Acoustical Society of America Conference, Cancun, Mexico 17/11/'10: ***Invited Presentation at Special Session on Forensic Voice Comparison***

3aSC3

Special Session on Forensic Voice Comparison and Forensic Acoustics @ 2nd Pan-American/Iberian Meeting on Acoustics, Cancún, México, 15–19 November, 2010

http://cancun2010.forensic-voice-comparison.net

Combining linguistic and non-linguistic information in likelihood-ratio-based forensic voice comparison

phil rose AAFS

School of Language Studies,
Australian National University

Joseph Bell Centre for Forensic Statistics and Legal Reasoning,

University of Edinburgh

This presentation was researched as part of *Australian Research Council Discovery Grant* No. DP0774115.

Background

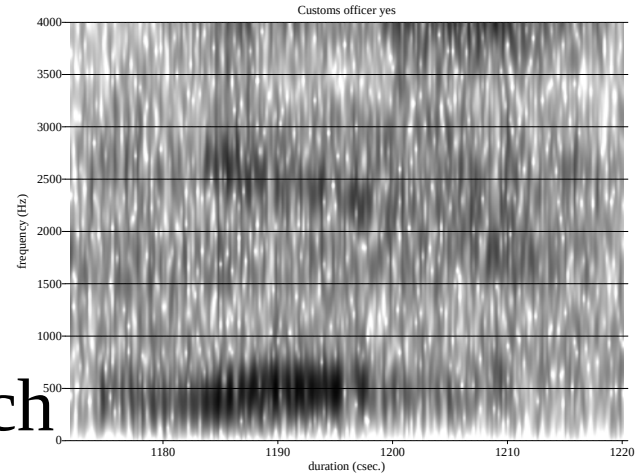
- Assumption: LR-based FVC framework:
  - Logically & Legally correct
  - Testable & Tested (cf *Daubert*)
  - Many other advantages (e.g combining evidence)
- Having your FVC cake and eating it:
  - ‘traditional’ & automatic LR-based approaches
  - both must be missing information,
  - so why not combine them?
- Neglected trad. FVC parameters:
  - Sonorant consonant F-pattern ([l ɹ n …])
  - Fricative consonant F-pattern ([s ʃ …])
  - Nasals, frics some non-deformable aspects in articulation

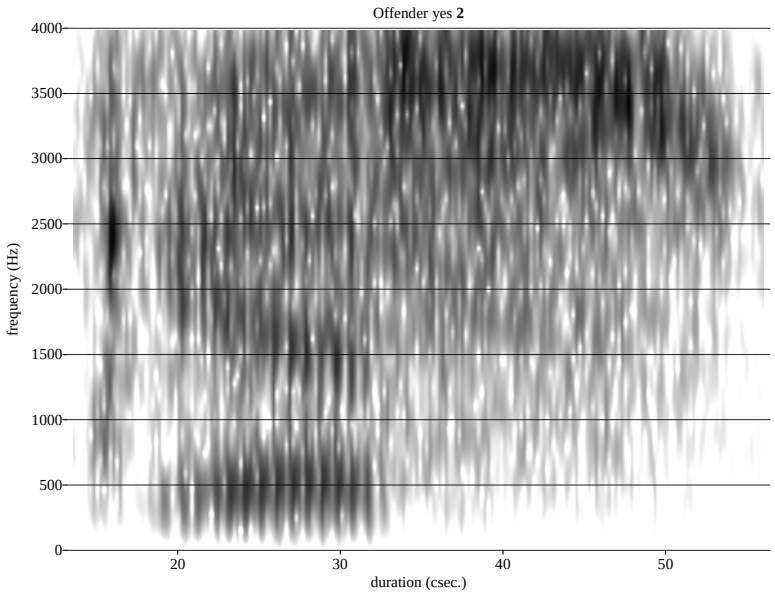
"… DNA profile evidence is now seen as setting a standard for rigorous quantification of evidential weight that forensic scientists using other evidence types should seek to emulate."
Balding: *Weight of Evidence for Forensic DNA Profiles* 2005.

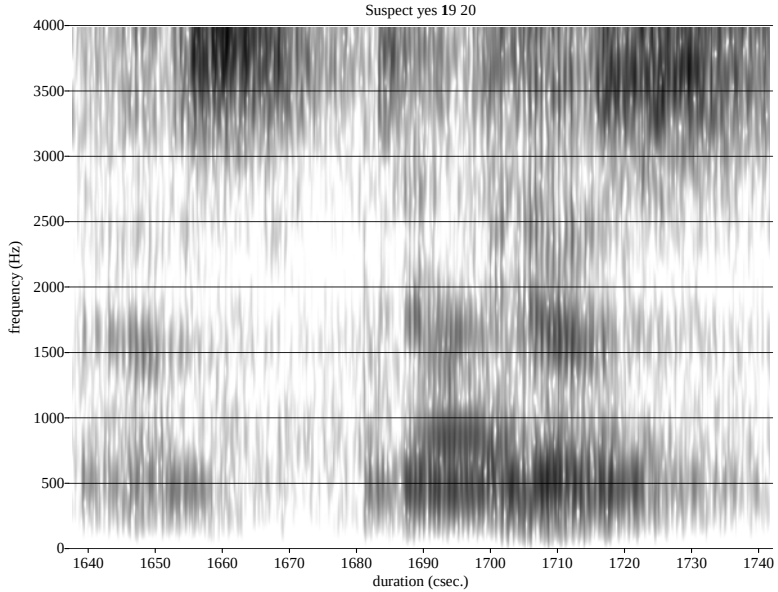
'Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition', Gonzalez-Rodriguez et al.: *IEEE TASpLP* 2007.

Fricative spectra in FVC

- R v Huffnagl et al. 2008
- $150 million telephone fraud case
- Small amount of offender speech
- Adequate amount of suspect speech
- But off. and sus. speech highly comparable in many linguistic features, incl. /s/ spectrum in *yes*.







