System performance and speaker individuality in LR-based forensic voice comparison

Bruce X. Wang, Vincent Hughes and Paul Foulkes
Department of language and linguistic science, University of York

Introduction

Previous studies

- Largely focused on **generic system** testing using
  - Cost log likelihood ratio (C_{llr} , Brümmer & du Preez, 2006): SS LR / DS LR
  - Decision Error Tradeoff (DET) graph (Martin, 1997): false hit rate / right miss rate

- Limited study looked at the individual speaker's behaviour/performance (Lo, 2021)

Introduction

However,…

What is more important for forensic phoneticians?

- **Generic system** testing ?
  - i.e. in the context of research or a generic validation exercise, e.g. FP *um* is a good variable for FVC

- **Case-specific** testing?
  - i.e. individual speaker's behaviour, e.g. **speaker A** gives good performance using the FP *um*, how about **speaker B**?

- How generalisable is generic testing to case conditions?

Questions

Under different conditions, how is overall performance affected and how do individual speakers behave?

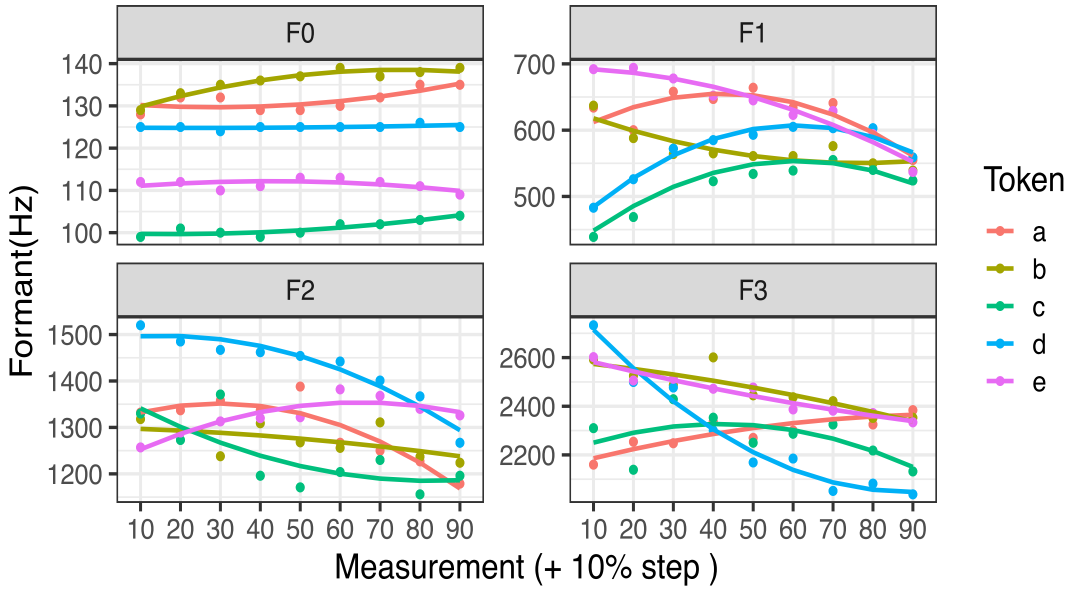
E.g.

- Change in the configuration of training and reference speakers
  - Given all speakers are from the **relevant population** (Hughes & Foulkes, 2015)
- Change in the use of parameters
  - Only use one parameter or use different combinations of different parameters

Method

Material

- 90 SSBE speakers (DyViS; Nolan et al., 2009)
  - Task 1: mock police interview
  - Task 2: telephone conversation
- Variable
  - FP ***um***
- Parameters
  - F0,F1, F2, F3
  - Nasal and vocalic duration
- Features
  - Quadratic coefficients
  - Duration



Five tokens, speaker 114 DyViS Task 1.

