Making Conversational Vowels More Clear

Seyed Hamidreza Mohammadi
Alexander Kain
Jan van Santen

Center for Spoken Language Understanding

OREGON HEALTH & SCIENCE UNIVERSITY
OHSU

Background

- 28 million in U.S. have hearing loss
  - understanding speech in noise is difficult
  - primary strategy is hearing aid
    - frequency band amplification and shifting
    - dynamic processing
    - can we do more?

Clear speech

- speakers adopt special *clear* speech style (vs. *conversational*) when talking to impaired listeners
  - due to hearing loss, background noise, or both
- studies have shown the increased intelligibility of clear speech in many conditions [Picheny 1985, Ferguson 2002 & 2004]
  - speakers make CLR speech in different ways, and some can not produce it [Ferguson 1999]

CLR speech: prosody

- fundamental frequency with increased range and mean
- increased consonant-vowel energy ratio
- prolonged phoneme durations
- longer pauses
- decreased speaking rate

CLR speech: spectrum

- expanded vowel formant frequency space
- overall higher energy in the 1000-3150 Hz frequency band
- more aspirated stops and fewer alveolar flaps

Intelligibility features

- previously, we investigated the degree to which various acoustic features *cause* an increase in speech intelligibility [Kain 2008]
- *hybridized* CNV and CLR speech
  - energy trajectory, F0 trajectory, phoneme durations, short-term spectra, pauses
- perceptual intelligibility testing

Hybridization results

- For one male en_US speaker,
  - CNV = 72%
  - CLR = 85%
  - CNV with CLR spectrum and duration = 82%
    - energy trajectory, F0 trajectory, and pauses not relevant
    - duration alone was not statistically significant

Research question

- can we automatically increase the intelligibility of CNV vowels?
  - use a voice conversion method that maps CNV spectra towards CLR spectra
  - no duration modeling for now
  - speaker-dependent

Outline

- data corpus
- formant space analysis
- LSF space analysis
- perceptual experiment
- conclusion & future work

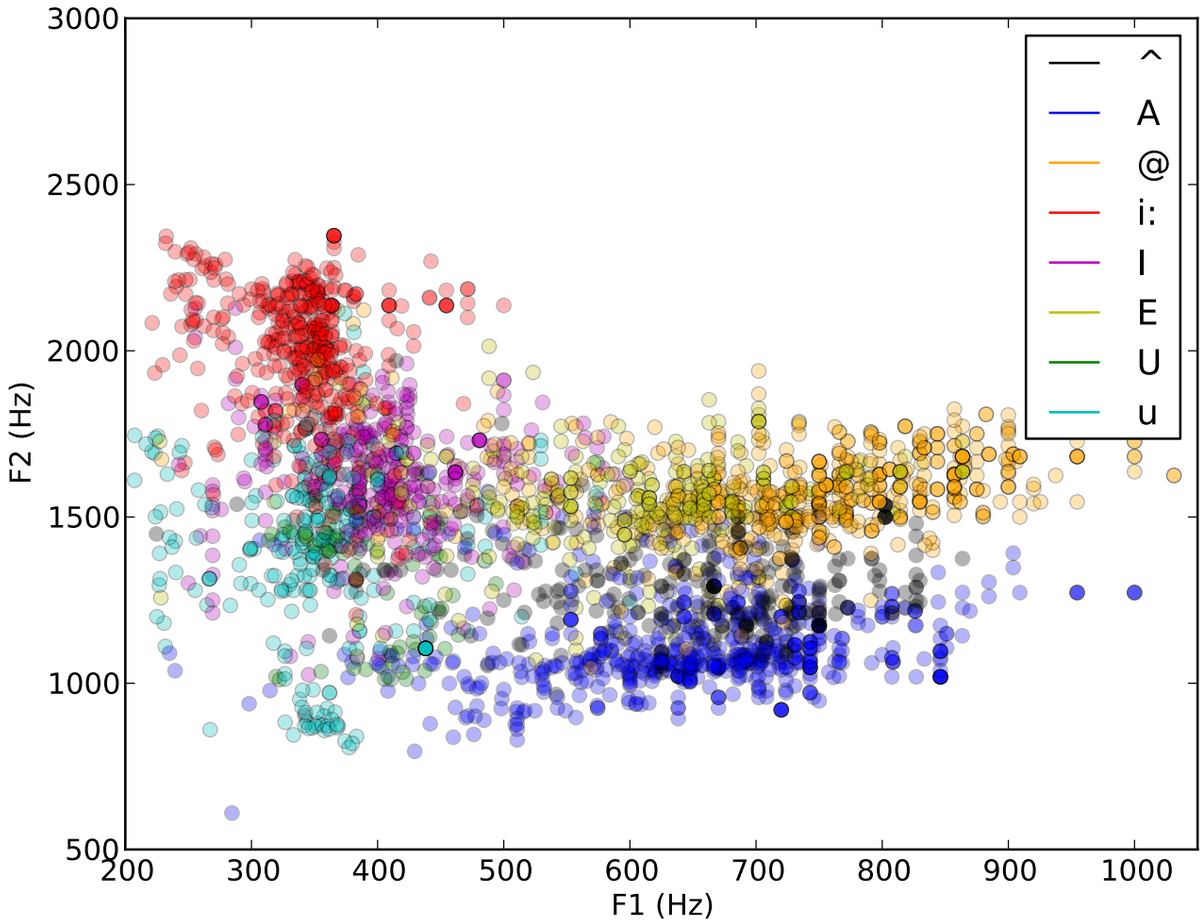
Corpus

- recorded a corpus of 242 en_US CVC words, uttered by one male speaker
  - each word recorded in CNV and CLR styles, and two renditions (242 x 2 x 2=968 tokens)
  - for CNV speech, the speaker was asked to speak as if talking with a friend at a natural pace
  - for CLR speech, the speaker was asked to "enunciate consonants more carefully and with greater effort than in CNV speech and avoid slurring words together"

