Can speakers make themselves more recognisable?: Voice dynamics and its influence on voice recognition

Volker Dellwo

Phonetics & Speech Sciences Group

Department of Computational Linguistics

University of Zurich

+ Thayabaran Kathiresan (Telepathy Labs)
+ Elisa Pellegrino (U Zurich)
+ Alexis Hervais-Adelman (U Zurich)
+ Valeriia Perepelytsia (U Zurich)
+ Leah Bradshaw (U Zurich)
+ Sandra Schwab (U Zurich)
+ Lei He (U Zurich)
+ Simon Townsend (U Zurich)
+ Moritz Daum (U Zurich)
+ Dieter Maurer (Zurich U of the Arts)
+ Rushen Shi (U Quebec at Montreal)
+ Laura Dilley (Michigan State U)
+ Sarah Lim (U Zurich)

Swiss National Science Foundation

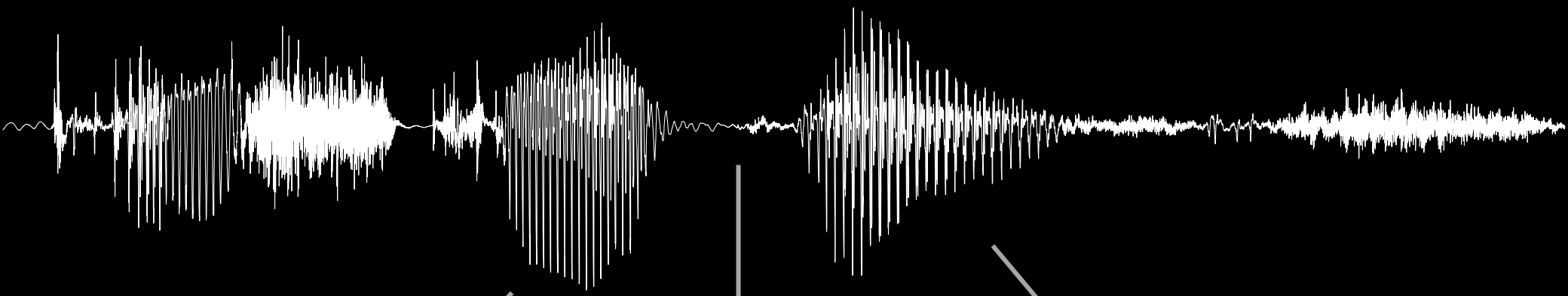
nccr evolving language

Linguistic Research Infrastructure

University of Zurich | CFPA

CENTER FOR FORENSIC PHONETICS & ACOUSTICS

Speech signals contain several types of information



What is being said?
(linguistic information)

- messages
- words
- utterances
- etc.

How is it said?
(speaker-state
or paralinguistic information)

- mood
- emotion
- state of health
- etc.

Who says what?
(speaker-individual or
indexical information)

- gender
- age
- origin
- etc.

From a communicative point of view, indexical information is typically viewed as ...

… unwanted information or ‘noise’.

… Index = pointer (Peirce; semiotic theory)

- No direct communicative intent
- E.g. smoke is an index of fire
- E.g. nasality is an index of a cold

… an uncontrolled by-product:
speakers involuntarily give it away

… static: does not change over the course of an utterance/discourse

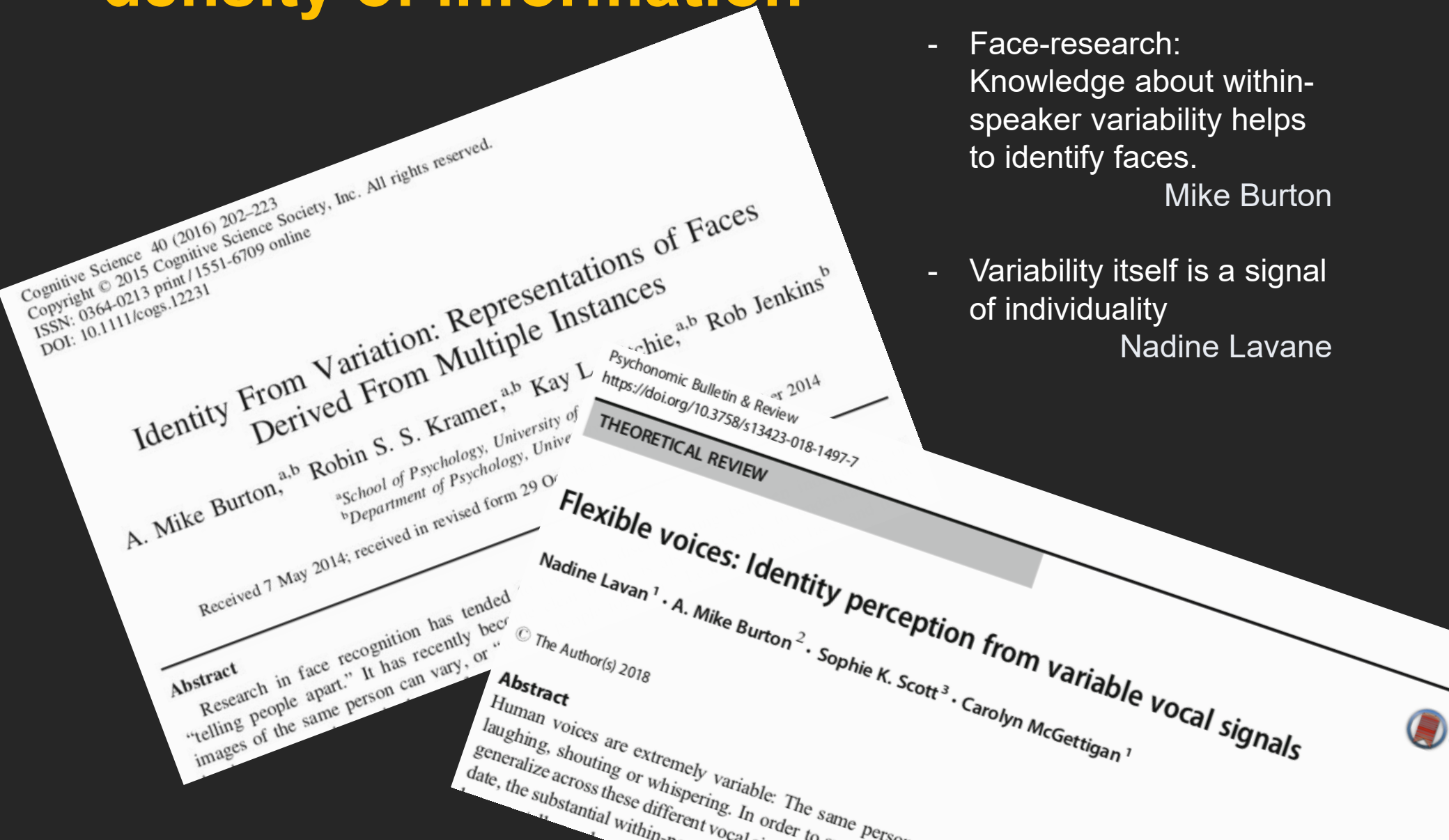
Indexical information HAS communicative function

