Visual Comparison of Speaker Groups (VICOS)

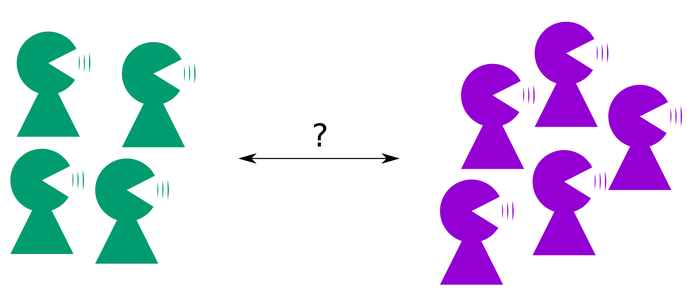
Sebastian Wankerl[1], Florian Hönig[1], Anton Batliner[1],
Juan Rafael Orozco-Arroyave[1,2], Elmar Nöth[1]

[1]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[2]Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia
{sebastian.wankerl, 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

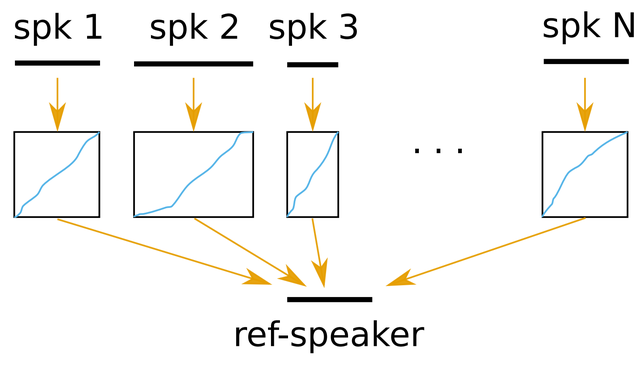
The VICOS approach

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

Method

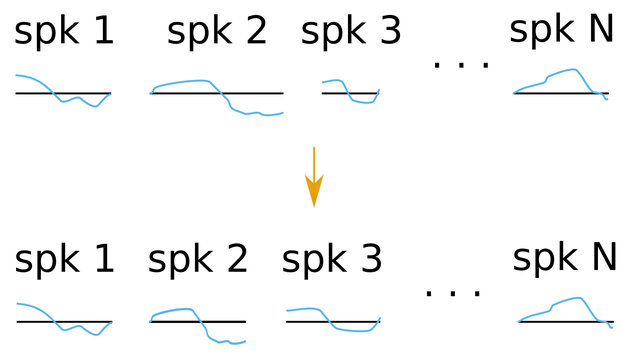
Common time basis



- Mapping of corresponding speech segments
- Penalized dynamic time warping [1, 2]
- 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

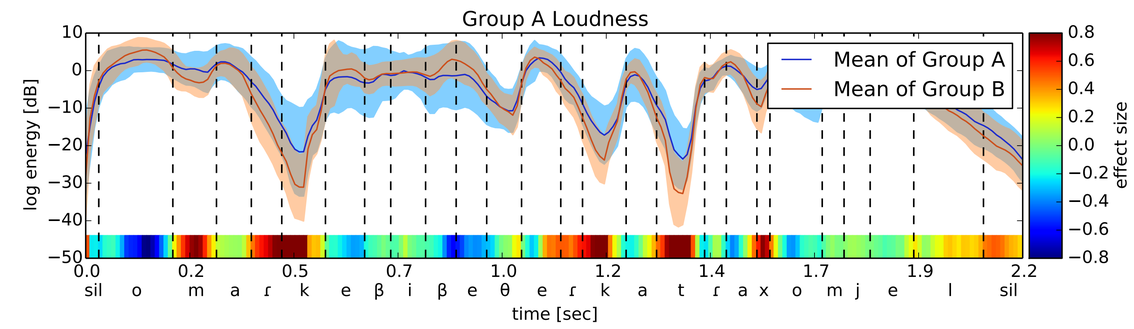
Database

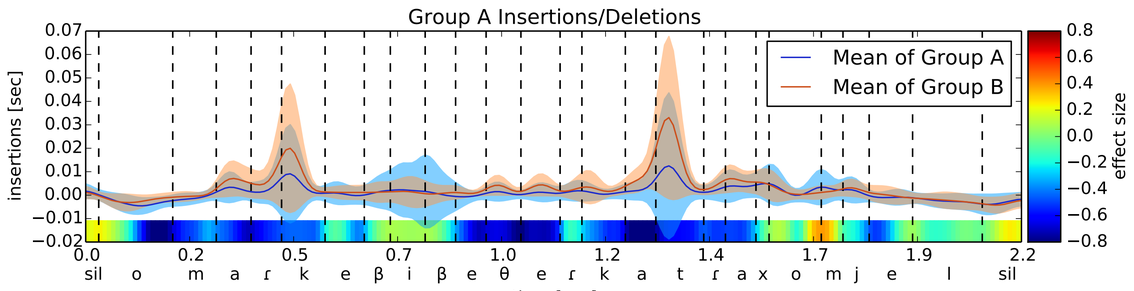
- Spanish speech from 50 speakers suffering from Parkinson's disease (PD) + 50 age-matched healthy control speakers [3].
- Selected sentence:

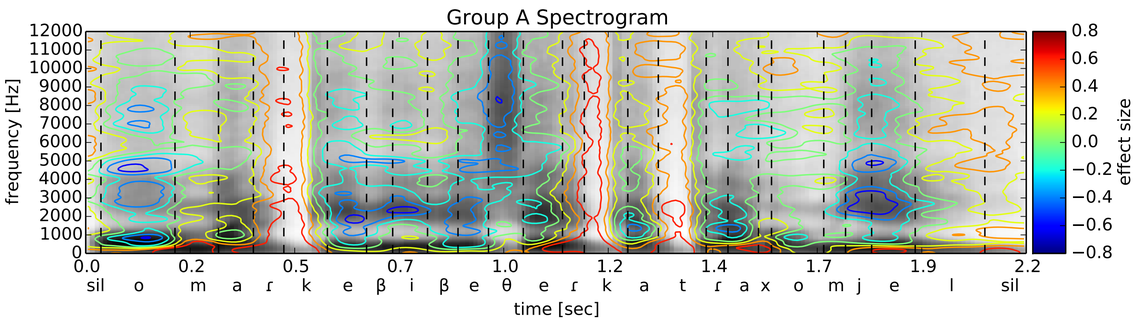
  Omar, que vive cerca, trajo miel
  (Omar, who lives nearby, brought honey)

- Removing reading errors $\rightarrow$
  42 PD speakers (21f, 21m) – **Group A**
  49 healthy controls (25f, 24m) – **Group B**

Results







Pertinent effects [4] observed

1. Rushed speech generally
2. Slowing down at phonemes difficult for PD speakers – /β i β e/, /ɾ k/, /om/
3. Less control of fricative – /θ/
4. Impaired control of velum: higher energy before closures – /k/
5. More nasality generally: more energy in the frequency band of nasal formants

Conclusions

- Rapid assessment of speaker group differences
- Generic
- Interpretability through locality
- Application to PD speech reproduced known characterizations
- Recent improvements: pitch, harmonicity, resynthesis feature
- Available as open-source python code at `www5.cs.fau.de/vicos`

References

[1] 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.

[2] 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.

[3] J.R. Orozco-Arroyave et al. "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease". In: *LREC*. 2014, pp. 342–347.

[4] Uwe Jürgens. "Neural pathways underlying vocal control". In: *Neuroscience & Biobehavioral Rev.* 26.2 (2002), pp. 235–258.