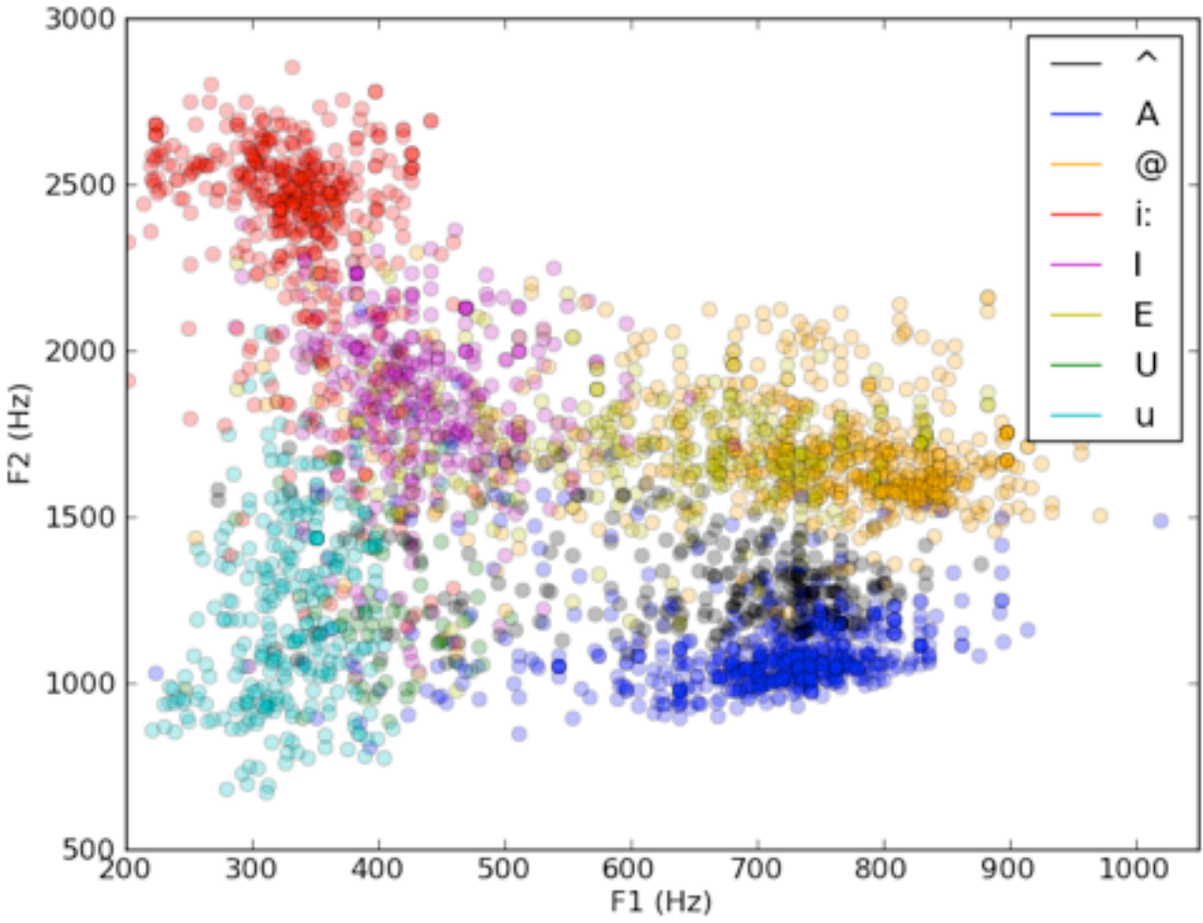
Corpus

- CVC words were located in a neutral carrier phrase immediately following the phoneme /d/:
  - "I know the meaning of the word **moon**"
- phonetically labeled by hand
- formants for the last word are extracted and hand corrected