Oh dear, I did not know you split up.
Yes, the whole relationship was a nightmare.
What was the problem?
Well, he was just not ready to let me into his live.
Why was that?
Well, you know, societal pressure.
He just could not take who I really was.

Recognizing (discriminating between) speakers is crucial in processing speech communication.

Consequently, it maybe beneficial to be able to modify our recognizability by controlling indexical properties.

Recognizable information varies with density of information



- Face-research: Knowledge about within-speaker variability helps to identify faces.

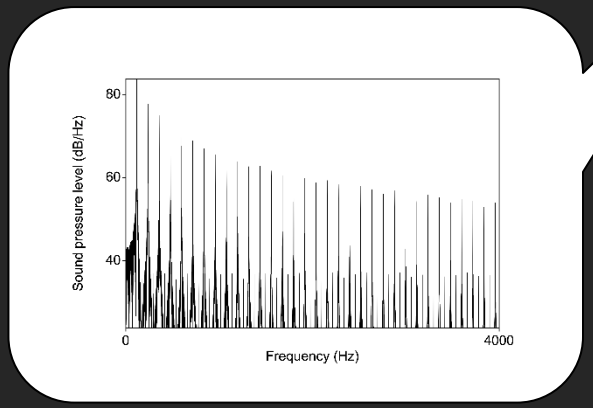
  Mike Burton

- Variability itself is a signal of individuality

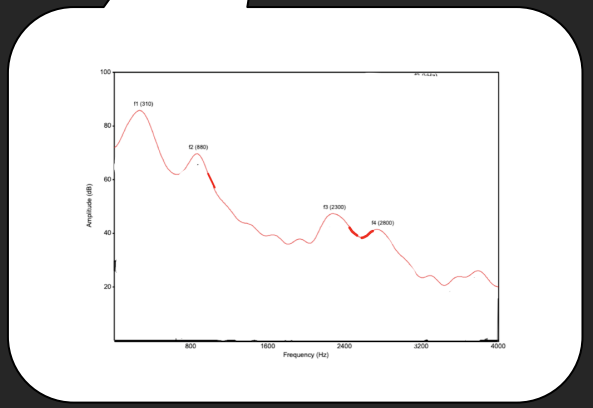
  Nadine Lavane

How may indexicality be controlled?

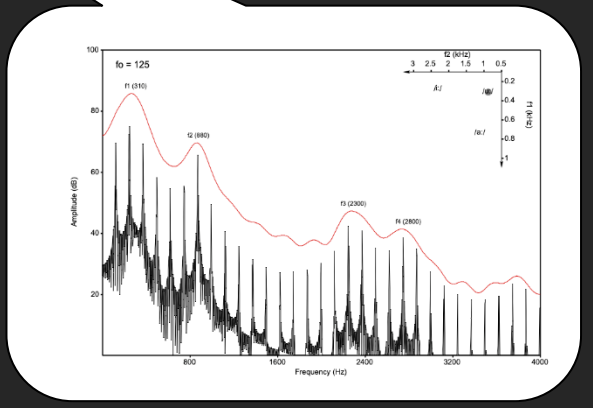
The vocal face



Source



Filter



Speech

For the vocal face to be be maximally identifiable it is essential that vocal tract detail is rich.

How can vocal tract detail be exposed?

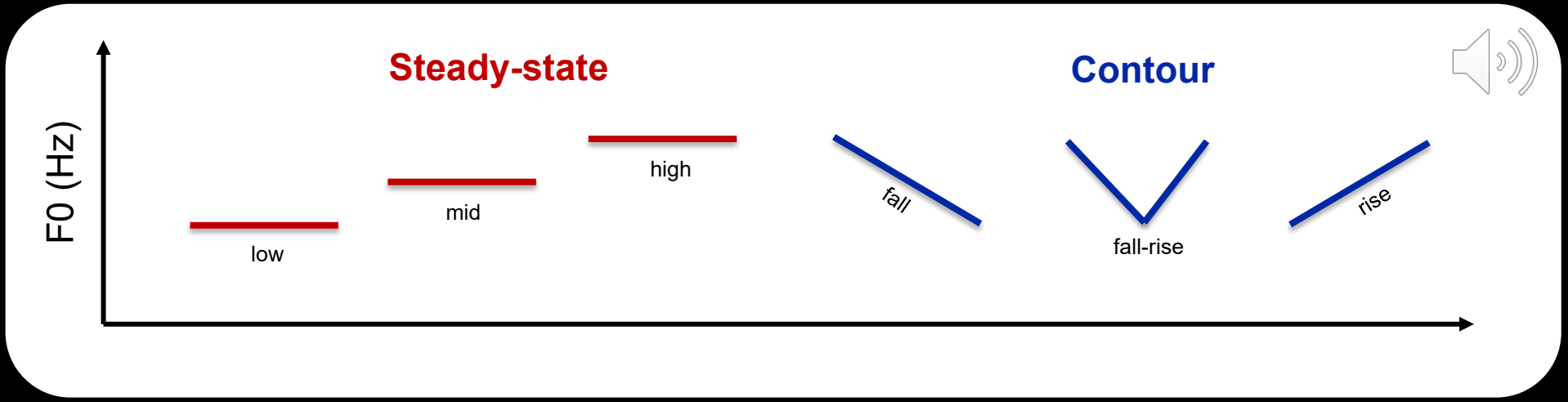
Sweeping harmonics through the vocal tract should provide richer indexical information.

