Pinpointing the Difference – Visual Comparison of Non-Native Speaker Groups

Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany

Florian Hönig, Sebastian Wankerl, Anton Batliner, Elmar Nöth

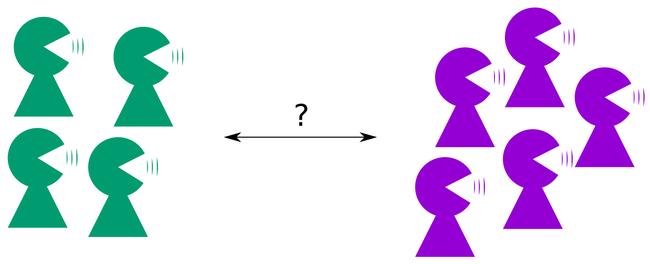
florian.hoenig@fau.de

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Introduction

How to characterize traits of speaker groups?



- Literature – possibly incomplete
- Manual inspection – time-consuming, subjective
- Data-driven – difficult to interpret

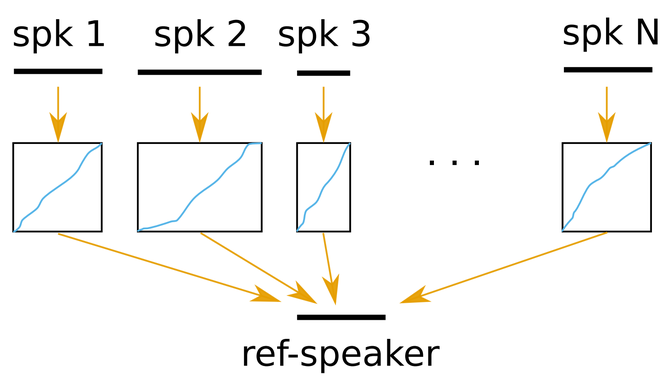
Visual Comparison of Speaker Groups (VICOS)

Collapse a whole corpus of recordings into a **single visualization**

- Generic – all kinds of speech
- Local – relatable to individual phonemes
- Restricted to realizations of the same word sequence
  $\rightarrow$ no repetitions, insertions and deletions
- Originally developed for pathological speech [1]

Method

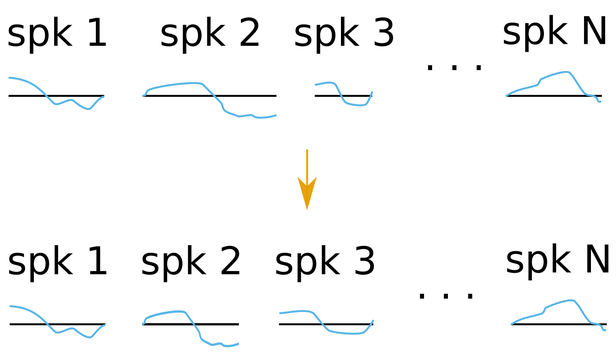
Common time basis



- Mapping of corresponding speech segments
- Penalized dynamic time warping [2, 3]
- MFCCs + deltas + delta-deltas
- Each dimension normalized to $\mu = 1, \sigma = 1$ $\rightarrow$ costs for insertions/deletions: 1

Mapping parameters of interest to fixed length

- 1-D case (loudness, tempo):



- 2-D case (spectrogram, mel spectrogram) analogously
- Group prototypes: **average**
- Group difference: **effect size**

$$d = (\mu_A - \mu_B)/\sqrt{(\sigma_A^2 + \sigma_B^2)/2}$$

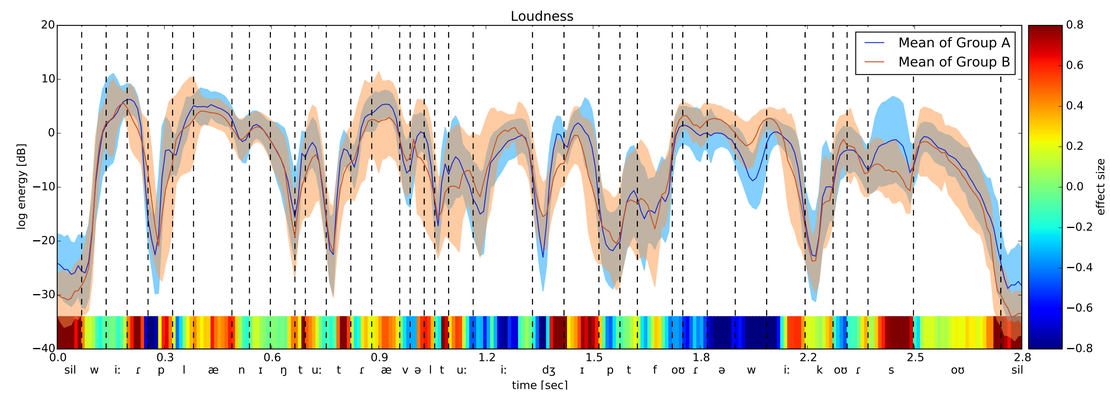
- Appropriate normalization, transformation, and smoothing
- Tempo: insertions/deletions