CNV & CLR Formants

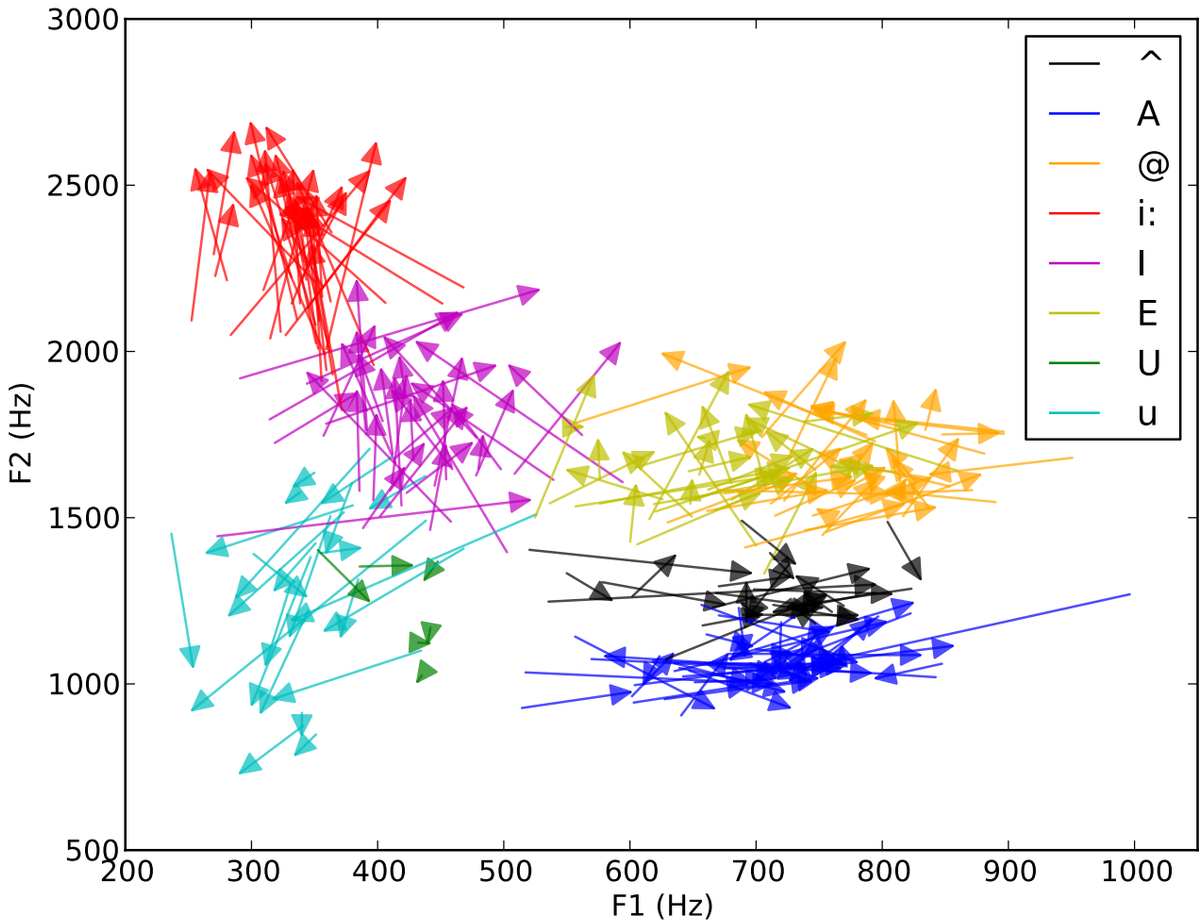


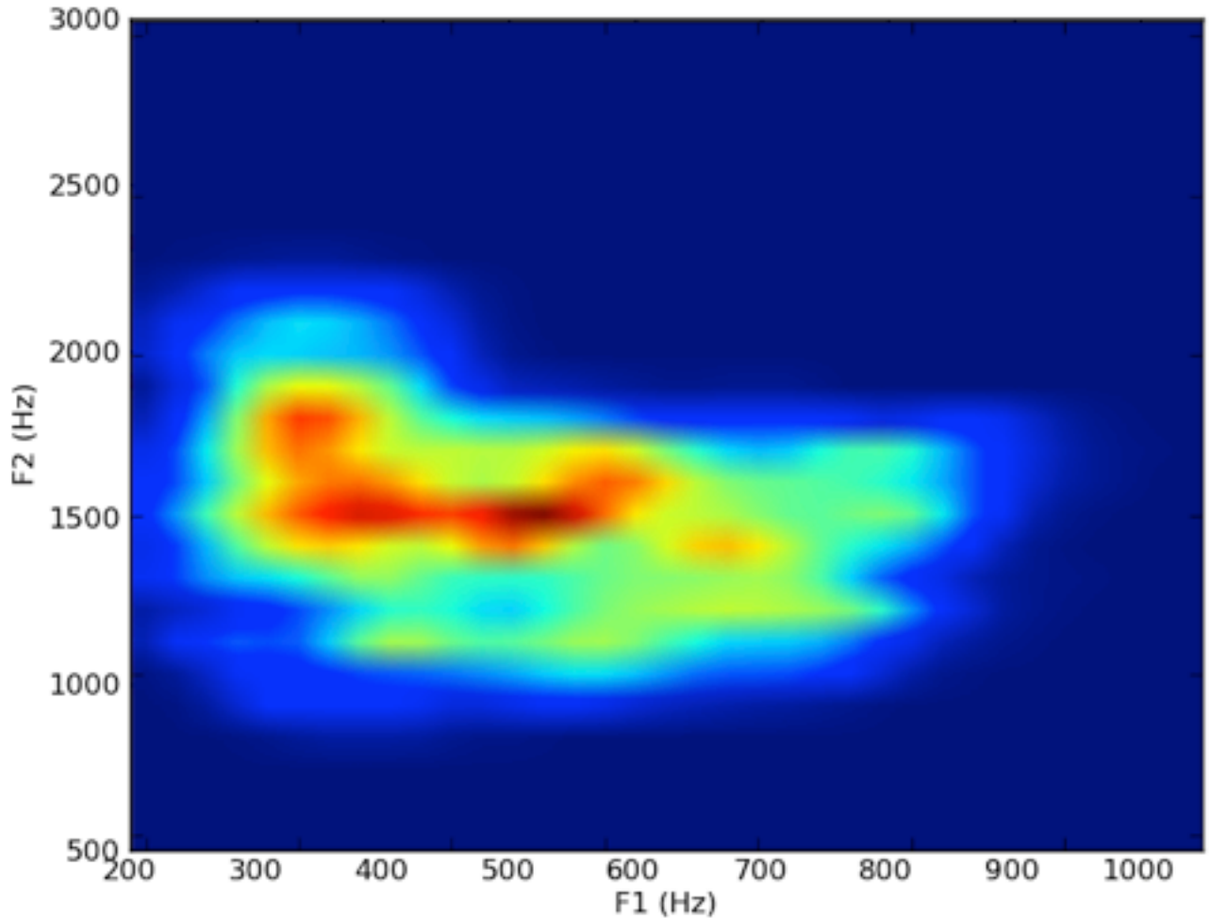
CNV



CLR

Mapping CNV to CLR





CNV to CLR

Consistency

Consistency measure

- determinant of the weighted covariance matrix (volume)