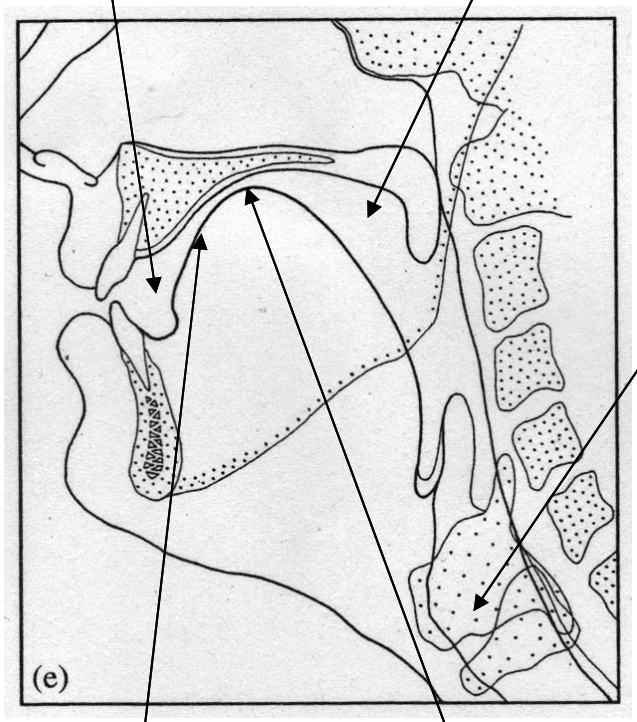
Aim(s)

- How well can same-speaker speech samples be discriminated from different-speaker speech samples using voiceless sibilant [ɕ] spectral features with LR as discriminant function?
- i.e. should we make use of these features in FVC?
- Can performance be enhanced by combining linguistic ([ɕ]) and non-linguistic LRs?

Integration of Traditional and automatic approaches

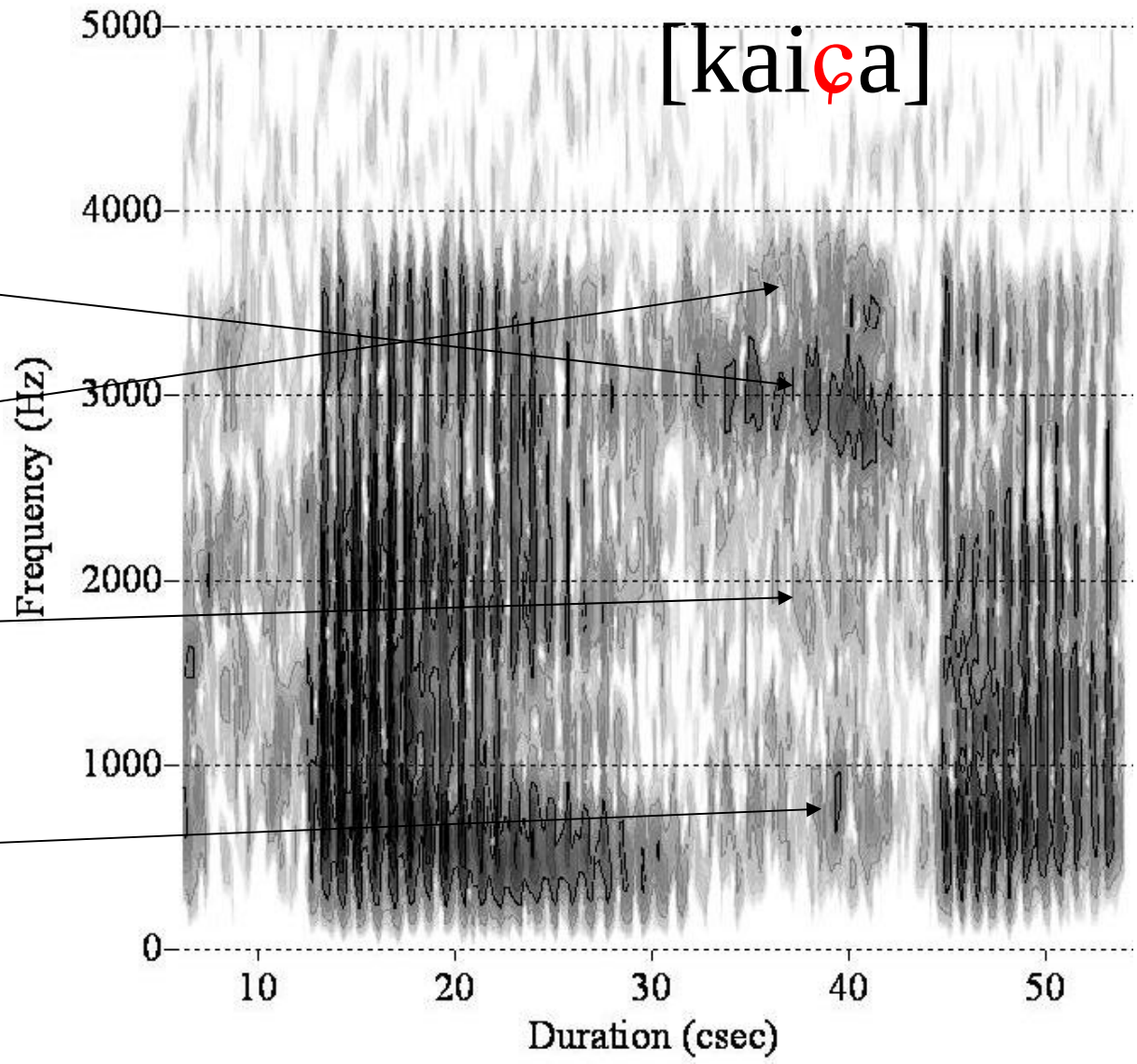
- Two senses:
  - Use automatic backend processing (fusion, GMM)
  - Use automatic features (e.g.MFCCs)
  - But locally
  - That’s what this talk is about
- Pull out and process comparable linguistic units
- Do the rest globally
- Combine results

Alveolopalatal fricative [ɕ]: articulation



Alveolopalatal fricative [ɕ]: acoustics

- Sources at incisors, constriction
- λ/2 resonance
  < front cavity
- λ/4 resonance
  < palatal channel
- λ/2 resonance
  < back cavity
- Helmholz resonance < SLVT
- subglottal resonances
- zeros



Data

- (Japanese) National Research Institute of Police Science (NRIPS) database (ca.2004)
- 300 male policemen; first **84 speakers** used
- **Ca. 70-80 secs. net speech per recording**, Sf = 10 k
- Set of vowels plus
- Single and polysyllabic word utterances
- **Non-contemporaneous landline recordings**
- **Separation ca. 3 – 4 months**
- Two repeats per recording
- **Channel not controlled, but likely similar**

"I've planted a bomb", "don't tell the police", "get the money ready"