Method

LR computation

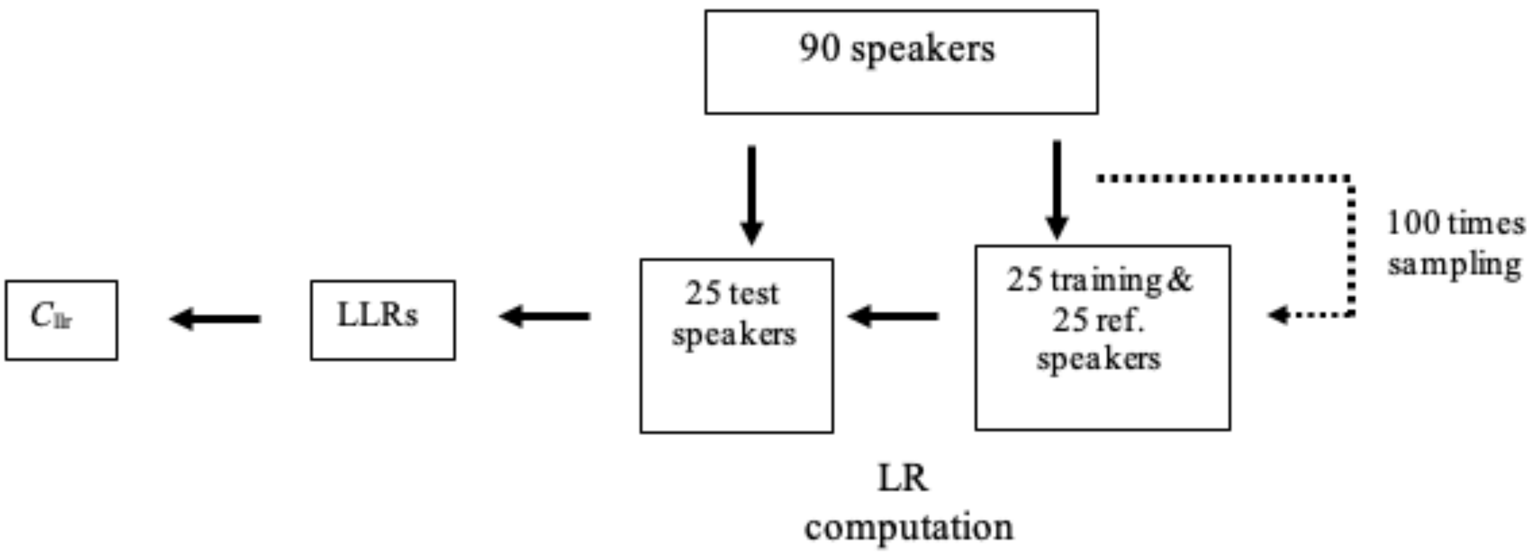
- 90 speakers divided into sets:
  - Test (25 speakers)
  - Training (25 speakers)
  - Reference (25 speakers)

- SS and DS LRs computed
  - Suspect data = Task 1
  - Offender data = Task2
  - MVKD (Aitken & Lucy, 2004)
  - Logistic regression (Brümmer et al., 2007)

Method

Change in the configuration of training and reference speakers

- Experiments replicated 100 times using same set of test speakers



Method

Change in the use of parameters

- Possible combinations of parameters tested (25 systems)

SYSTEM	F0	F1	F2	F3	DUR.
F0	X				
F1		X			
F2			X		
F3				X	
DUR					X
F01	X	X			
F02	X		X		
F03	X			X	
F0DUR	X				X
F12		X	X		
F13		X		X	
F1DUR		X			X

SYSTEM	F0	F1	F2	F3	DUR.
F23			X	X	
F2DUR			X		X
F3DUR				X	X
F012	X	X	X		
F013	X	X		X	
F01DUR	X	X			X
F123		X	X	X	
F12DUR		X	X		X
F23DUR			X	X	X
F0123	X	X	X	X	
F012DUR	X	X	X		X
F123DUR		X	X	X	X
F0123DUR	X	X	X	X	X