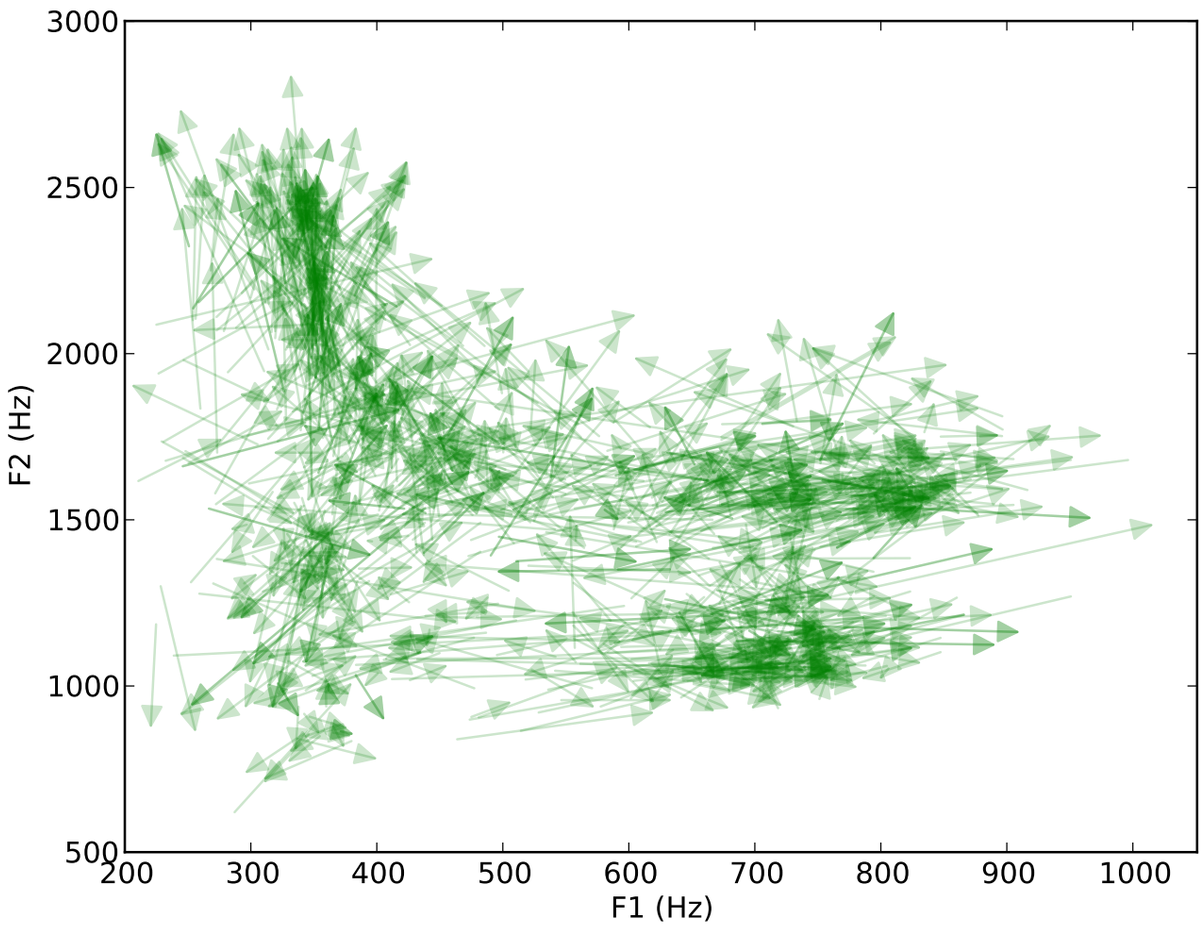
$$WeightedCov_{x'} = \frac{1}{\sum^{N} w_{i,x'}} \sum_{i=1} w_{i,x'}(x_i - \overline{x})(y_i - \overline{y})$$

$$w_{i,x'} = \exp(\frac{-\|x_i - x'\|^2}{2\sigma^2})$$

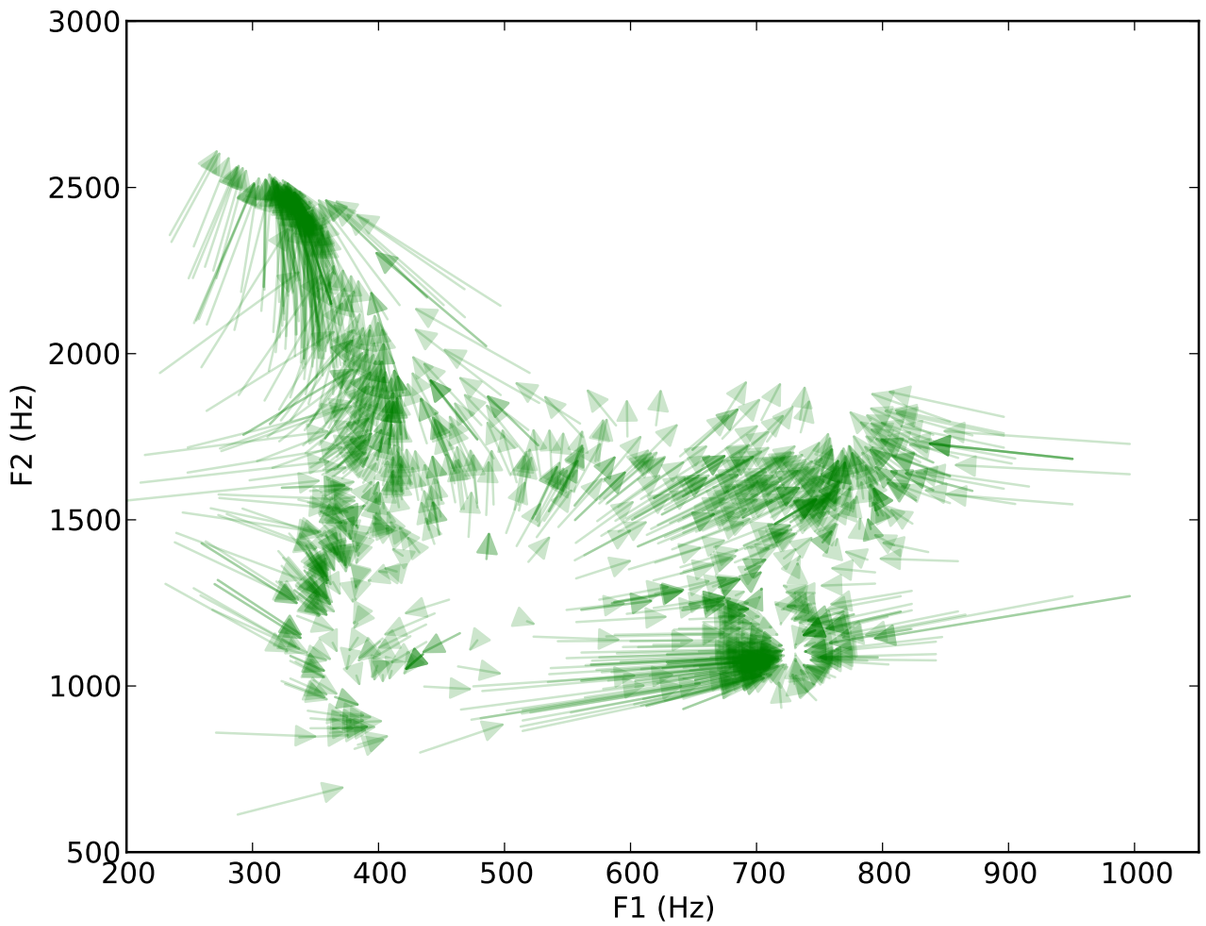
Mapping

- frame-by-frame mapping of CNV F1-F2 formant feature sequences of vowel trajectories, using a joint-density Gaussian mixture model (JDGMM) [Kain 1998]
- divided all available data into training (218 vowels) and test (24 vowels) sets using a 10-fold cross-validation scheme