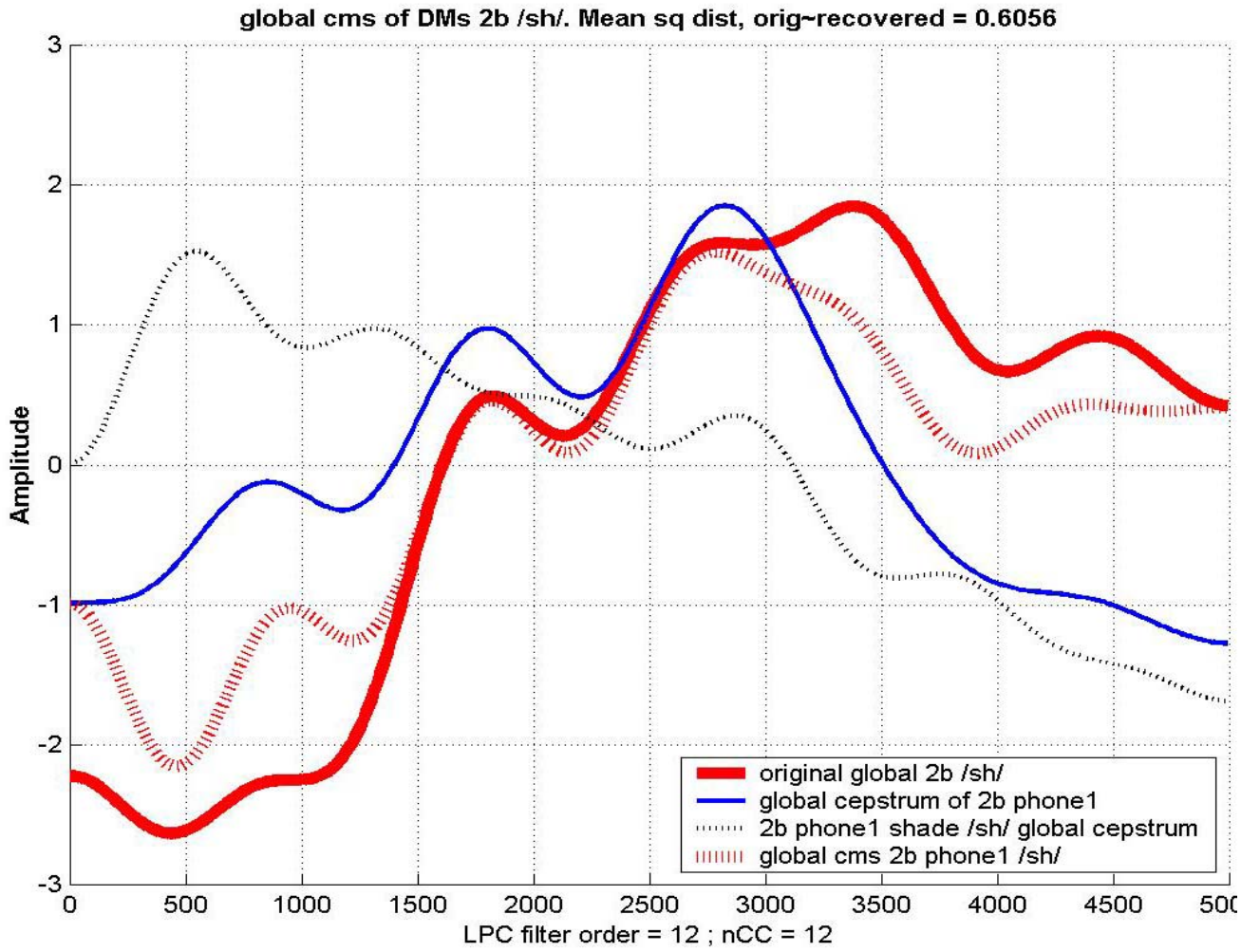
Data: [ɕ]

- 10 tokens of [ɕ] per repeat, various env'ts, e.g.
  - kaisha [kaiɕa] *firm*
  - ashita [aɕ:ta] *tomorrow*
  - shikaketa [ɕ:kaketa] *plant*
  - yooishiro [jo:iɕiɾo] *prepare*
- **20 tokens per recording**

Processing

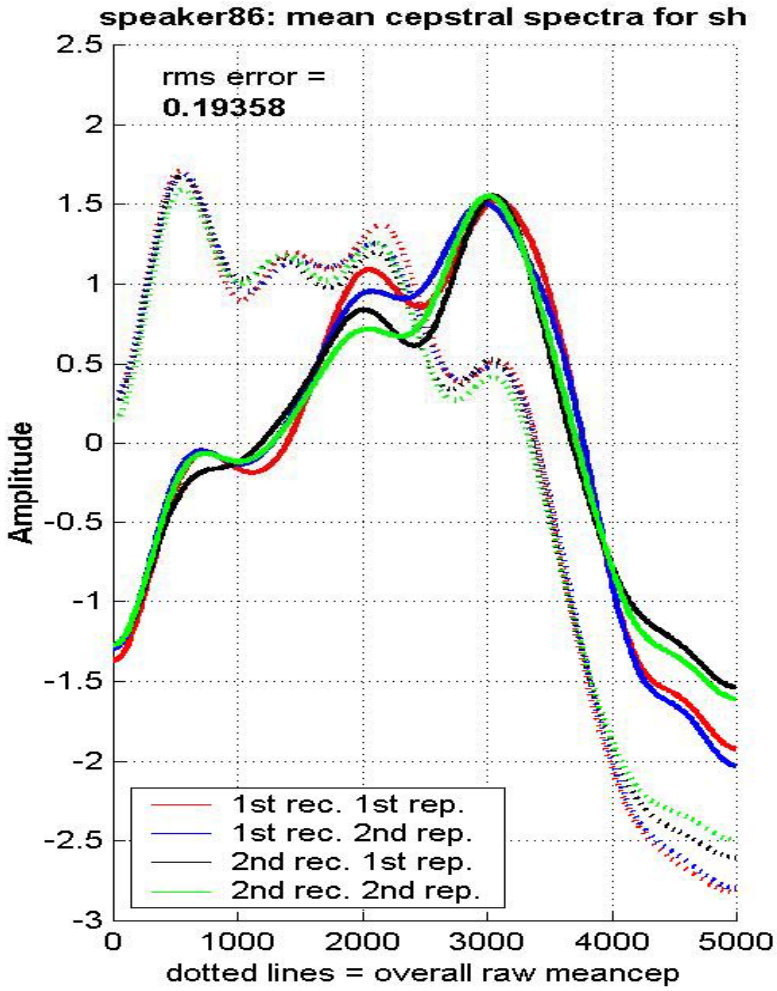
- Very basic front-end
- Non-linguistic:
  - LPC CCs 1 - 12
  - Mean cepstral vector
- Linguistic ([ɕ]):
  - Locate utterances with [ɕ], eyeball, *Praat* script to extract quasi steady-state (ca, 4 to 20+ csec.)
  - LPC CCs 1 – 12
  - Mean cepstral subtraction from non-linguistic mean vector

