 will be lower

Cords vibrating like a spring

- \( F_0 = \frac{1}{2\pi} \sqrt{\frac{k}{m}} \)
- \( m \) = vocal cord \textit{mass}

Since F0 is proportional to cord mass, \textit{other things being equal}, if the speaker's cords are bigger, their F0 will be lower
Forensic Voice Comparison

• Self-evidently the differences between speakers that are important
• Absent BSD’s not possible to recognise someone by their voice
• FVC = comparing speech samples wrt any aspect of voice (not just phonetics!) to help trier-of-fact decide whether suspect said incriminating speech
On Christmas Eve 2003 a fraudulent fax was sent to the investment bank *JP Morgan Chase* in Australia requesting the transfer of $150 million to accounts in Switzerland, Greece and Hong Kong.

About 10 minutes before the closing of business, the bank received a phone call from a Craig Slater, asking for a *call-back* on the fax — a procedure confirming the details of the fax and verifying that the transfer could go ahead.

Here is part of the money-making phone-call:

**The Crime**
“JP Morgan Greg speaking”

“Yeah hello Greg this is Craig Slater here mate”

“Oh g’day how are you?”

“Not too bad I bin havin a bit of trouble here…”
Out goes the money …

“em.. And we’re going to pay Hong Kong dollars 118,678,543 spot 29 to HSBC em…Hong Kong?”

“Correct”

Hong Kong I think Hong Kong Power Limited six three six double oh three oh five five double oh one [$636,003,055,001] ?

“Yes”
• That is how you make $150 million in one phone call
• And also how the *Australian Commonwealth Superannuation Scheme* account administered by the bank lost $150 million.
The Suspect

• 15 intercepted telephone calls containing “not too bad”, e.g.
• “…mate, how are you?”
• “Oh not too bad, everything’s good.”
The (Intonational F0) Evidence

• Both suspect and offender contain the utterance “not too bad” said with same H.L.LH intonation
  – rise nuclear tone on bad (“supportive interest encouraging further conversation”).
  – high head on not (the suspect’s not high/low head)

• Therefore F0 highly comparable

• Usually F0 not much good in FVC

• < high within-speaker variation

• > disadvantageous variance ratio.
罪犯的 “not too bad” F0

H on not
L on too
LH on bad
F0 曲线的相似程度 Degree of similarity between suspect and offender’s not too bad F0
You want to know the probability the suspect said the incriminating speech, given the similarity between the suspect and offender data? $p(H|E)$

By my theorem, that is proportional to the strength of your evidence …

... and the probability that the suspect said the incriminating speech BEFORE the evidence is taken into account …

Bayes’ Theorem:
Posterior Odds = Prior Odds * Likelihood Ratio
The Likelihood Ratio

- Strength of Evidence in support of one hypothesis over another:
- Probability of evidence under competing hypotheses =
- \( p(E \mid H_{same\ spk}) / p(E \mid H_{diff\ spk}) \)
- Probability of the difference between suspect and offender F0 in *not too bad* assuming the suspect said it, vs. the probability of the difference, assuming it was said by someone else randomly chosen from the relevant population.

LR denominator is where the between-speaker differences come in!
So we have to collect a Reference Sample of “not too bad”s

- Natural responses to “how’s it going?” etc
- Do any two samples sound as if they are from the same speaker?
- Relatively easy to find speakers with very similar voices!!
Reference sample: non-contemporaneous variation in 30 males’ “not too bad” F0.