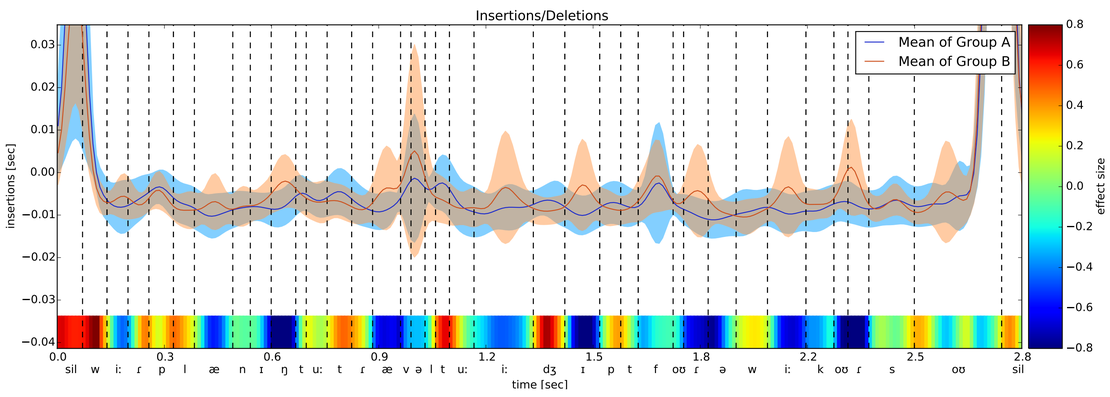
Experiments and Results

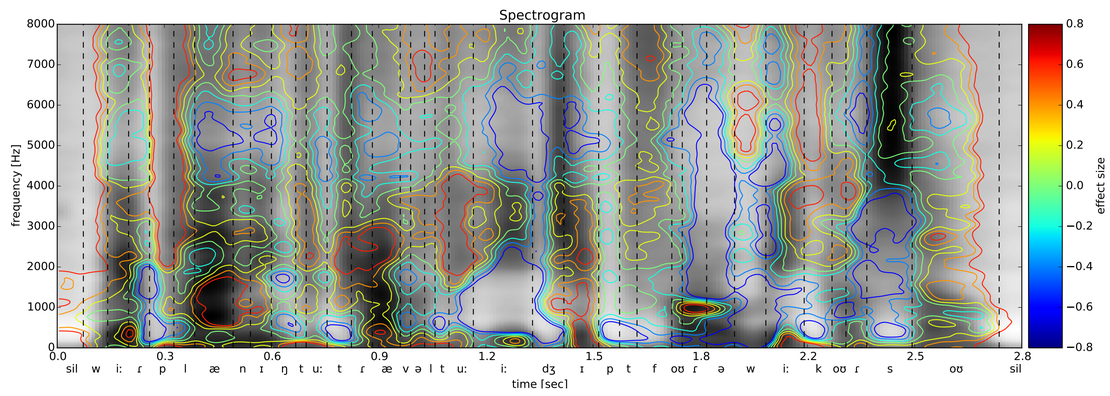
Database

- A sentence from the ISLE corpus [4]:
  We're planning to travel to Egypt for a while or so
- Removing reading errors $\rightarrow$
  19 German speakers (7f, 12m) – **Group A**
  22 Italian speakers (4f, 18m) – **Group B**

Results







Observed differences

- German speakers produce /t/ more articulate – steeper slope in loudness
- Syllable /iː/ in Egypt (word + phrase accent) louder in Italian – related to accents and/or phoneme substitution (see Spectrogram)
- German speakers: /s/ more sharp and loud – loudness + spectrogram

Conclusions

- VICOS suitable for non-native speech, too
- Rapid assessment of speaker group differences
- Generic
- Interpretability through locality
- Recent improvements: pitch, harmonicity, resynthesis feature
- Available as open-source python code at `www5.cs.fau.de/vicos`

References

[1] Sebastian Wankerl et al. "Visual Comparison of Speaker Groups". In: *INTERSPEECH 2015 (Show and Tell)*. to appear. 2015.

[2] Kim Roberts et al. "Enhancement and dynamic time warping of somatosensory evoked potential components applied to patients with multiple sclerosis". In: *Biomedical Engineering, IEEE Transactions on* 6 (1987), pp. 397–405.

[3] Haibin Sun, John CS Lui, and David KY Yau. "Distributed mechanism in detecting and defending against the low-rate TCP attack". In: *Computer Networks* 50.13 (2006), pp. 2312–2330.

[4] W. Menzel et al. "The ISLE corpus of non-native spoken English". In: *Proc. LREC*. Athens, 2000, pp. 957–964.