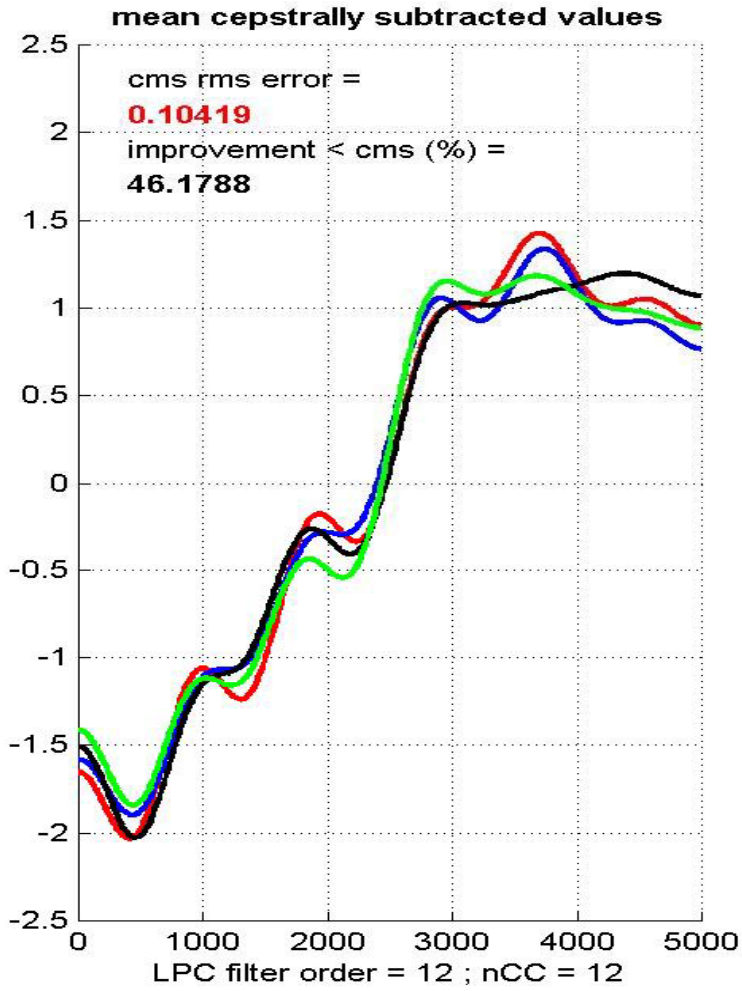
Cepstral mean subtraction



Cepstral spectra of [ʃ] in *shape*

Typical mean cepstral spectra (spk. 86)





Back-end-processing

- Two types of LR:
  - Multivariate LR
  - GMM/(U)BM LR
- All 84 speakers (i.e. intrinsic), cross-val.
- Log-reg fusion/calibration of LRs/scores from linguistic and non-linguistic data (Brümmer's *FoCal* toolkit)
- Evaluation with Cllr / EER
- Empirically discard CCs 4 6 8 9.

Generative LR developed at *Joseph Bell Centre for Forensic Statistics and Legal Reasoning* (Aitken & Lucy)

•Morrison's Matlab implementation of Reynolds Quaterieri & Dunn (2000) Adapted GMM Speaker Verification (Discriminative LR).

Cllr

Performance of LR-based detection systems is currently evaluated with the *Log Likelihood Ratio Cost* (Cllr):

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

- Simple scalar metric with 2 hypothesis-dependent log cost functions
- Idea is to severely penalise highly misleading LRs
- Cllr < unity considered “good”:
- > system is delivering some information

MVLR formula

numerator of MVLR =

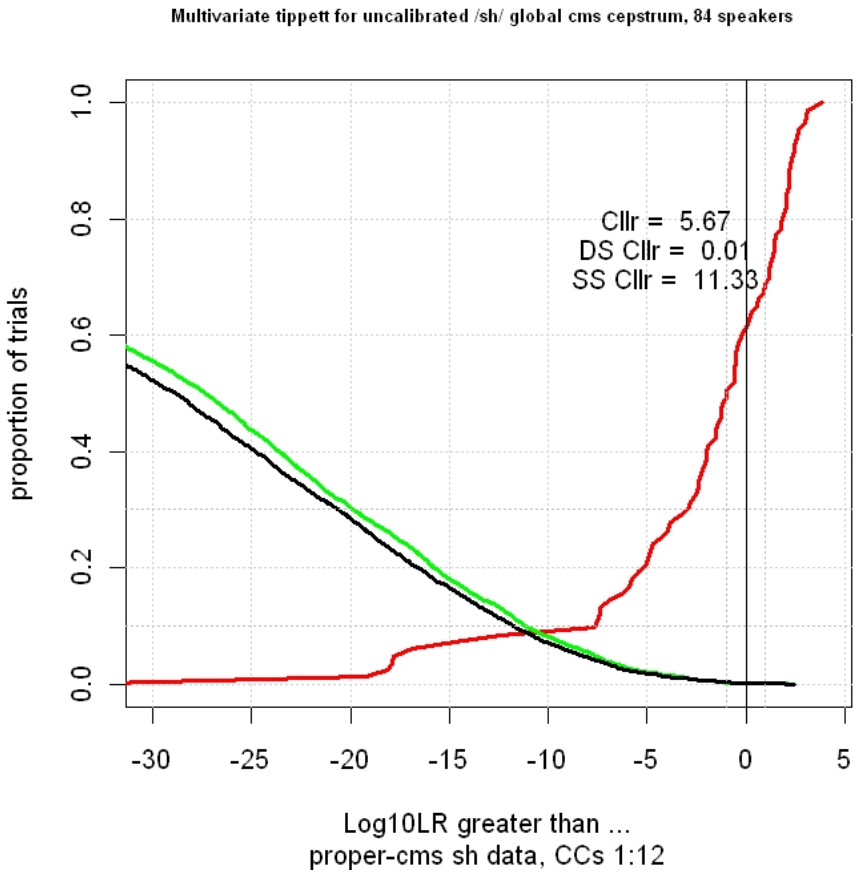
$$(2\pi)^{-p}\left|D_1\right|^{-1/2}\left|D_2\right|^{-1/2}\left|C\right|^{-1/2}\left(mh^p\right)^{-1}\left|D_1^{-1}+D_2^{-1}+\left(h^2C\right)^{-1}\right|^{-1/2}$$
$$\times\exp\left\{-\frac{1}{2}\left(\bar{y}_1-\bar{y}_2\right)^{\mathrm{T}}\left(D_1+D_2\right)^{-1}\left(\bar{y}_1-\bar{y}_2\right)\right\}$$
$$\times\sum_{i=1}^{m}\exp\left[-\frac{1}{2}\left(y*-\bar{x}_i\right)^{\mathrm{T}}\left\{\left(D_1^{-1}+D_2^{-1}\right)^{-1}+\left(h^2C\right)\right\}^{-1}\left(y*-\bar{x}_i\right)\right]$$