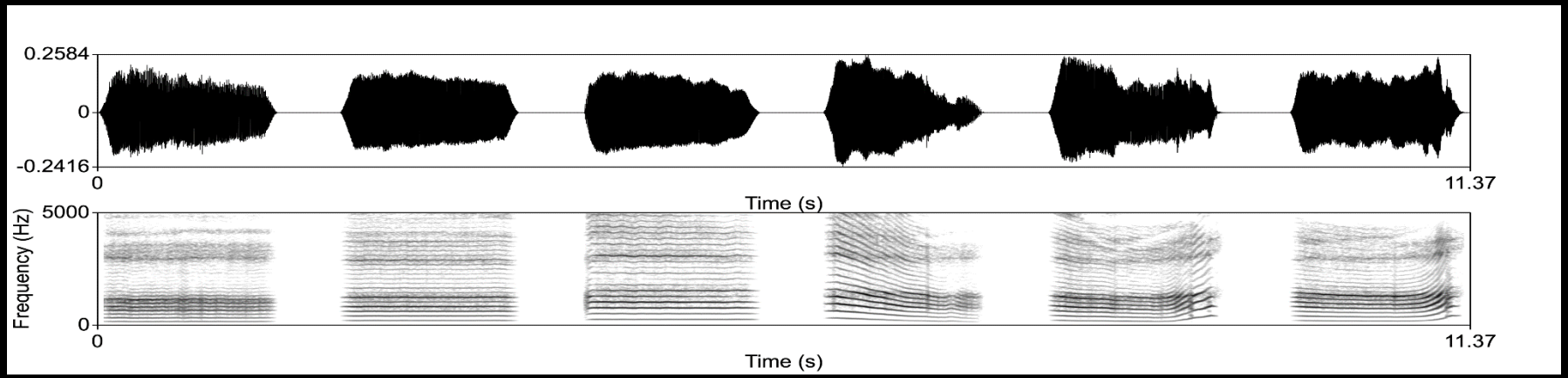
Hypothesis: Speakers are better recognizable when sweeping as compared to when producing steady state vowels.

Experimental design

(a) Training: sentence utterances read by 15 speakers

(b) Test material:

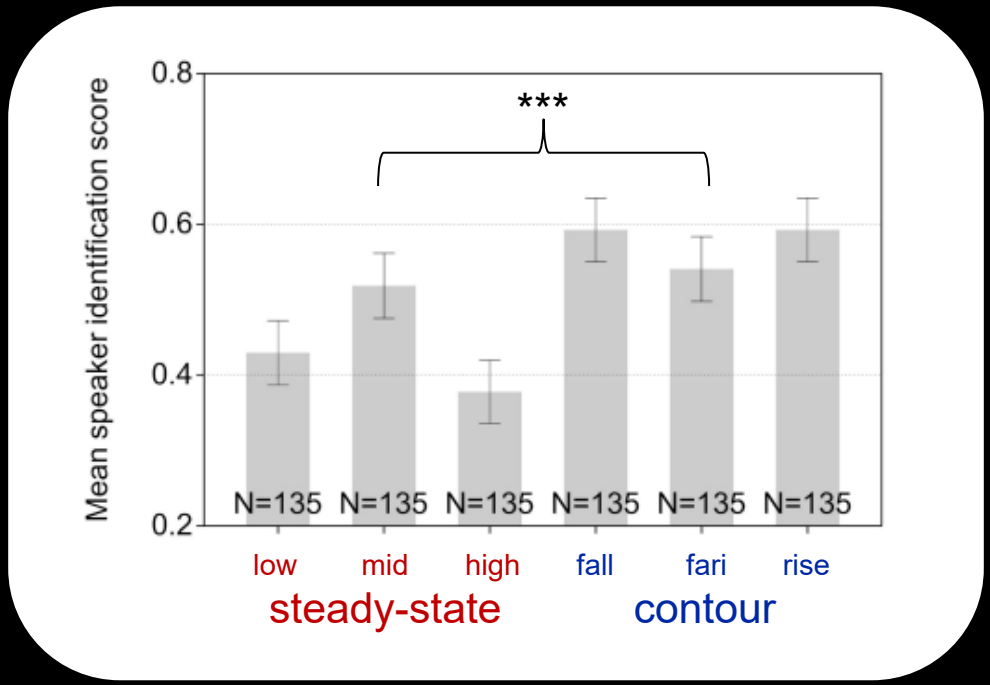




Recognition results

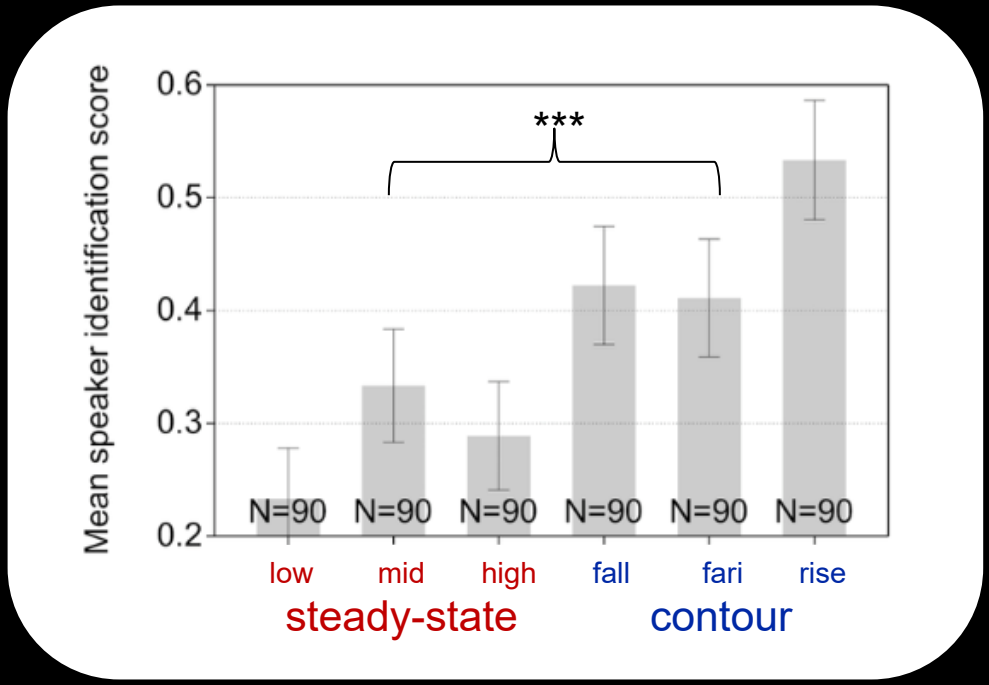
Computer (15 voices)