Method

Evaluation

System

- C_{llr} (Brümmer & du Preez, 2006)
  - Mean C_{llr} over 100 runs
  - Overall range, i.e. max. C_{llr} – min. C_{llr}

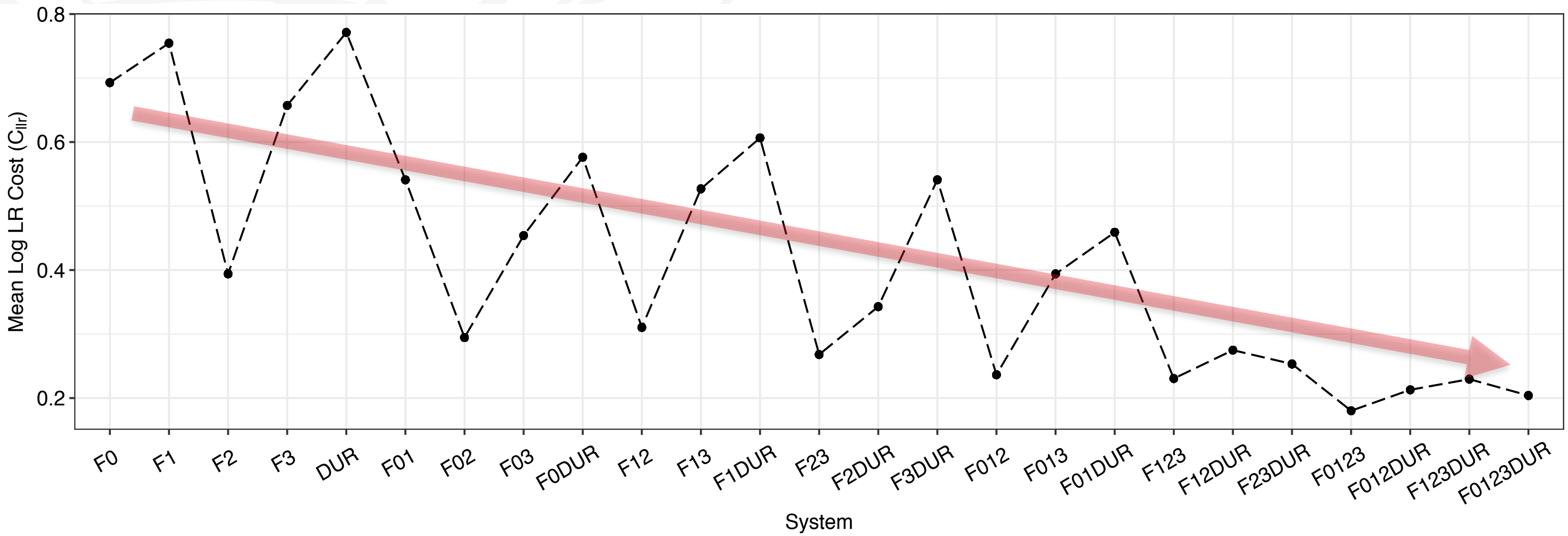
Individual speaker

- SS and DS RMSE for each speaker
  - $RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_i - y_i)^2}$
  - x_i : individual LLR in each comparison
  - y_i : mean LLR of each individual speaker over 100 runs
  - Captures how variable the results are for each speaker