denominator of MVLR =

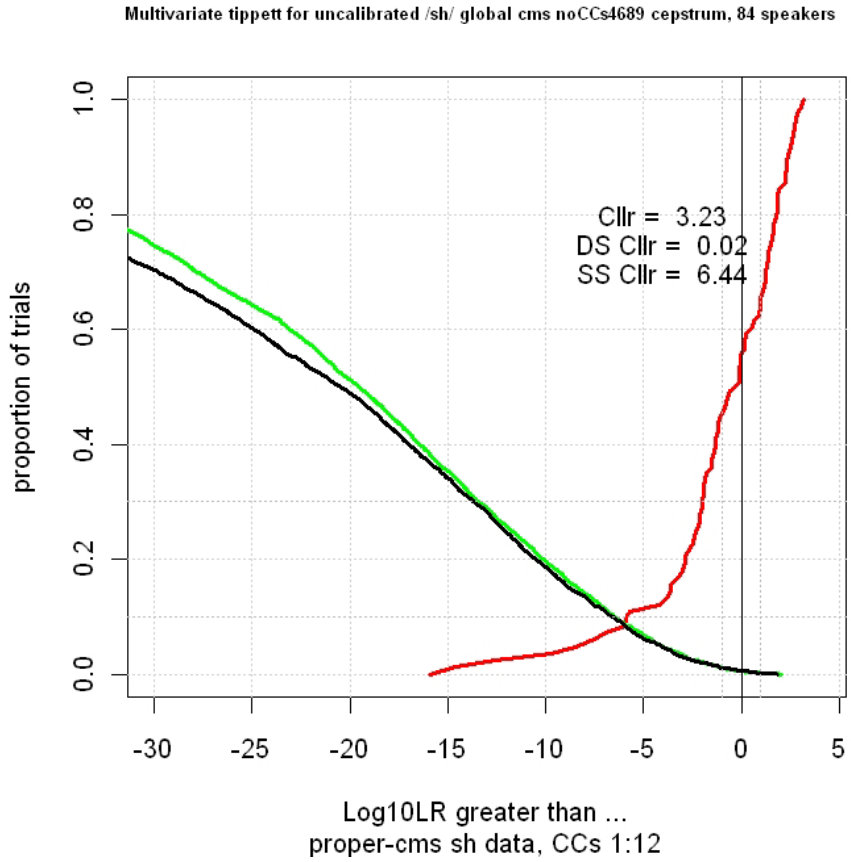
$$(2\pi)^{-p}\left|C\right|^{-1}\left(mh^p\right)^{-2}\prod_{l=1}^{2}\left[\left|D_l\right|^{-1/2}\left|D_l^{-1}+\left(h^2C\right)^{-1}\right|^{-1/2}\times\sum_{i=1}^{m}\exp\left\{-\frac{1}{2}\left(\bar{y}_l-\bar{x}_i\right)^{\mathrm{T}}\left(D_l+h^2C\right)^{-1}\left(\bar{y}_l-\bar{x}_i\right)\right\}\right]$$

MVLR Results …

Uncalibrated Tippetts (MVLR)

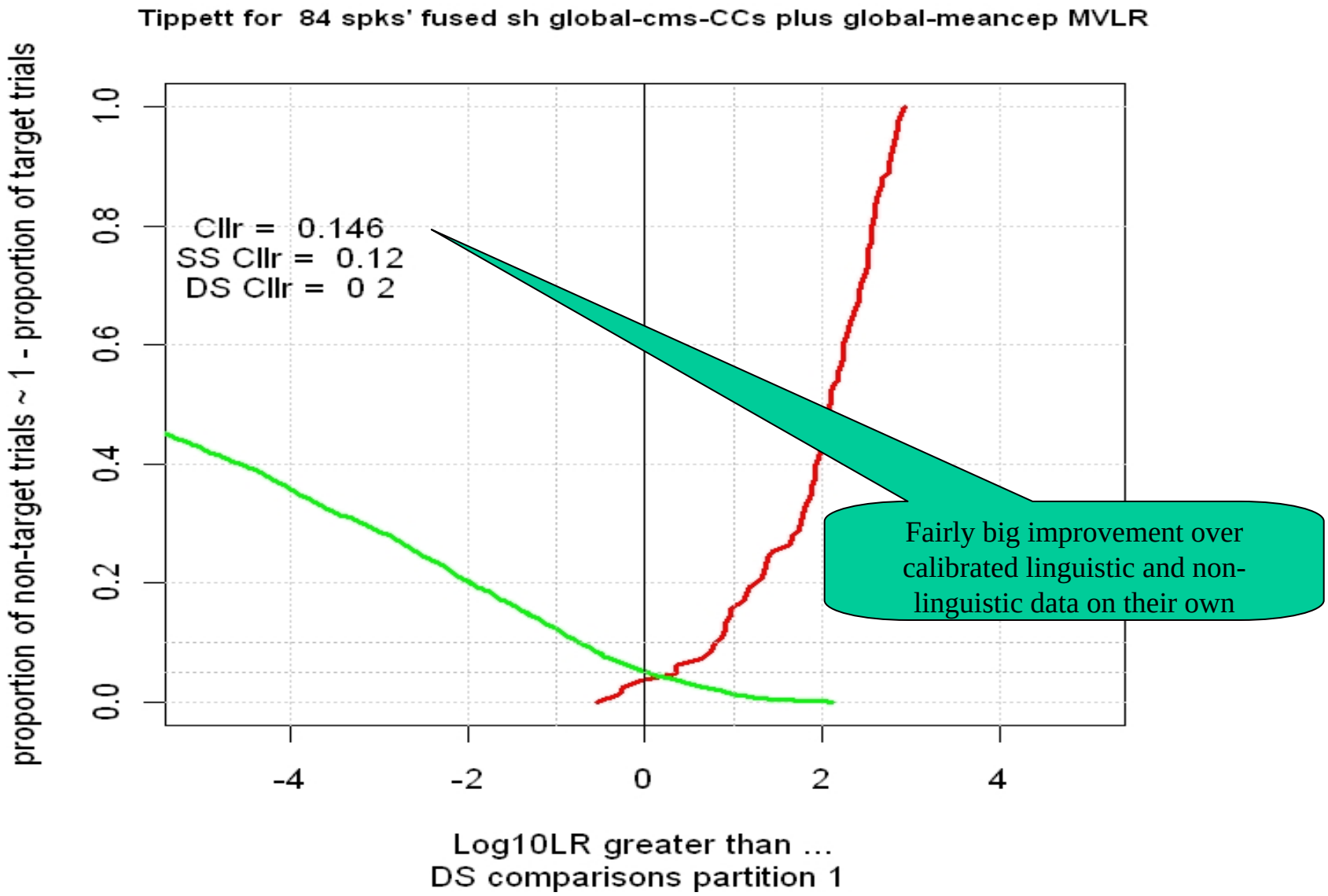


[ɕ]



