MODELING INSTANTANEOUS INTONATION FOR SPEAKER IDENTIFICATION
USING THE FUNDAMENTAL FREQUENCY VARIATION SPECTRUM

Kornel Laskowski and Qin Jin
interACT, Carnegie Mellon University, Pittsburgh PA, USA
Cognitive Systems Lab, Universität Karlsruhe, Karlsruhe, Germany

Goal

Improve speaker identification by modeling **intonation bias**
(the distribution of speaker-preferred pitch change in octaves per 8 ms).

HOW?

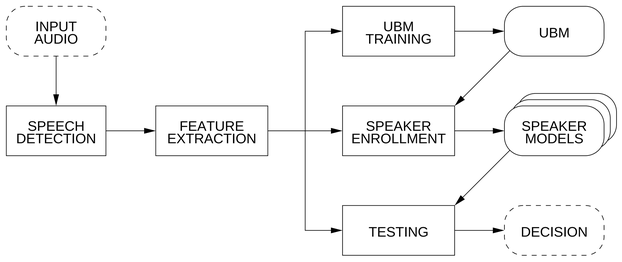
Compute the fundamental frequency variation (FFV) spectrum
and model alongside "traditional" spectral (MFCC) features, ie.

$$\log \boldsymbol{P}\left(\textsc{Mfcc13} \oplus \textsc{Ffv7} | \mathcal{M}\right) = \log \boldsymbol{P}\left(\textsc{Mfcc13}, \textsc{Ffv7} | \mathcal{M}_{\boldsymbol{stacked}}\right)$$

or

$$\log \boldsymbol{P}\left(\textsc{Mfcc13} \otimes \textsc{Ffv7} | \mathcal{M}\right) = \lambda_{\textsc{Mfcc13}} \log \boldsymbol{P}\left(\textsc{Mfcc13} | \mathcal{M}_{\textsc{Mfcc13}}\right) + \lambda_{\textsc{Ffv7}} \log \boldsymbol{P}\left(\textsc{Ffv7} | \mathcal{M}_{\textsc{Ffv7}}\right)$$

Baseline GMM-MFCC SID System



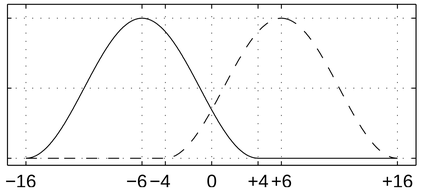
- 32 ms frames every 8 ms, low-energy frames excluded
- $\textsc{Mfcc13} \equiv \{\boldsymbol{MFCC_0}, \boldsymbol{MFCC_1}, \cdots, \boldsymbol{MFCC_{12}}\}$
- cepstral mean subtraction
- Gaussian mixture models (4096)
- ML estimation of universal background model (UBM)
- MAP estimation of speaker GMM means

Data

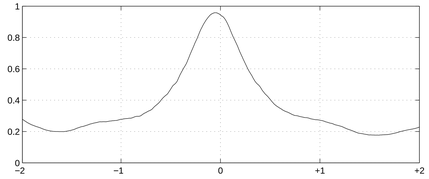
- Wall Street Journal, LDC CSR-I (WSJ0) & LDC CSR-II (WSJ1)
- read sentences and some spontaneously produced utterances
- 16 kHz wideband audio (Sennheiser HMD414)
- 102 female speakers & 95 male speakers
  TRAINSET: 5 minutes of speech per speaker
  TESTSET: 3 trials $\times$ 1 minute of speech per speaker
- UBMSET: remaining speakers' speech (70 hours)

What Is the Fundamental Frequency Variation Spectrum?

- dot product of two spectra, $\mathbf{F}_L$ and $\mathbf{F}_R$ ...



- ... after dilating one of $\mathbf{F}_L$ and $\mathbf{F}_R$ by factor $\mathbf{2}^{\rho}$
- FFV spectrum is value of dot product over range of ρ



- after 7-filter filterbank, $\textsc{Ffv7} \equiv \{\boldsymbol{FFV_{-3}}, \cdots, \boldsymbol{FFV_{+3}}\}$

Findings

1. Raw FFV7 features offer better than chance performance.
2. Sphering FFV7 features dramatically improves performance.
3. Model-space combination of MFCC13 with sphered FFV7 yields 54% and 40% relative error reductions for female and male speakers, respectively.
4. Model-space combinations of MFCC13 with FFV7 relative to that with MFCC7 yields 24–35% and 26–30% relative error reductions for female and male speakers, respectively.
5. Feature-space combinations of MFCC13 with FFV7 offer no benefit.

Experiments (identification accuracies, %)

System	Female	Male
MFCC13	82.0	92.3
MFCC7	43.8	68.8
FFV7	27.8	45.3
PCA(MFCC13)	84.3	91.2
PCA(MFCC7)	44.4	65.2
PCA(FFV7)	**62.7**	64.2
MFCC13 $\otimes$ MFCC7	86.3	92.6
MFCC13 $\otimes$ FFV7	80.7	92.3
PCA(MFCC13) $\otimes$ MFCC7	87.3	92.3
PCA(MFCC13) $\otimes$ FFV7	84.6	**92.6**
MFCC13 $\otimes$ PCA(MFCC7)	89.2	93.7
MFCC13 $\otimes$ PCA(FFV7)	**91.8**	**95.4**
PCA(MFCC13) $\otimes$ PCA(MFCC7)	86.9	93.0
PCA(MFCC13) $\otimes$ PCA(FFV7)	**91.5**	**95.1**

(kornel|qjin)@cs.cmu.edu