Gaussian Mixture Model based on 13-dimensional MFCCs



Humans (4 f voices)

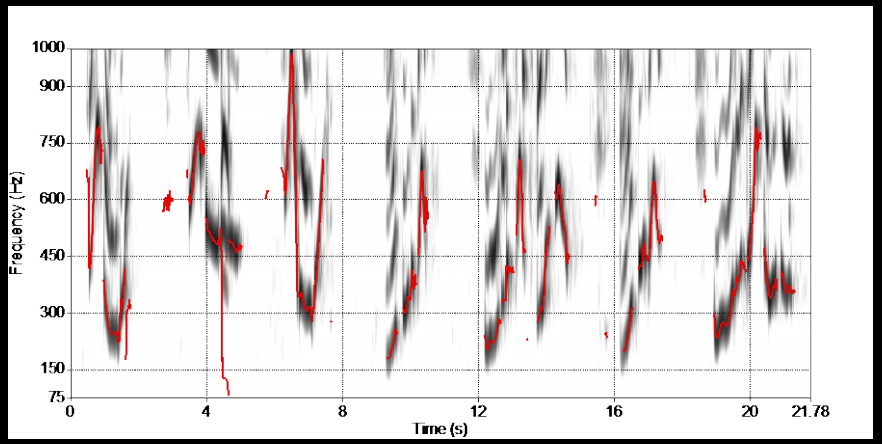
Training to 75%C before test.



Interim conclusions & questions

Sweeping harmonics reveal more information about the vocal tract that leads to better recognition performance.

Under what natural circumstances would speakers do this?



Hypothesis: Speaking styles should have an impact on recognition depending on their use of sweeping

Speaking style variability

Infant Directed Speech

- High-sweeping
- high-intonation

Social Bonding between mother and infant.

Data UZH/
Michigan State U

Clear Speech

- Low-sweeping
- High-intonation

Targeted at intelligibility

UCL LUCID Corpus

Deceptive Speech

- Mumbled speech
- Low variability

Speaker has no interest in revealing identity.

Columbia Deceptive Speech

How to test?

Focus on mismatch conditions:

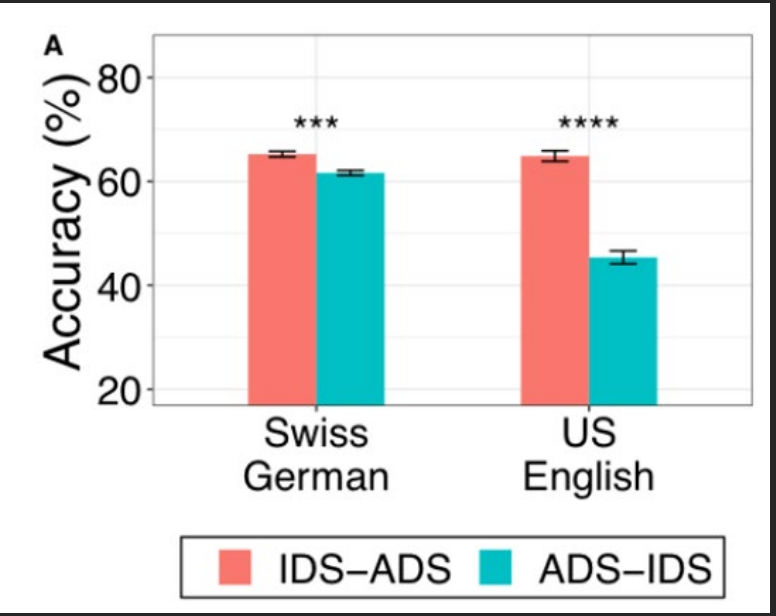
- Train system with one style, test in another (e.g. train with IDS (high variability) test with ADS (low variability) and vice-versa)

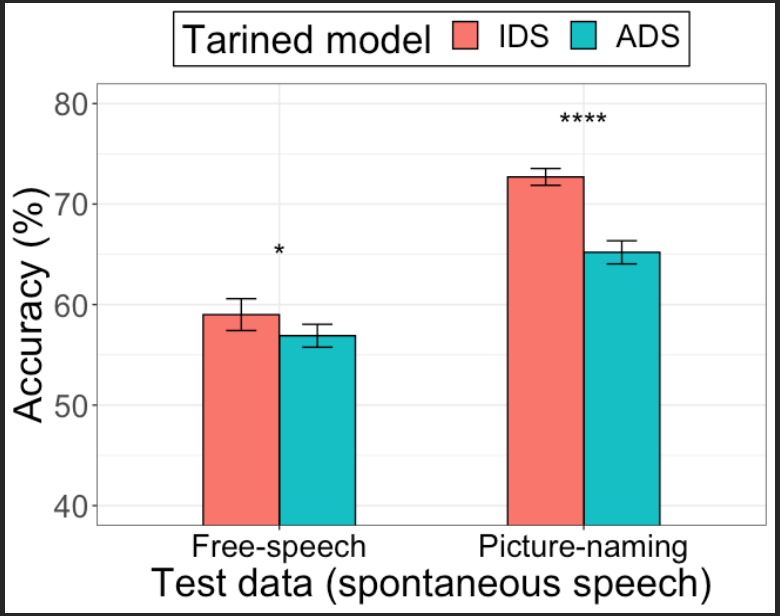
Test effects of high and low variability on non related styles:

- Train with IDS or ADS (high and low variability) test with other spontaneous speech

Recognition here: different forms of automatic models (typically GMM based on MFCCs acoustic modelling) and/or human listeners.