You have to go and get this!
The Formula

多变量似然率计算公式 Multivariate Likelihood Ratio (Aitken & Lucy 2002)

MVLR的分子 =

\[
(2\pi)^{-p} |D_1|^{-1/2} |D_2|^{-1/2} |C|^{-1/2} \left( m h^p \right)^{-1} |D_1^{-1} + D_2^{-1} + (h^2 C)^{-1}|^{-1/2} \\
\times \exp \left\{ -\frac{1}{2} \left( y_1 - \bar{y}_2 \right)^T (D_1 + D_2)^{-1} \left( y_1 - \bar{y}_2 \right) \right\} \\
\times \sum_{i=1}^{m} \exp \left\{ -\frac{1}{2} \left( y^* - \bar{x}_i \right)^T \left( (D_1^{-1} + D_2^{-1})^{-1} + (h^2 C) \right)^{-1} \left( y^* - \bar{x}_i \right) \right\}
\]

MVLR的分母 =

\[
(2\pi)^{-p} |C|^{-1} \left( m h^p \right)^2 \prod_{l=1}^{2} |D_l|^{-1/2} |D_l^{-1} + (h^2 C)^{-1}|^{-1/2} \times \sum_{i=1}^{m} \exp \left\{ -\frac{1}{2} \left( y_l - \bar{x}_i \right)^T \left( D_l + h^2 C \right)^{-1} \left( y_l - \bar{x}_i \right) \right\}
\]
Multivariate LR values for comparison between suspect and offender samples using F0 in “*not too bad*” against reference population of 30 males. About 20 times more likely to get this difference in *not too bad* F0 if suspect said it than if someone else had said it.

NOT the suspect is about 20 more likely to have said it than someone else!!
By combining LRs from different features, one can get quite large strengths of evidence in support of either defence or prosecution hypotheses.

In this case the acoustics (F-pattern) in “yes” were also used. Offender Suspect Reference sample.

They gave a LR of about 70.

Combined with not too bad F0 the LR is now 1400.

All the acoustic voice evidence in the case gave a LR of about 11 million.

The Other Voice Evidence
• I don’t know the prior odds (= the other evidence in the case), but
• The suspect was found guilty
• Most of the money was recovered
Forensic Voice Comparison with Tonal F0?

Yes, small contribution from tones – improves /i/ Cllr on fusion.

Log-LR cost (0.51)

LRs from same-speaker comparisons

LRs from different-speaker comparisons

EER = 15%

Tippett/reliability plots for F-pattern and tonal F0 in Cantonese yih ‘two’ for 26 young male Cantonese speakers’ non-contemporaneous natural speech.
Theme 2:
Using Between-speaker differences to get quantified Linguistic-Tonetic description of tones of a variety
– For tonal typology
– Acoustic reconstruction
Modelling tones

- Wu dialect tones
- Merit in complexity
- Some typically complex data from Wencheng, Jinyun
- Can all be easily modelled with a continuous model (e.g. Fujisaki)
- But perhaps not quite so easily with discrete phonological Bao-type model
Jinyun tones (Steed & Rose 2009)

<table>
<thead>
<tr>
<th>Contour</th>
<th>Register</th>
</tr>
</thead>
<tbody>
<tr>
<td>fall</td>
<td>Whole range</td>
</tr>
<tr>
<td>depressed fall</td>
<td>low</td>
</tr>
<tr>
<td>rise-fall</td>
<td>low</td>
</tr>
<tr>
<td>dipping</td>
<td>mid</td>
</tr>
<tr>
<td>rise</td>
<td>low</td>
</tr>
<tr>
<td>fall-rise</td>
<td>low</td>
</tr>
<tr>
<td>fall</td>
<td>low</td>
</tr>
<tr>
<td>rise-fall</td>
<td>high</td>
</tr>
</tbody>
</table>
Wencheng 文成
tonal acoustics (Rose 2010)

- Mean tonal F0 as function of mean tonal duration

<table>
<thead>
<tr>
<th>Contour</th>
<th>Register</th>
</tr>
</thead>
<tbody>
<tr>
<td>level</td>
<td>upper mid?</td>
</tr>
<tr>
<td>level</td>
<td>low</td>
</tr>
<tr>
<td>depressed level</td>
<td>mid</td>
</tr>
<tr>
<td>short rising</td>
<td>upper</td>
</tr>
<tr>
<td>rising</td>
<td>lower</td>
</tr>
<tr>
<td>falling-rising</td>
<td>mid?</td>
</tr>
<tr>
<td>falling-rising</td>
<td>low</td>
</tr>
</tbody>
</table>
Observations