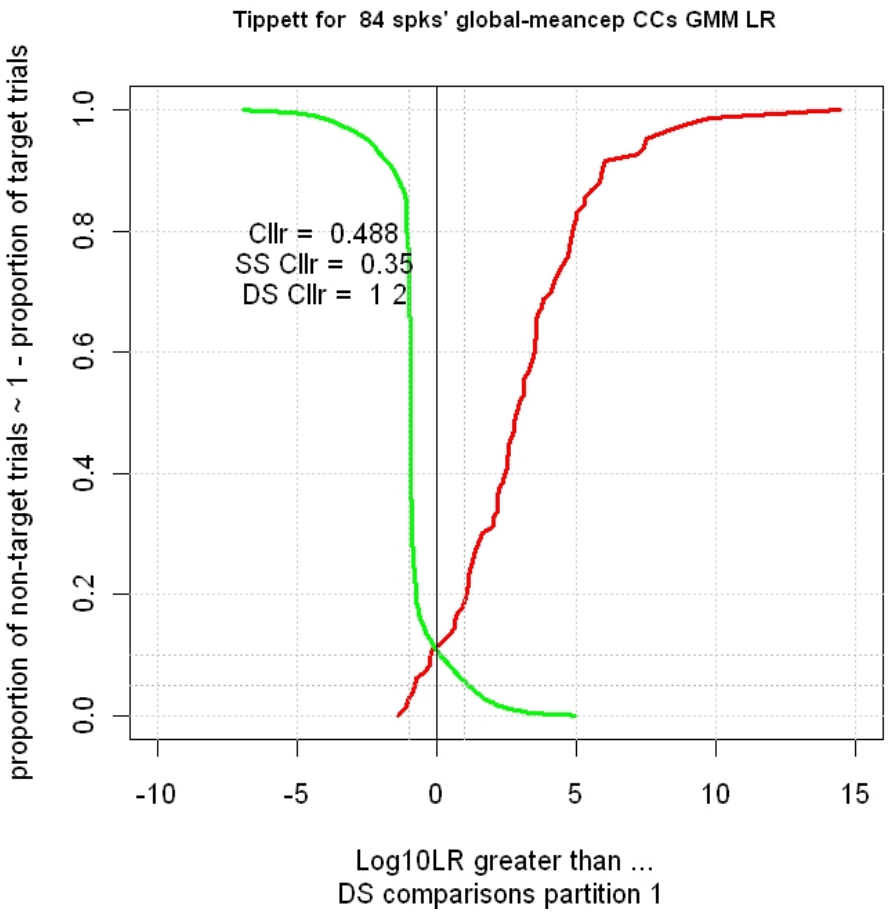
Non-linguistic

Fused & calibrated Tippett (MVLR)