Infant- and adult-directed speech

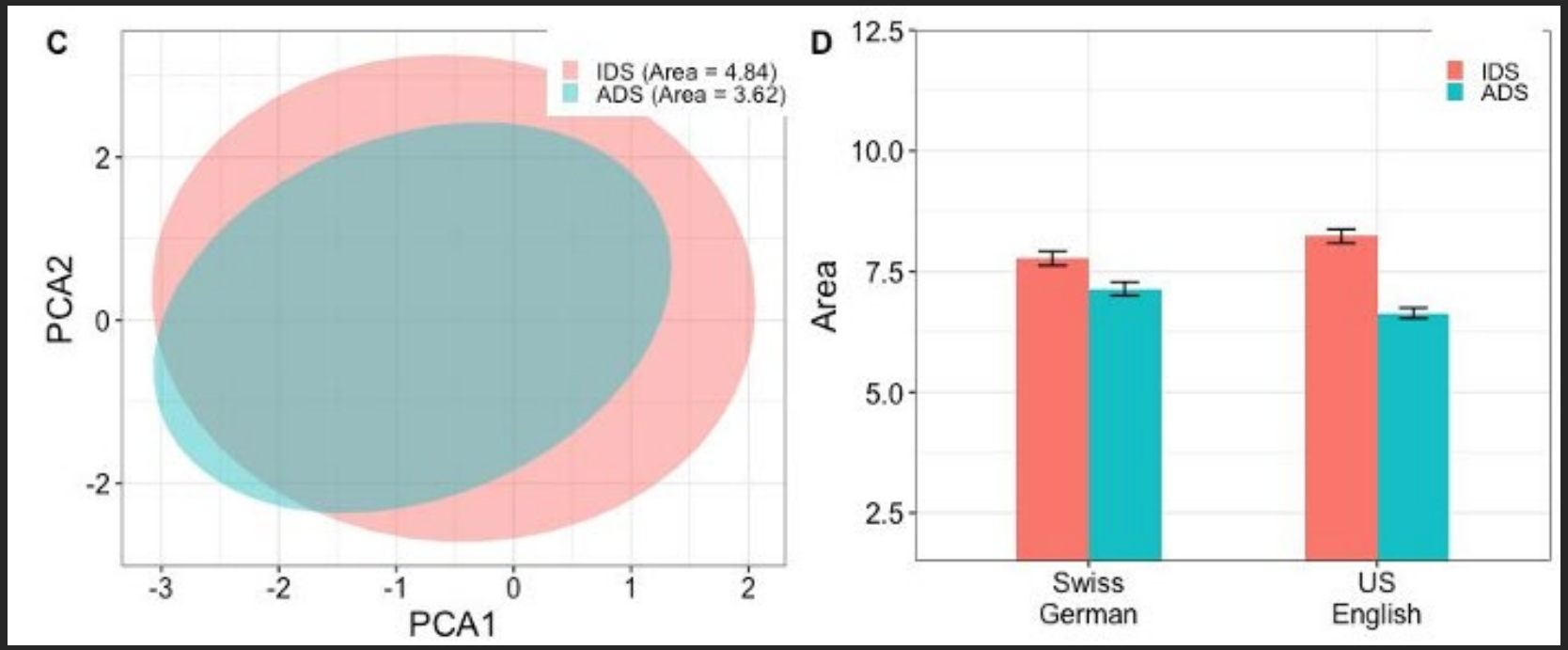




GMM based on MFCC acoustic modelling

- → **Recognition advantages of using IDS as training.**
- → **Plausible in terms of language evolution.**

Acoustic explanation:



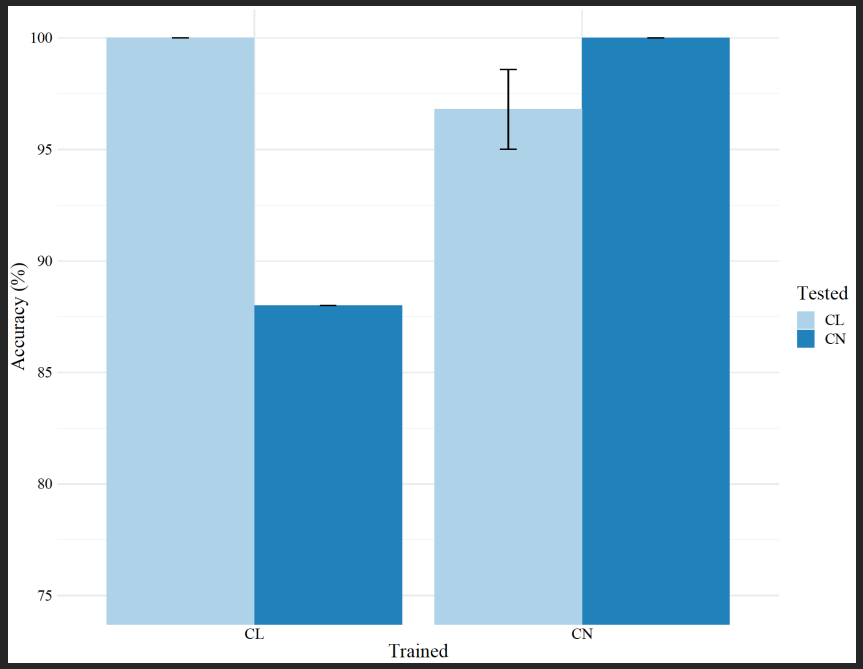
→ **GMM (32 clusters on 13-dimensional MFCCs)**
→ **Acoustic 'space' in IDS is larger than in ADS**
→ **ADS is a 'subspace' of IDS, i.e. knowing the speaker under IDS means knowing the speaker under ADS but not vice-versa.**

Clear- and conversational speech

High variability between segments but low within segment variability

Clear-speech targeted at intelligibility, i.e. speaker specific variability should be reduced.