• These systems mostly do not behave in the way theory tells us to expect
• Simple tones are rare; complex contours abound; phonation type contrasts are found in nearly all possible different interactions with tone
• The rules/constraints relating the isolation tones to tone sandhi forms lack phoneticity.
• Why would such systems evolve?
• Difficult to avoid idea of tones as indexical features
Normalisation

• Before you can answer these typological questions you need to be able to characterise varieties’ acoustics quantitatively

• Let’s look at a simple single variety - Shanghai
Shanghai raw tonal acoustics

Unstopped tones:
“high falling”
“mid dipping”
“low rising”

8 male (thick lines)
8 female

Controlled for intrinsic vowel F0
Controlled for intrinsic consonantal F0
Shanghai normalised tonal acoustics

(Rose 1993)

8 males
8 females

normalisation:
F0 - intrinsic z-score
duration – percent

NB not equalised!

Coloured lines = mean normalised F0, duration
Solid = male
Dotted = female

Note sex related differences in high falling tone

Normalisation index (Earle 1975):
How much does the normalisation reduce the original tonally-related between-speaker F0 variance?
With this normalisation, about 9.5 times
Comparing varieties

• If we want to find out how languages differ in their tonal acoustics, and how they are the same (Anderson’s 1973 “linguistic-phonetic properties”), we need to compare varieties.

• Problem: Comparing different varieties with normalisation is not straightforward: you need to be sure that your normalisation parameters are comparable across varieties! for example:

• How many linguistic-tonetically shared tones are there between Standard Thai (5) and Southern Thai (7)?
What is the correct relationship between these two sets?
Using bilingual’s tones

The female speaker is a Southern Thai educated professional, bilingual in both Southern & Standard Thai (Rose 1997).

So our normalisation strategy must adequately reflect the relationship between her two sets of tones …
testing z-score normalisation with bilingual’s tones

Speaker is controlling 11 different tones
(conservative HK) Cantonese

- six contrasting pitch shapes on unstopped syllables (subminimal sextuplet from CF4):

<table>
<thead>
<tr>
<th>Pitch Shape</th>
<th>Pinyin</th>
<th>Character</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>high level “[55]”</td>
<td>ku</td>
<td>姑</td>
<td>father's sister</td>
</tr>
<tr>
<td>mid level “[33]”</td>
<td>ku</td>
<td>故</td>
<td>cause</td>
</tr>
<tr>
<td>lower mid level “[22]”</td>
<td>pōu</td>
<td>部</td>
<td>part</td>
</tr>
<tr>
<td>falling from low “[21], [1 ↓]”</td>
<td>fu</td>
<td>扶</td>
<td>support</td>
</tr>
<tr>
<td>low to high rise “[24]”</td>
<td>ku</td>
<td>古</td>
<td>ancient</td>
</tr>
<tr>
<td>low to mid rise “[23]”</td>
<td>fu]</td>
<td>婦</td>
<td>woman</td>
</tr>
</tbody>
</table>
Z-score normalised Cantonese unstopped tones
(Rose 2000)

5 males, 5 females, controlled for intrinsic vowel F0

Normalised Tonal F0 and duration values for Cantonese isolation tones
Comparing Cantonese, Shanghai tones

8 different tones, “low rising” shared?
Is *anything* the same??

Comparing tones *across varieties*: high falling tones in Yongjiang 涌江 & Oujiang 甌江 sub-groups of Wu.

The amount of variance around these normalised curves is less than that for a single variety (Shanghai). They are demonstrably linguistic-tonetically the same tone.

Problem: these are two different Middle Chinese tonal cognates.
Summary

• Talk has focussed on quantified comparison of tonal/intonational F0 shapes..
• And testing of hypotheses about them!
• It has shown that BSD’s (from a lot of data!) are crucial for doing this.
THANK YOU FOR LISTENING