Results

System

Accuracy

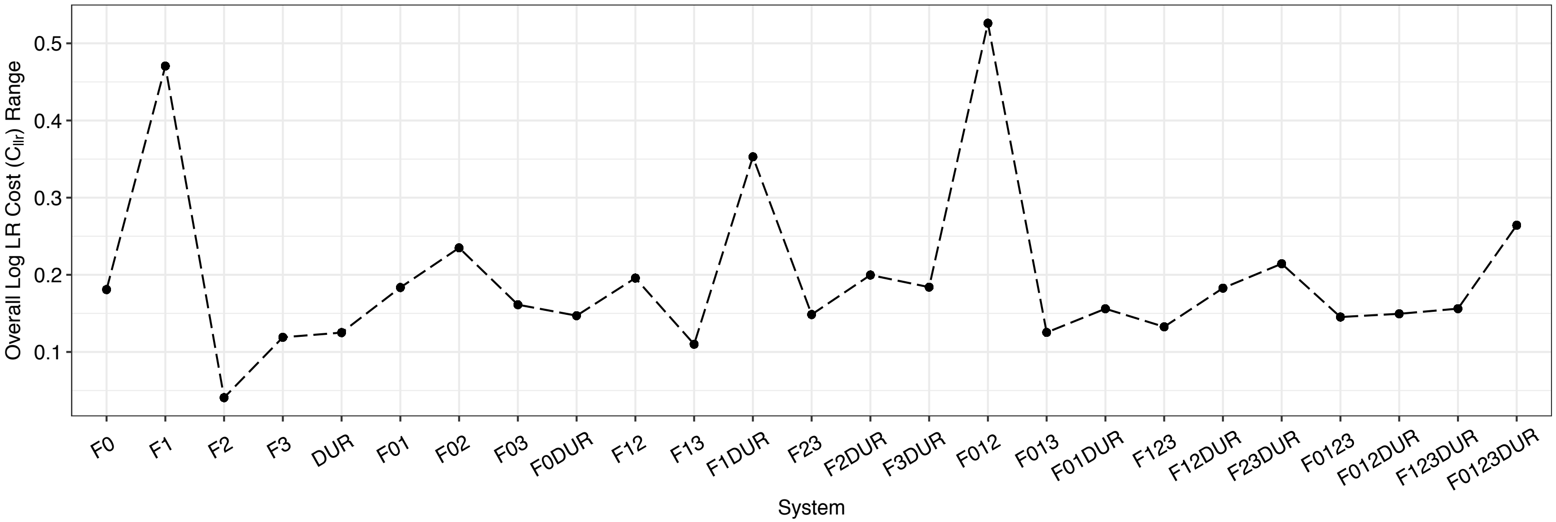


- Systems with more parameters yield higher accuracy
- Starts to stabilise when four or more parameters are used
- Adding extra parameter does not necessary improve system accuracy, e.g. duration
- Systems with F2 involved yield higher accuracy
  - McDougall & Nolan (2007) obtained higher classification rate using the F2 of /u:/ from DyViS speakers

Results

System

Stability



- Systems with four or five parameters in general have lower C_{llr} OR than the rest
  - e.g. F0123, F012DUR, F123DUR, F0123DUR vs. F1DUR and F012
- Exceptions

System/C_{llr}	Min.	Max.	OR
F2	0.37	0.41	0.04
F13	0.48	0.59	0.11
F013	0.33	0.46	0.13

VS.