→Use UCL LUCID corpus to train and test



→Training clear and testing conversational: performance drops

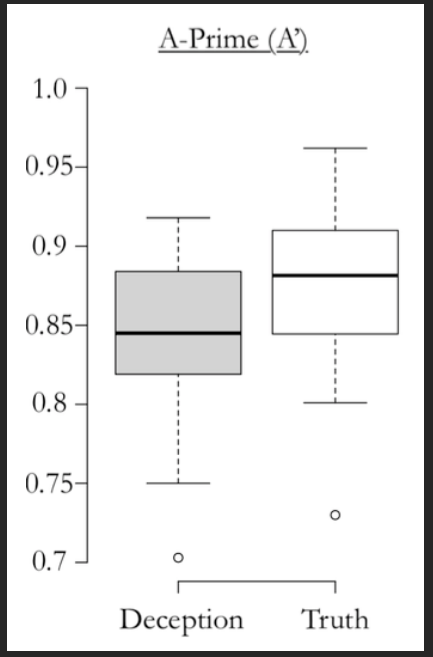
→Clear speech contains less information about the speaker.

Deceptive-speech

When speakers lie, they are not interested in revealing their identity.
Recognition ability also affects speaker memory.

→Use *Columbia University Deceptive Speech Corpus*

		Train	
		TU	LD
Test	TU	100	75
	LD	87.5	100



Human voice discrimination (same/different)

→ **Learning speaker in 'lie' reduces recognition ability**

Recognition summary

(in numerous tests)

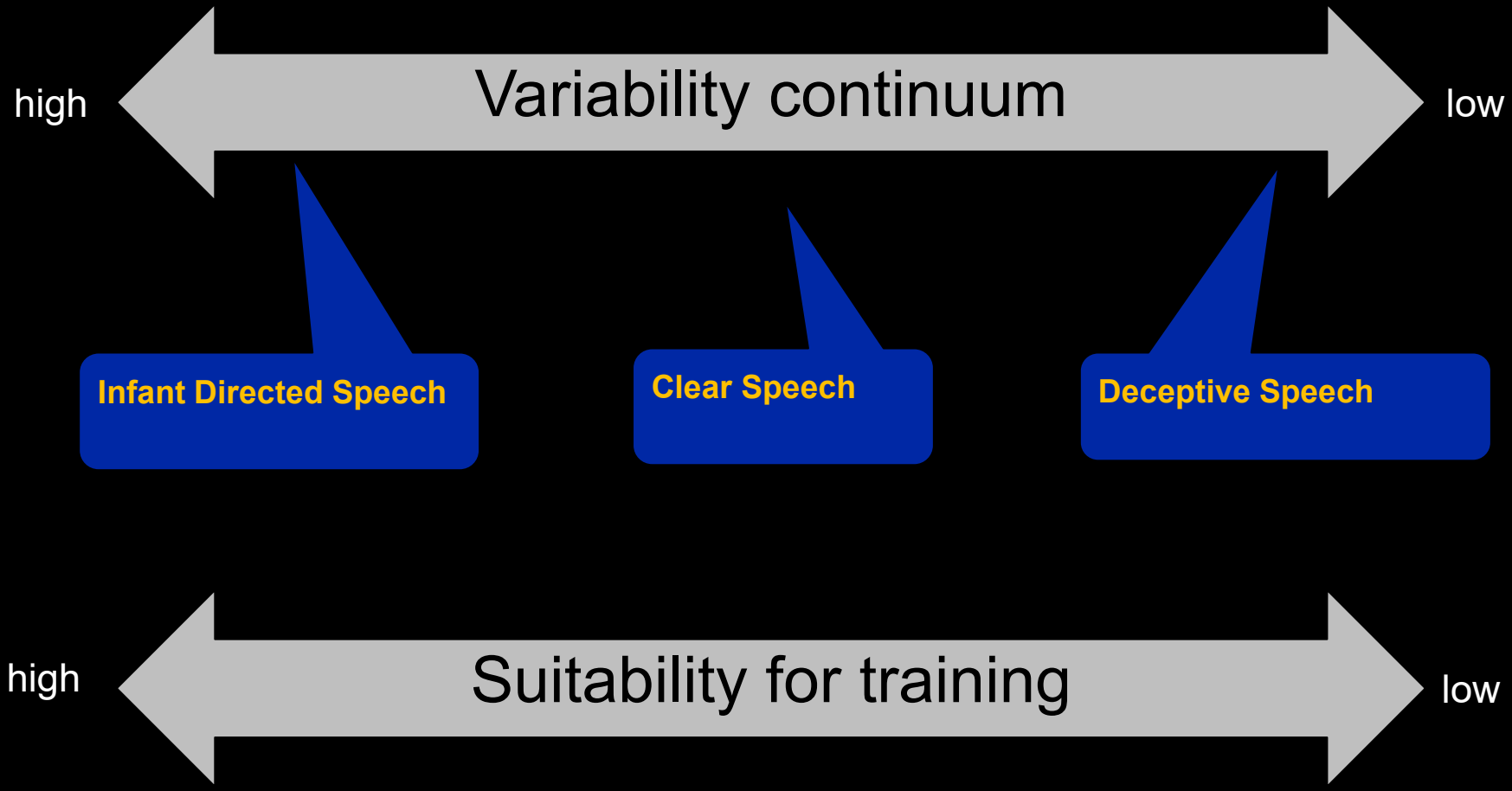
Train	*Test*	*Accuracy*
IDS	*various*	***high***
Clear	*various*	***low***
Deceptive	*various*	***low***

Acoustic similarity (PLDA on i-vectors):

IDS = high between and within speaker variability

Clear and deceptive = reduced between speaker variability

Conclusion


high
Variability continuum
low
Infant Directed Speech
Clear Speech
Deceptive Speech
high
Suitability for training
low

Conclusion

- Higher variability may contain more speaker specific detail (e.g. sweeping of pitch)
- Speaking styles vary in their use of these features
- This variability contributes to recognizabilty of speakers
- Supports the view: identity can be controlled by the speaker

Future work:
Can speakers control recognizability?