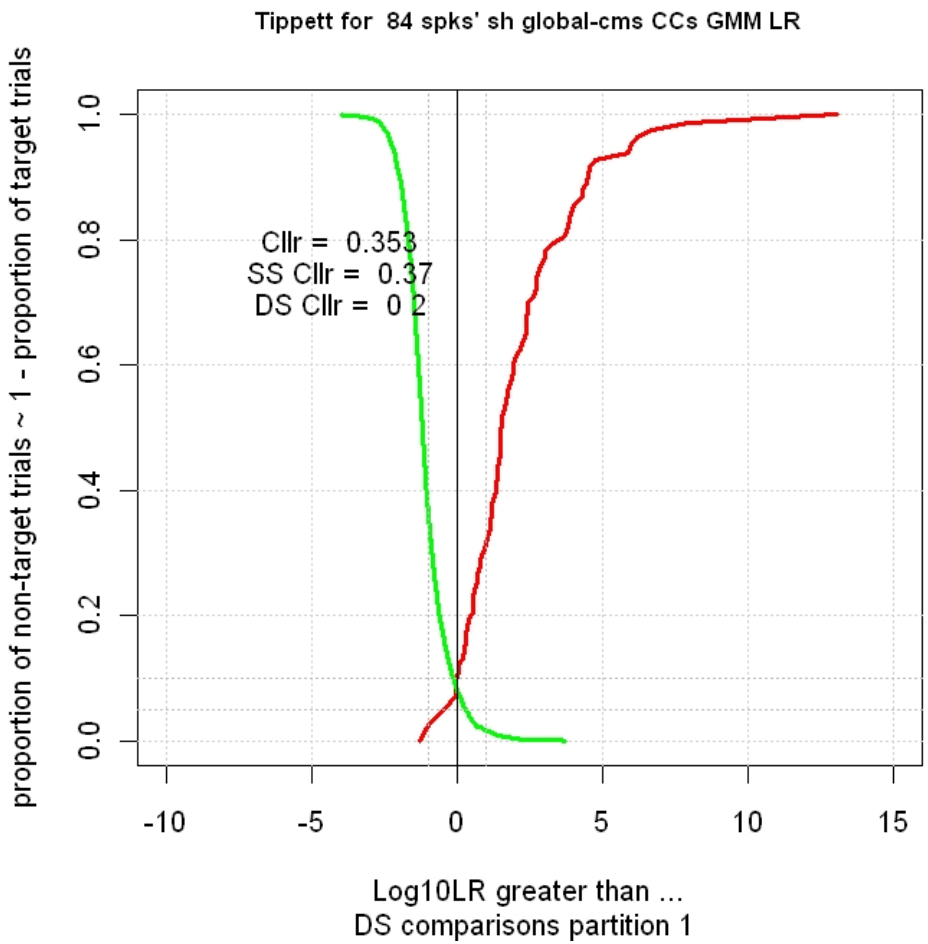


GMM/BM Results

Calibrated Tippetts: GMM/BM

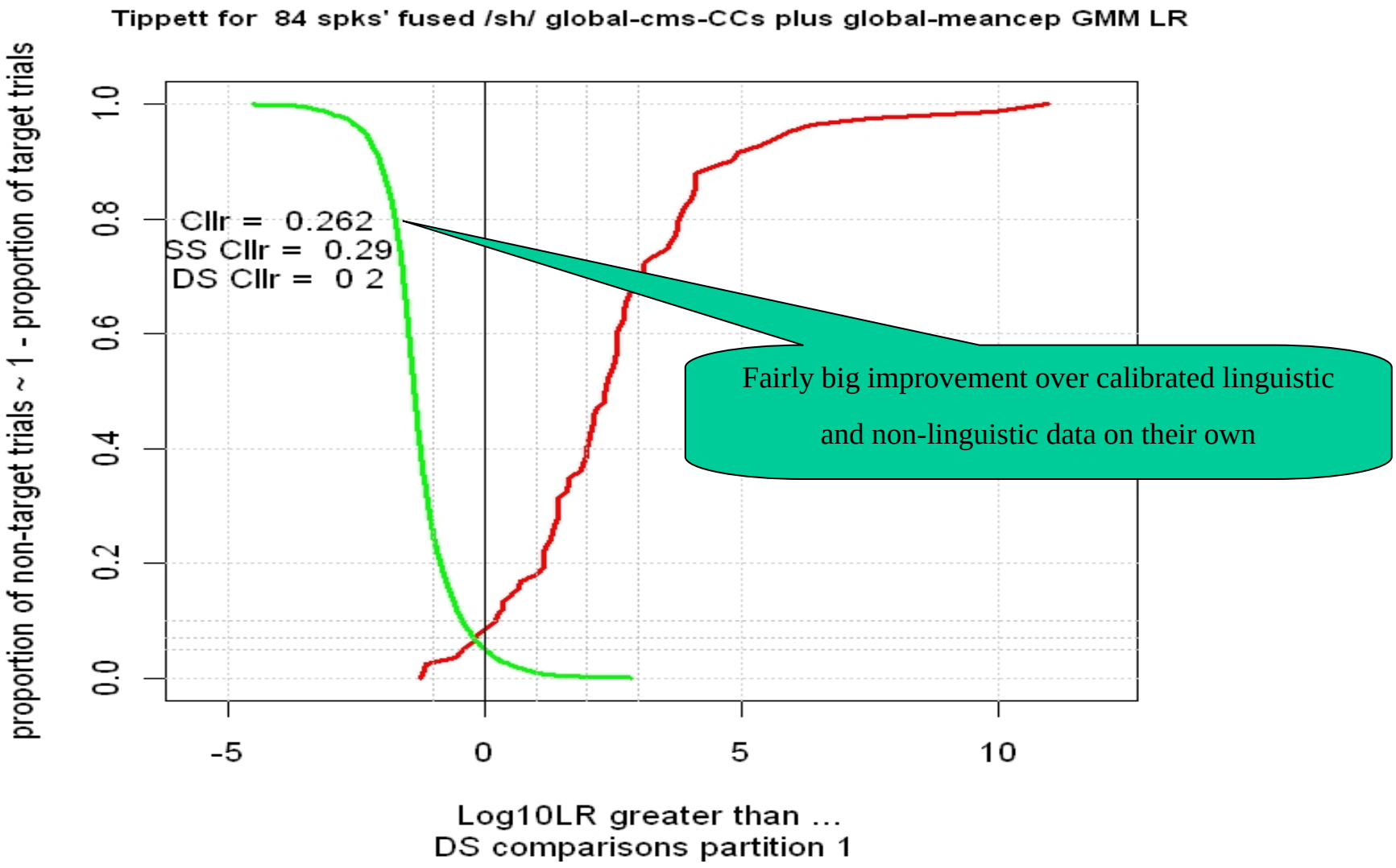


[ɕ]



Non-linguistic

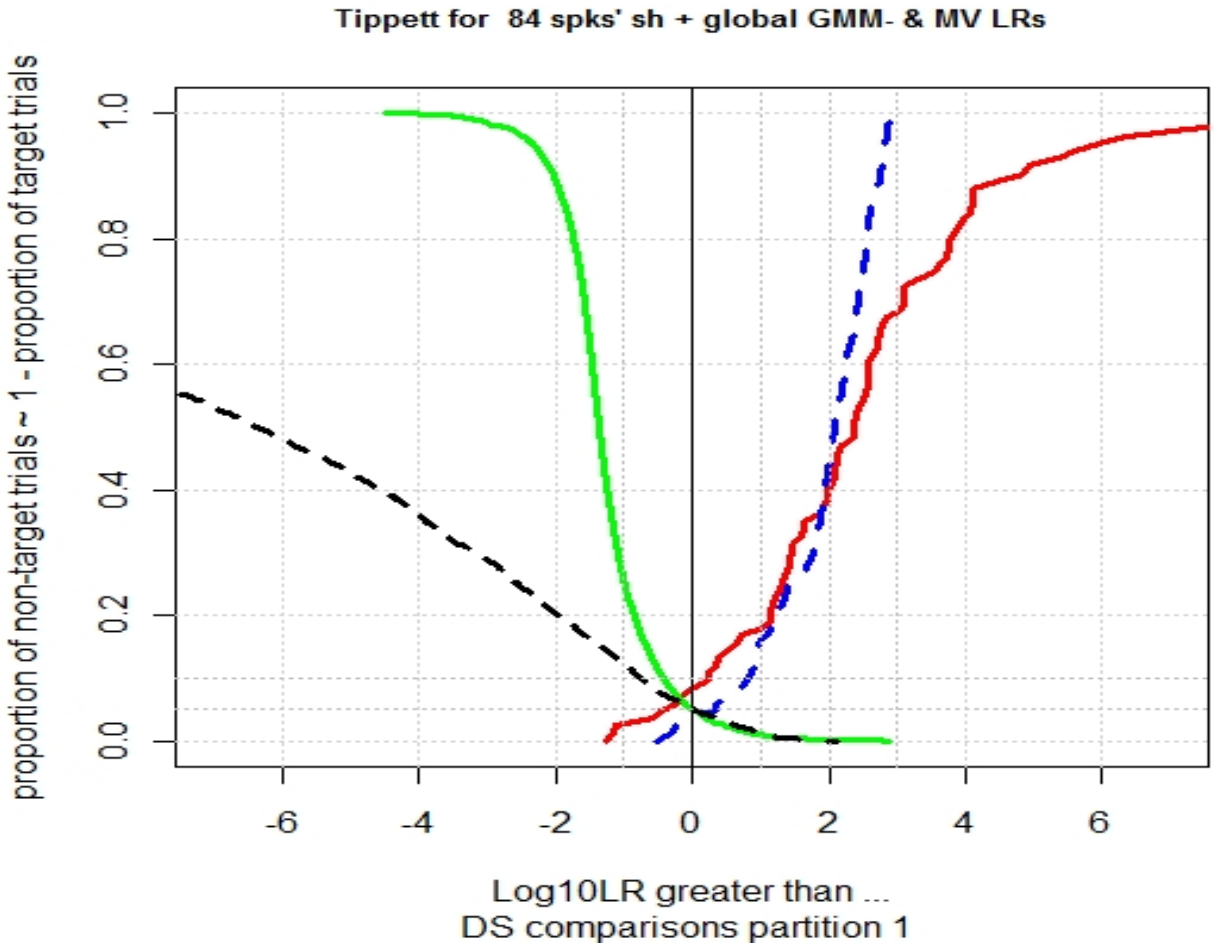
Fused & calibrated Tippett (GMM LR)