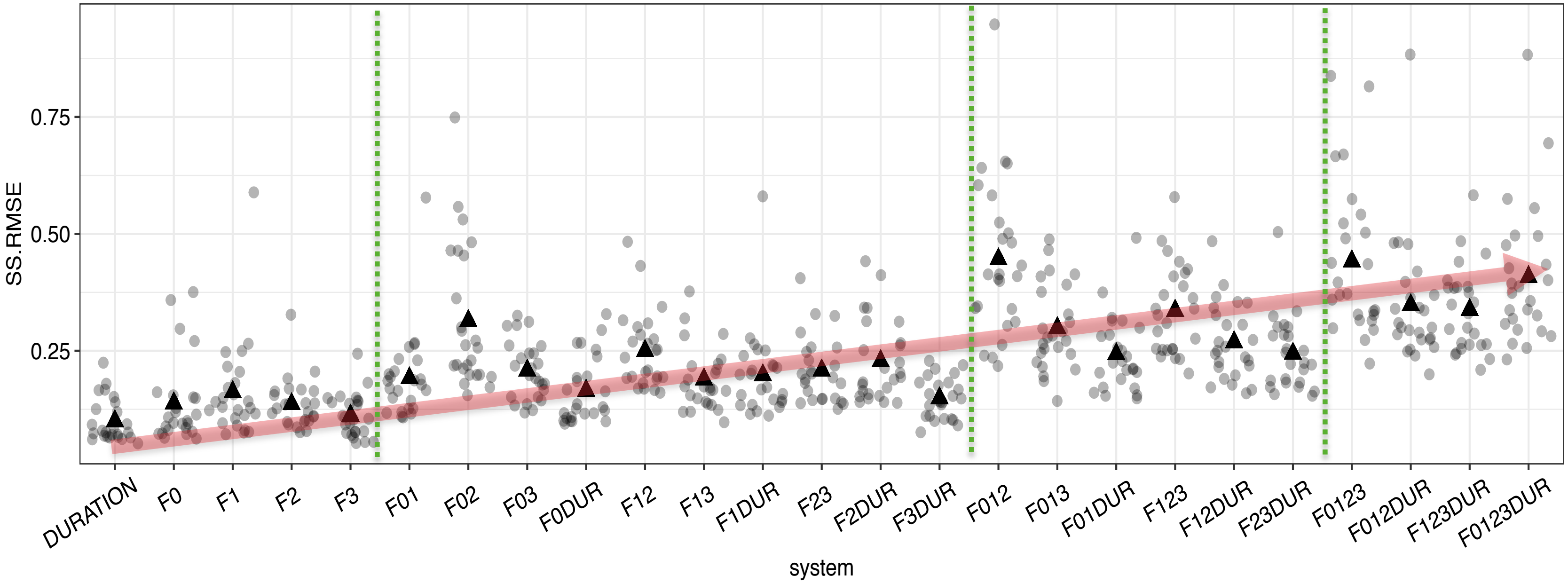
What is the trade-off?

System/C_{llr}	Min.	Max.	OR
F0123	0.12	0.27	0.15
F123DUR	0.16	0.32	0.16
F0123DUR	0.11	0.38	0.26