Find: Identity Marked Speech

Speech recognizer

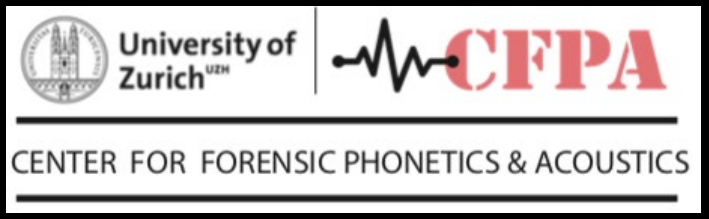
Humans communicate with a mock speech recognizer that performs numerous mistakes.

→ Speakers apply CLEAR SPEECH

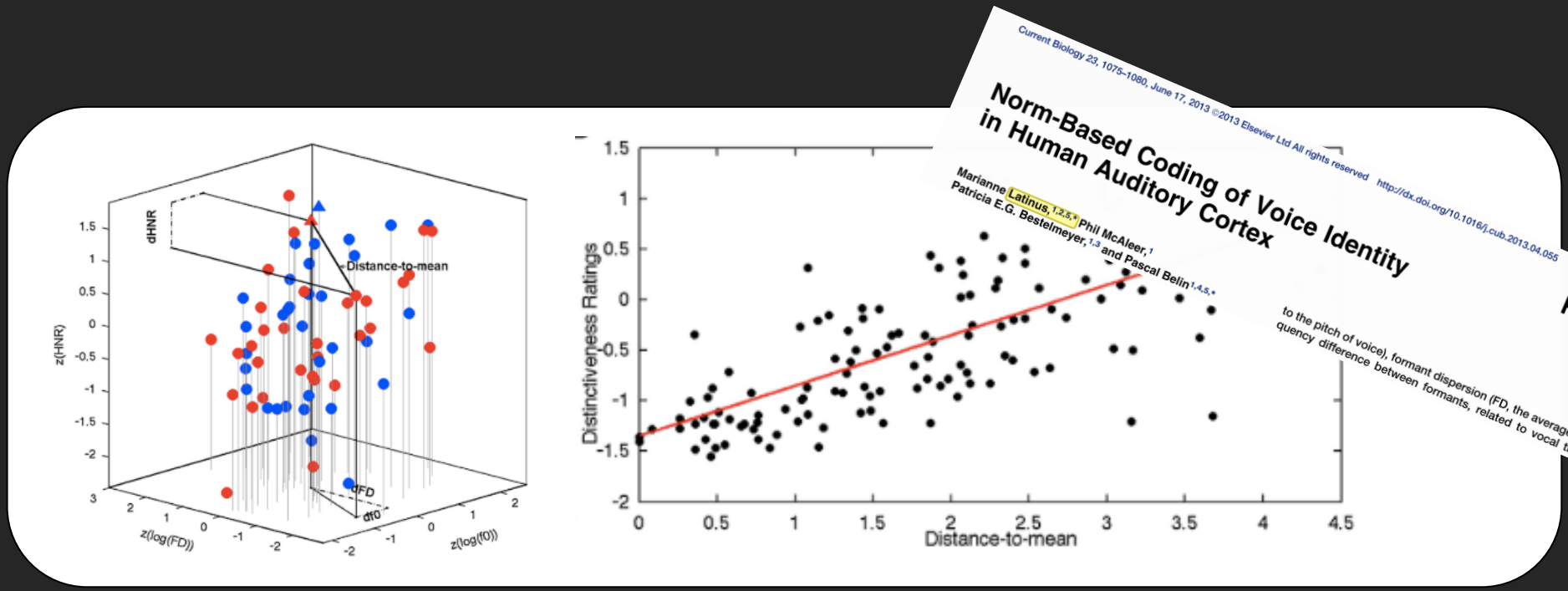
Voice recognizer

Humans verify their voice with a mock voice verification systems that often does not recognize them correctly.

→ Speakers apply IDENTITY MARKED SPEECH (?)



Fit into models of voice recognition



→ Currently not for within-speaker variability.

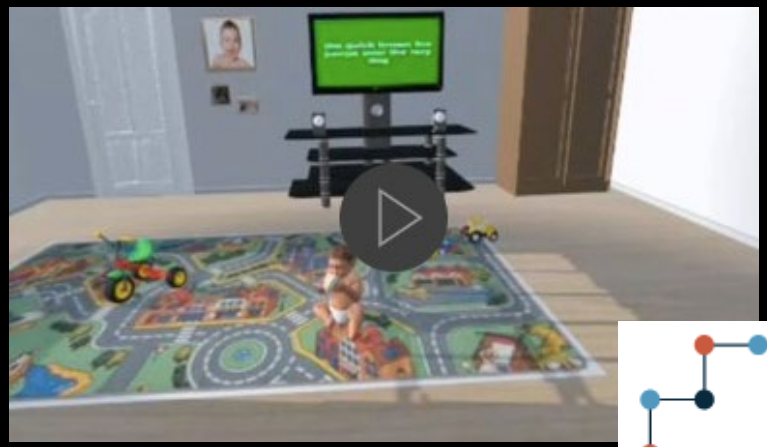
→ Hypothesis: varying the distance to mean of population within a speaker allows control of indexicality

Integration of identity marking in communication

Do speakers change their identity marking in communication when identity is at stake?