Mapping Function



Actual Mapping



JDGMM Mapping, Q=8

Experiment

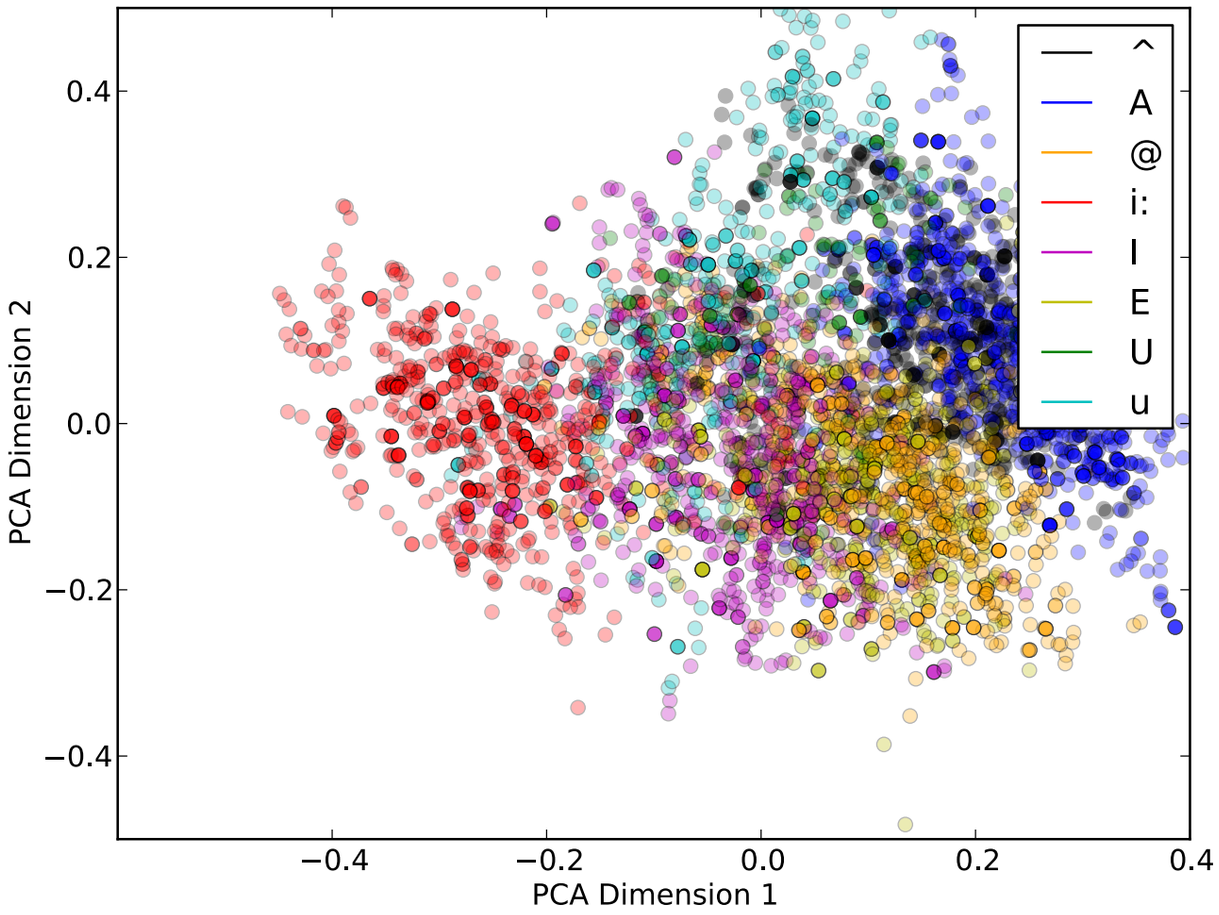
Table 1: Average RMSE errors in Hz

Method	test	train
no mapping	193.84	194.01
JDGMM Q=8, unsupervised	137.98	127.70
JDGMM Q=8, supervised	132.44	125.17