Results

Individual

SS LLR RMSE

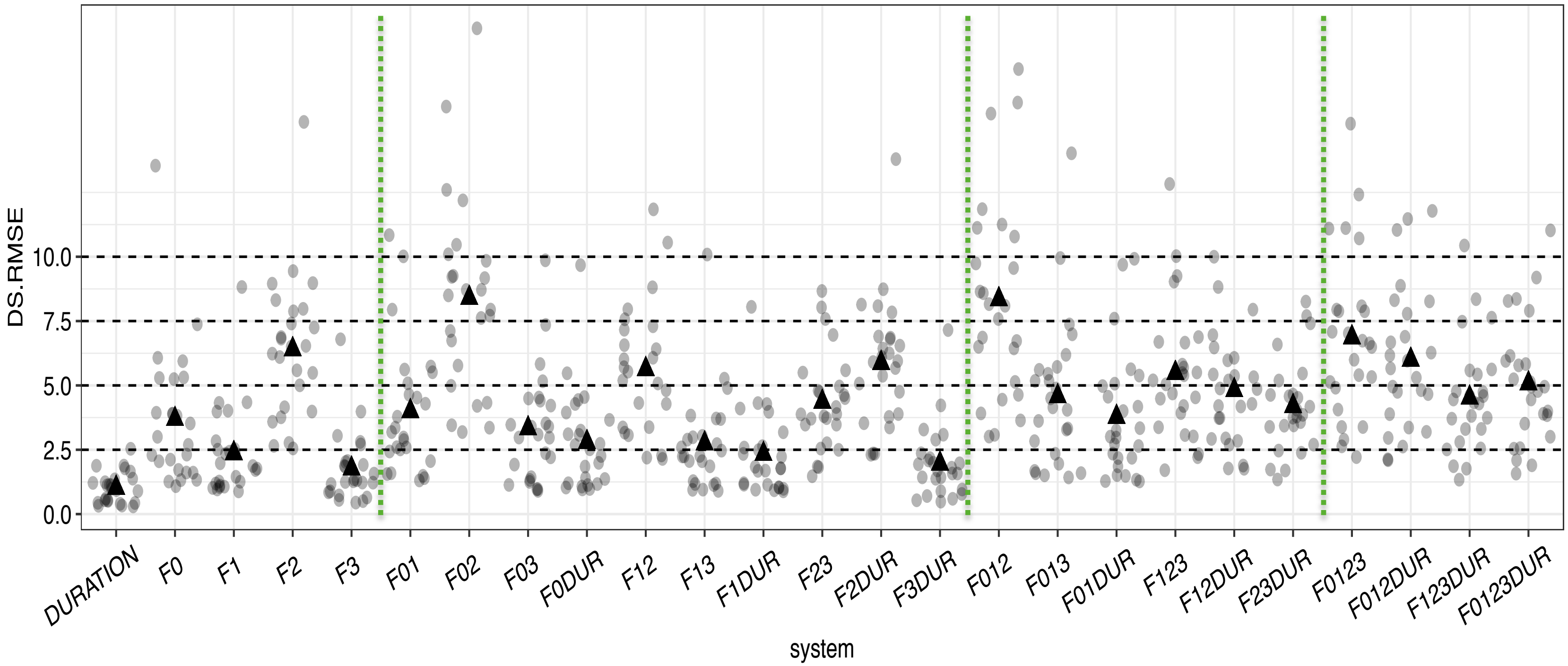


- Speakers fluctuate more with more parameters
- The SS LLR is stable in systems with equal number of parameters
  - i.e. different combinations of parameters do not have much effect on individual speakers' reliability in SS comparisons
- All SS.RMSE < 1
  - Seems to be less problematic in SS comparisons