Conclusions

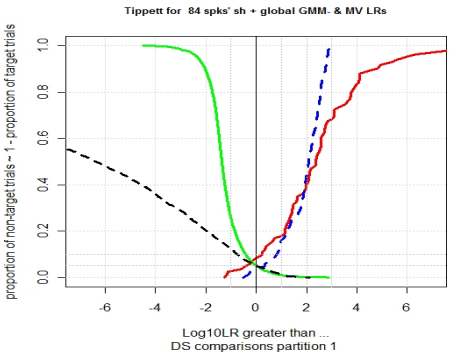
- Yes, it does improve strength of evidence estimates (both MV- and GMM- *both of which are good*) if you can combine linguistic with non-linguistic LRs.
- Spectrum of [ɕ] is useful forensic parameter IN CONJUNCTION WITH OTHERS
- This suggests that [ʃ ʂ ç] will also be of (perhaps greater) use;
- Perhaps also [s], but needs testing.
- But there is something else …

We have two rather different sets of LR estimates for the same data …



•Don’t chose … fuse!

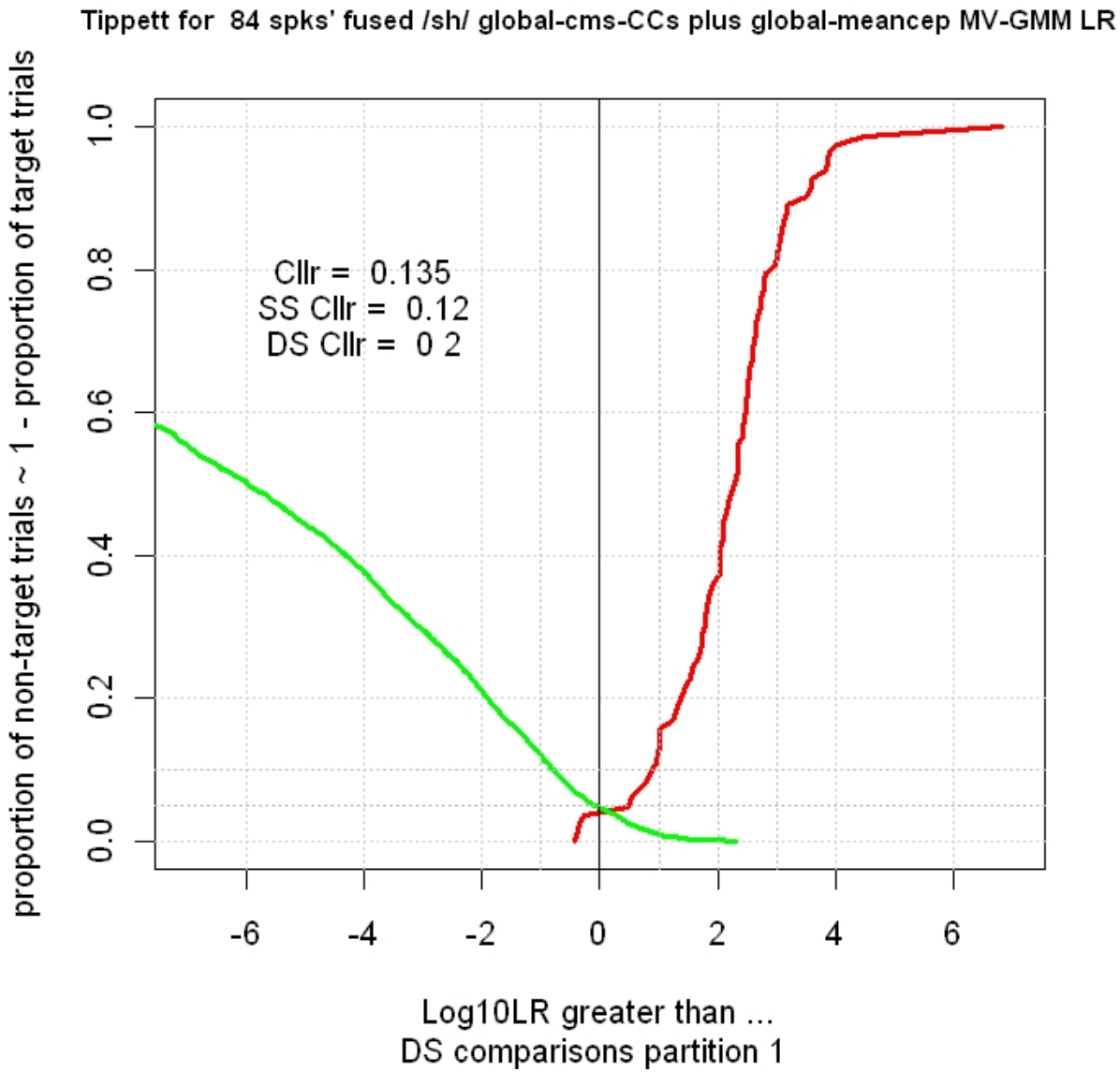
Fused hybrid-GMM-MV-LR Tippett



Cllr = 0.135

EER = 4.2%

Ca. 1% improvement over MV



Limitations

- Factors possibly contributing to too good results:
  - Training / test data not separated
  - Too much control over channel?
  - Jap. /ɕ/ *may* have inherently longer allos than, say, English /ʃ/ - easier for speaker to reach target

  (certainly the case before devoiced /i/)
- Also frics. not excluded from cepstral mean
- But, crude automatic processing:

  better channel compensation etc. would probably give better results

More Questions and further work

- MFCCs vs LPC CCs?? Might depend on segment.
- Channel compensation methods other than MCS? (or other types of MCS?).
- Band-limited cepstra …
- Incorporate formants (or peak-picked poles) …
- Do nasals, rhotics, laterals …

THANK YOU

Comments *very* welcome