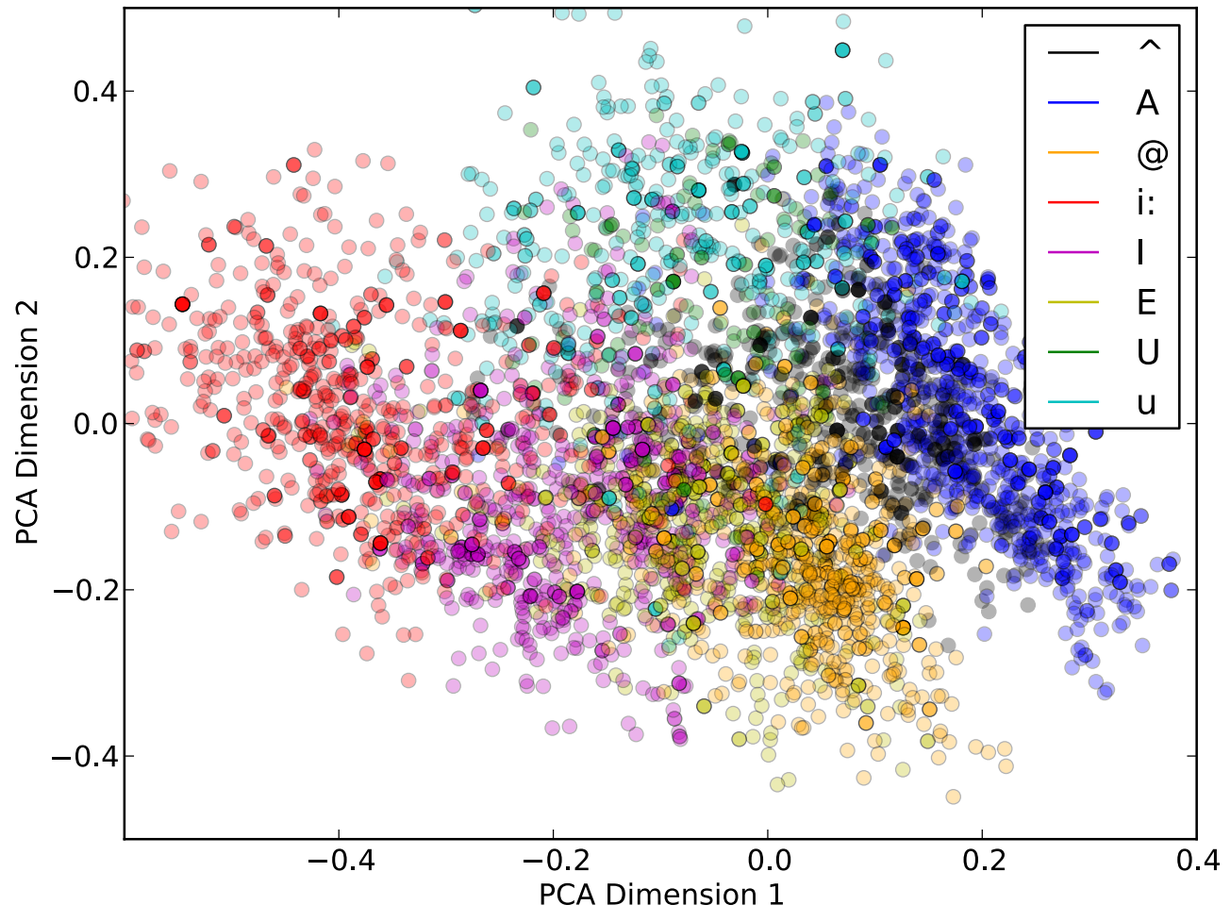
LSF domain

- repeated analysis for Line Spectral Frequencies
  - 18 LSF coefficients
- for visualization purposes ONLY, LSFs are reduced to 2 dimension using PCA

