A comparison of acoustic features for articulatory inversion

Chao Qin and Miguel Á. Carreira-Perpiñán

Dept. of Computer Science & Electrical Engineering, OGI/OHSU

(from July 2007 at University of California, Merced)

http://www.csee.ogi.edu/~cqin

*Interspeech 2007, Antwerp*
Introduction

• Articulatory inversion, a.k.a acoustic-to-articulatory mapping
  – Recover sequences of vocal tract shapes from the acoustics
  – Multi-valued mappings or nonuniqueness

Still unsolved!

• Applications
  – Improve speech recognition, synthesis, and coding
  – Provide visual aid for language learning and therapy
Approaches to articulatory inversion

- **Analysis-by-synthesis** (Flanagan *et al* ’80, Levinson’83)
- **Neural networks** (Soquet *et al* ’91)
- **Codebook** (Atal *et al* ’79, Schroeter and Sondhi’88)
- **Emsemble neural networks** (Rahim’93)
- **Conditional modes** (Carreira-Perpinan’99)
  - Learn conditional density model
  - Derive *multi-valued mapping* from modes of conditional density
  - DP minimize the continuity constraint
- **Extended Kalman filtering** (Deng’98)
- **Particle filtering** (future work)

AC space  \(\rightarrow\) ARTIC space

Tracking

hidden state observation
tracking algorithm
Articulatory data

- MOCHA-TIMIT database (Wrench and HardCastle’00)
  - Simultaneous audio + pellet movements
Investigation of acoustic features

- **Jaggedness** of acoustic features makes it difficult to define mappings
Investigation of acoustic features

**Acoustic feat.**
- LPC
- LSF
- LPCC
- MFCC
- FBANK
- PLP
- RASTA-PLP

**Dynamics**
- static only
- static+dynamic

**Winsize**
- 25 ms
- 35 ms
- 45 ms
- 64 ms
- 80 ms
- 96 ms

**Stepsize**
- 10 ms

**Smoothing θ**
- \( \theta = 1 \)
- \( \theta = 0.5 \)
- \( \theta = 0.25 \)

Smoothing method: **filtfilt**
Experimental setup

• Dataset
  - One female speaker fsew0 from MOCHA
  - 10000 frames for training
  - 2000 frames for testing

• Silence removal by energy-based endpoint detection

• Inversion method
  - A multi-layer perceptron with a single layer of 55 hidden units

• Performance metric
  - RMS error: $\sqrt{E((\hat{x} - x)^2)}$
  - Correlation: $\frac{\text{cov}(\hat{x}, x)}{\sqrt{\text{var}(\hat{x}) \cdot \text{var}(x)}}$
Experimental results
Effect of time delay

- Alignment of acoustic and articulatory frames
- Empirical study to find out the best time delay

- Optimal time delay is around 15 ms
Conclusions

• Best acoustic parameterisations help but not significantly
  – LSF + dynamic features + 64~80 winsize + smoothing ($\theta = 0.25$)

• Time delay (15 ms) helps but very insignificantly

• Relatively large windows and smoothing were shown to alleviate jaggedness of acoustic features

• Limitations
  – Used data from one speaker
  – Did not study sounds separately
Acknowledgement

- MACP and CQIN thank Korin Richmond for valuable discussions
- MACP and CQIN thank A. Wrench and CSTR for MOCHA data
- Supported by NSF CAREER award IIS-0546857
Performance comparisons with cond. mean

![Graphs showing performance comparisons with different smoothing values.](image)
Performance comparisons with cond. modes