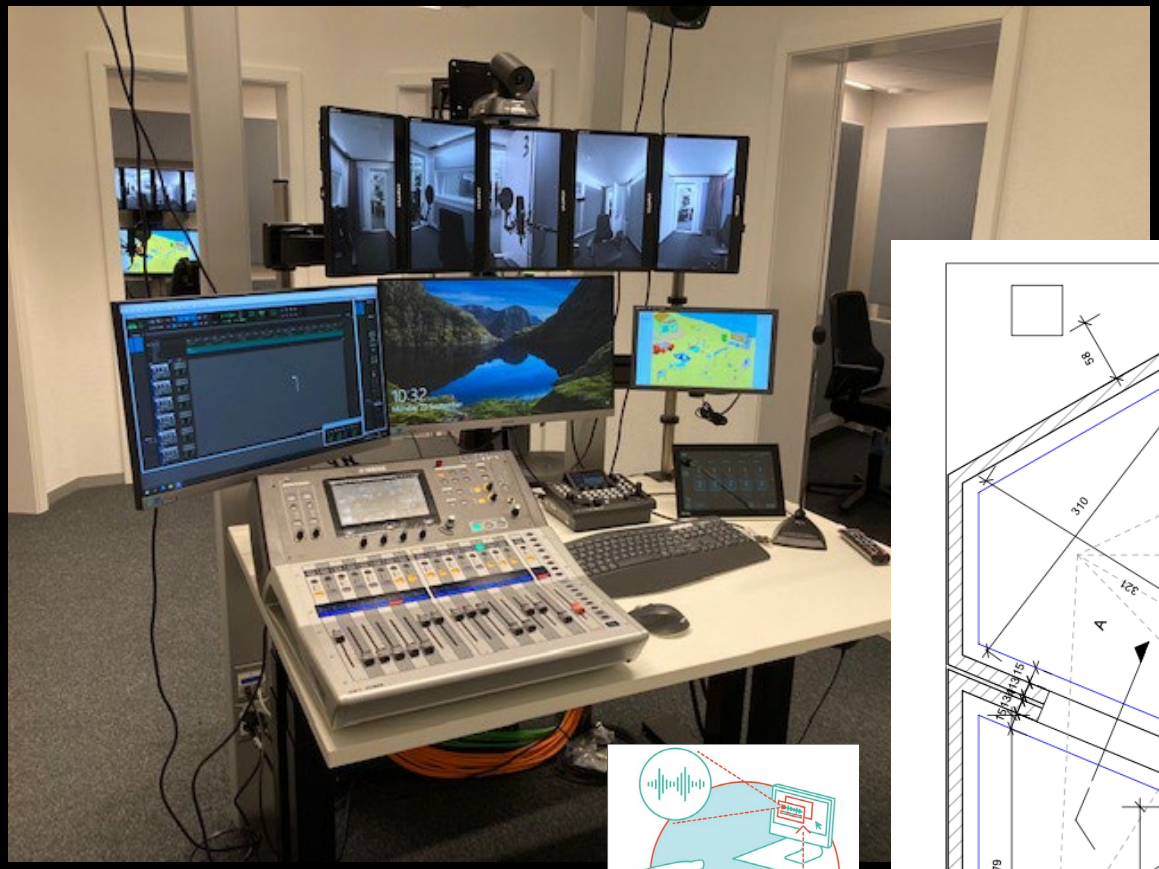
Does group-size play a role making voice more identifiable?

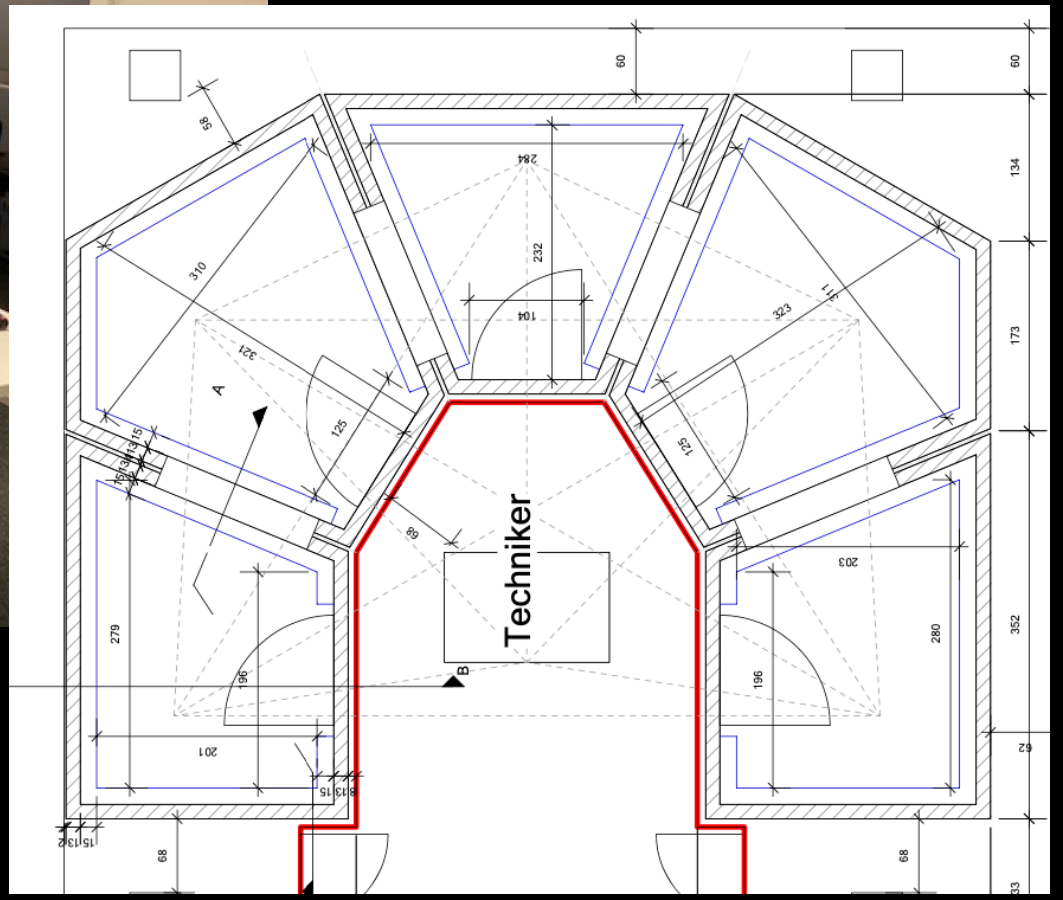
Eliciting speaking styles with VR

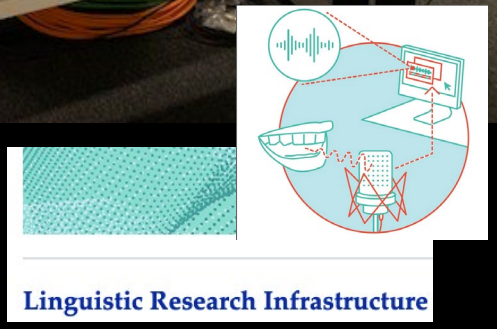


Swiss National Science Foundation

Controlled interaction lab:







Thank You!