CNV & CLR LSFs

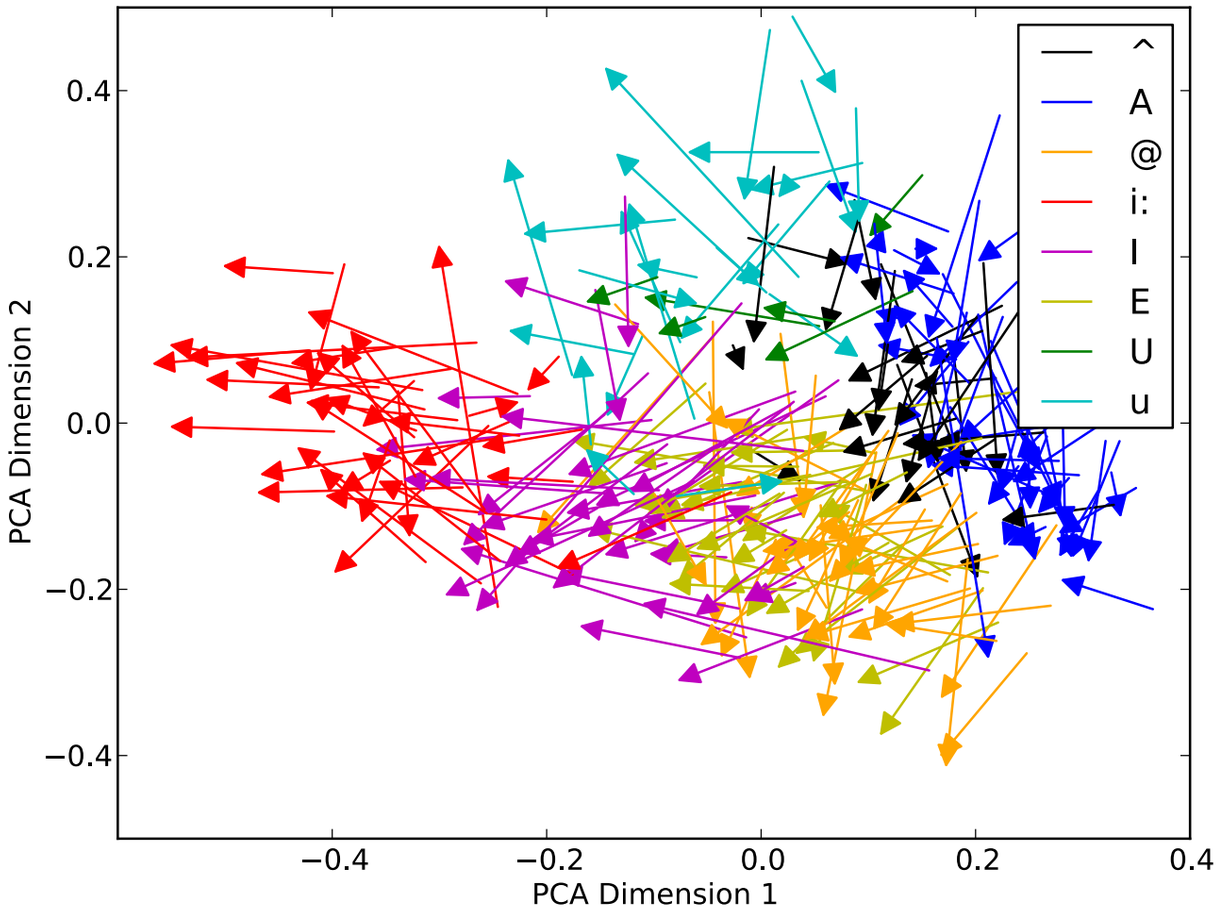


CNV

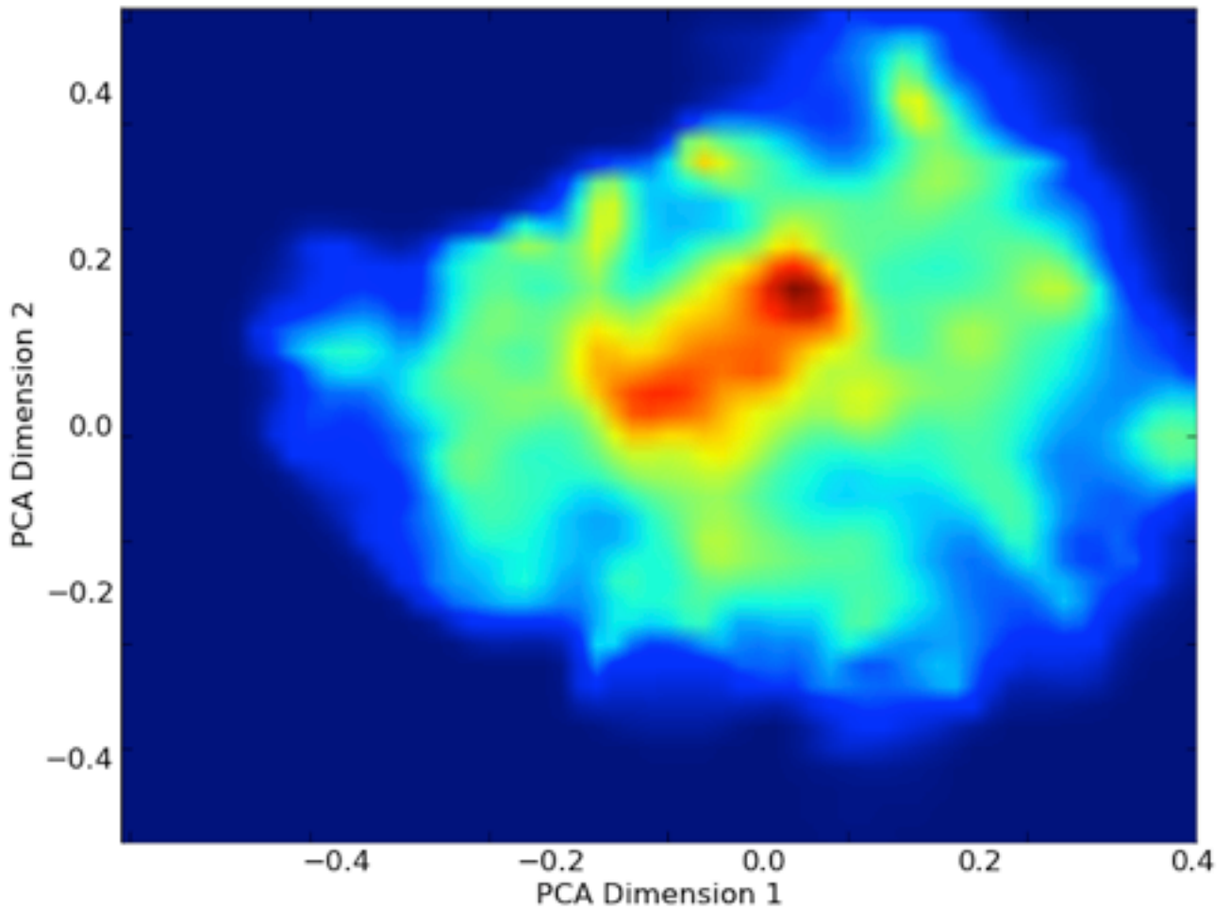


CLR

Mapping CNV to CLR

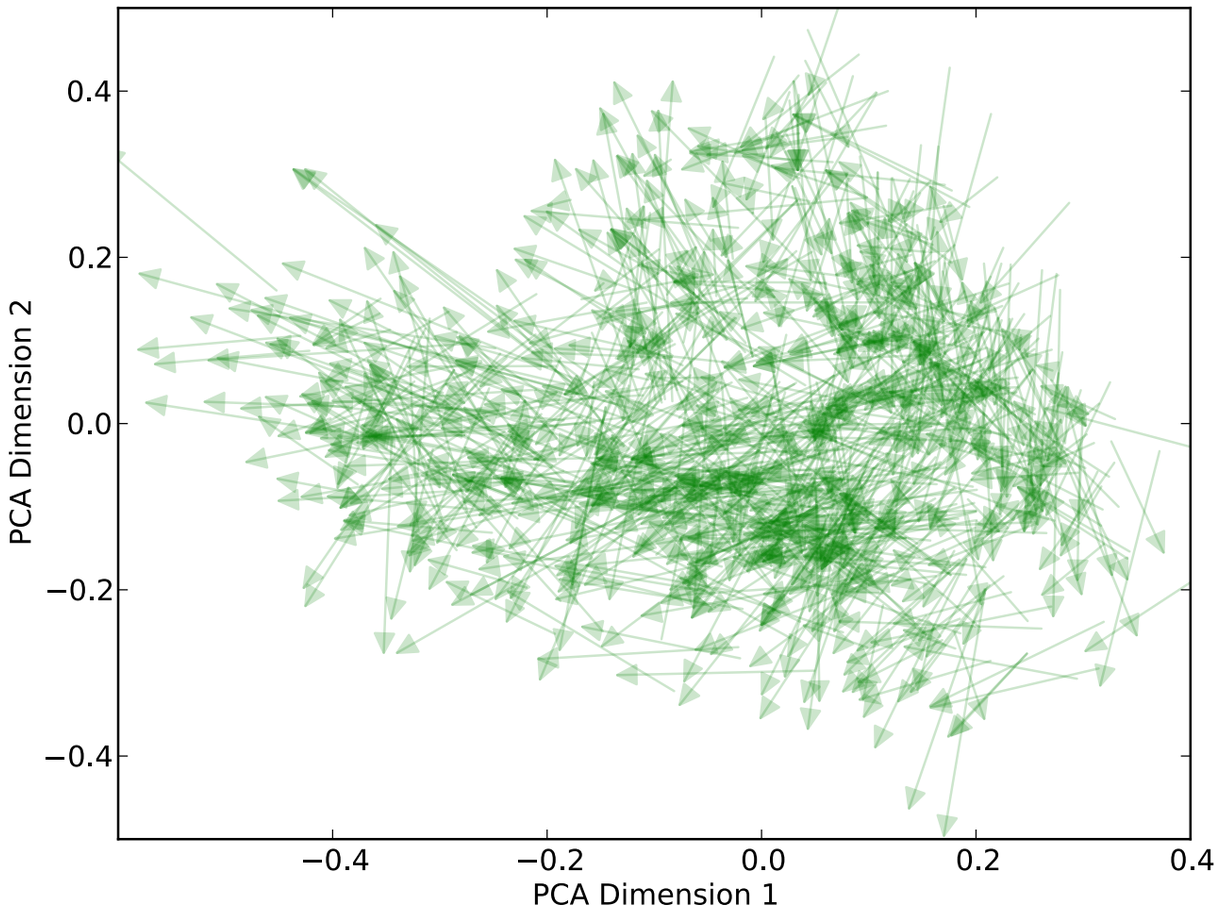


CNV to CLR

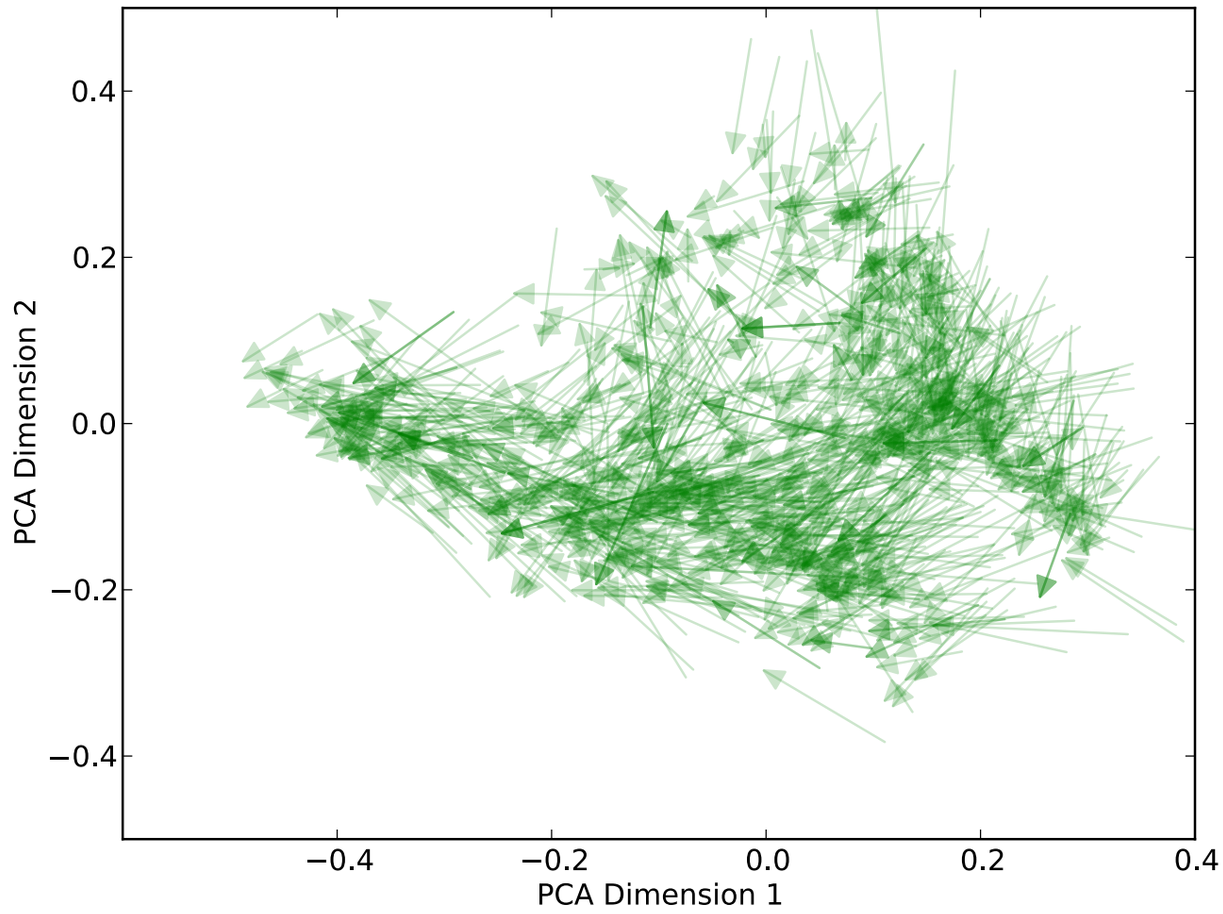


Consistency

Mapping



Actual Mapping



JDGMM Mapping, Q=3

Experiments

Table 2: Average LSD errors in dB

Method	test	train
no mapping	9.60	9.59
JDGMM Q=3	5.10	4.53
JDGMM Q=3, PCA=12	5.18	4.84

Perceptual experiment

- seven conditions
  1: CNV, 2: CLR,
  3: LSF-vocoded CNV (VCNV),
  4: LSF-vocoded CLR (VCLR),
  5: CNV with mapped spectrum (MAP-S),
  6: CNV with CLR "oracle" duration (MAP-D), and
  7: CNV with mapped spectrum and CLR "oracle" duration (MAP-SD)

- total of 49 test words x 7 conditions = 343 stimuli
  - loudness-normalized via rmsA
  - added 12-talker babble noise with SNR of +3 dB and -2 dB

Listeners

- 98 Amazon Mechanical Turk Listeners
  - approval ratings of at least 90% and located in the U.S.
- asked participant to "listen to the word in noise and select one of the vowel classes based on what you heard"
  - clean reference samples were available at any time

Intelligibility results

Table 3: Intelligibility rate of each condition

Configuration	−2 dB SNR	+3 dB SNR
CLR	74.92%	80.46%
VCLR	71.42%	78.71%
MAP-SD	56.26%*	58.60%†
MAP-S	49.85%	59.76%†
MAP-D	48.10%	56.26%
VCNV	45.18%	52.47%
CNV	45.48%	56.55%

Conclusion

- analyzed CNV and CLR data in formant frequency domain and LSF domain
- trained a mapping function for converting CNV vowels to resemble CLR vowels
- results show a modest increase in intelligibility

Future work

- use a duration model
- this experiment is for one speaker only – can the mapping be speaker-independent?
- use a higher quality vocoder and a more sophisticated mapping method

Thank you!