Results

Individual

DS LLR RMSE

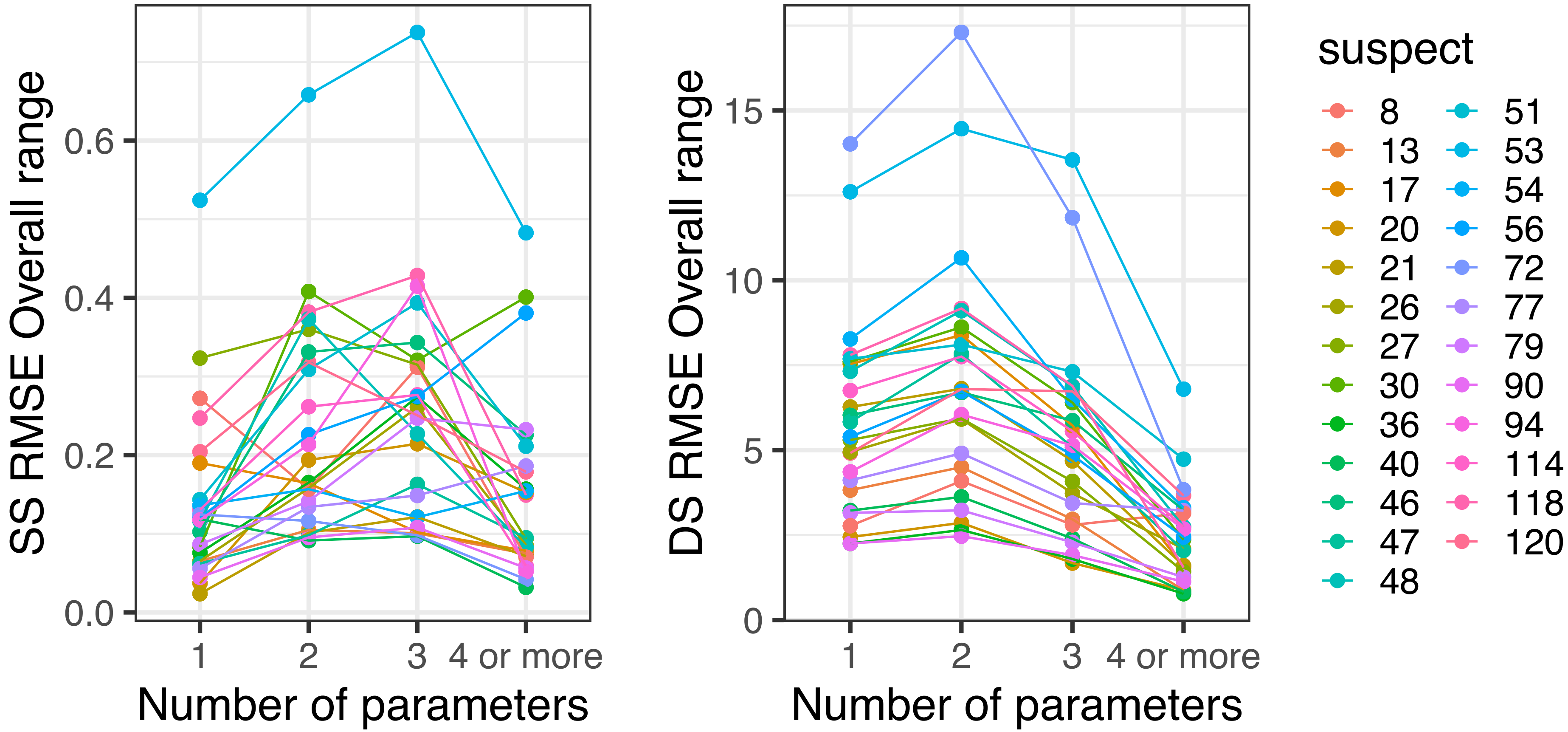


- No general pattern of individual speakers' reliability when more parameters are used
- Individuals fluctuate more when F2 involved

Results

Individual

- Less fluctuating when 4 or more parameters are used ?



Conclusion

- Generic system testing and case-specific testing are equally important

  - Systems are more accurate and stable with more parameters, irrespective of training and reference speaker used

  - Individuals are more stable with four or more parameters, irrespective of different combinations of parameters, especially in DS comparisons

  - The trade-off between system accuracy and stability is case specific

  - Individual speaker's performance is case specific

- Instead of using all parameters available under real case scenarios, system performance should be tested using different combinations of parameters

Thanks !
Questions ?

Special thanks to Justin Lo

Reference

Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(1), 109–122. https://doi.org/10.1046/j.0035-9254.2003.05271.x

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*(2–3), 230–275. https://doi.org/10.1016/j.csl.2005.08.001

Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2072–2084. https://doi.org/10.1109/TASL.2007.902870

Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication*, *94*, 15–29. https://doi.org/10.1016/j.specom.2017.08.005

Hughes, V., & Foulkes, P. (2015). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication*, *66*, 218-230. https://doi.org/10.1016/j.specom.2014.10.006

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). *The DET curve in assessment of detection task performance*. National Inst of Standards and Technology Gaithersburg MD.

Ishihara, S., & Kinoshita, Y. (2008). *How Many Do We Need? Exploration of the Population Size Effect on the Performance of Forensic Speaker Classification*. 4

Kinoshita, Y., & Ishihara, S. (2015). Background population: How does it affect LR based forensic voice comparison? *International Journal of Speech Language and the Law*, *21*(2), 191–224. https://doi.org/10.1558/ijsll.v21i2.191

Lo, J. (2021). Seeing the trees in the forest: Diagnosing individual performance in likelihood ratio based forensic voice comparison. *Proc. XVII National Conference of the Italian Association for Speech Sciences.* Zurich, Switzerland.

Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech Language and the Law*, *16*(1), 31–57. https://doi.org/10.1558/ijsll.v16i1.31