A General-Purpose 32 ms Prosodic Vector for Hidden Markov Modeling

Kornel Laskowski$^{1,2}$, Mattias Heldner$^3$ & Jens Edlund$^3$

$^1$Carnegie Mellon University, Pittsburgh PA, USA
$^2$Universität Karlsruhe, Karlsruhe, Germany
$^3$KTH — Royal Institute of Technology, Stockholm, Sweden

8 September, 2008
Imagine you had ...

- a local representation of tone
  - estimated from a **single** ASR-size analysis frame
- which would **not** require:
  - prior determination of voicing
  - speaker normalization

- with separable codeword clusters for
  - absence of voicing
  - presence of voicing, constant $F_0$
  - presence of voicing, falling $F_0$, with rate of change
  - presence of voicing, rising $F_0$, with rate of change
Then you could do lots of things cheaply ...

Examples include:

- online prosodic modeling
- improved ASR for tonal languages
- enriched ASR for other languages
- contrastive phone models
- variously accented same-word lexicon entries
- (word-conditioned) prosodic phrasing for free
Instead, currently you need to ...

1. **run a pitch tracker**, which
   1. computes a local estimate of voicing and of pitch
   2. applies dynamic programming over a long observation time
2. **heuristically correct its output**, by
   1. pruning outliers, based on long-observation-time trends, and/or
   2. applying a piecewise linear approximation
3. **normalize for the speaker**, by
   1. determining a long-observation-time speaker norm
   2. applying the normalization to each frame
4. **treat unvoiced regions** by
   - interpolating inside them, or
   - posting exceptions in downstream modeling/handling
5. **compute a first-order log-difference**
What we will present ...

1. Fundamental Frequency Variation (FFV)
2. Applicability of the FFV Representation
   - speaker change prediction
   - speaker classification
   - dialog act classification
3. Several Basic Questions
   - feature transformation
   - feature regularization
   - concatenation with other features
   - runtime improvements
   - acoustic model complexity
4. Summary
Computation, for Each (32 ms) Analysis Frame

- estimate the FFV spectrum $g[\rho]$
  - estimate the power spectra $F_L$ and $F_R$
  - dilate $F_R$ by a factor $2^\rho$, $\rho > 0$
  - dot product with undilated $F_L$
  - repeat for a continuum of $\rho$ values

- pass $g(\rho)$ through a filterbank to yield $G \in \mathbb{R}^7$
- decorrelate $G$
Computation, for Each (32 ms) Analysis Frame

- estimate the FFV spectrum $g[\rho]$
  - estimate the power spectra $F_L$ and $F_R$
  - dilate $F_R$ by a factor $2^{\rho}$, $\rho > 0$
  - dot product with undilated $F_L$
  - repeat for a continuum of $\rho$ values

- pass $g(\rho)$ through a filterbank to yield $G \in \mathbb{R}^7$
- decorrelate $G$
Computation, for Each (32 ms) Analysis Frame

- estimate the FFV spectrum \( g[\rho] \)
  - estimate the power spectra \( F_L \) and \( F_R \)
  - dilate \( F_R \) by a factor \( 2^\rho, \rho > 0 \)
  - dot product with undilated \( F_L \)
  - repeat for a continuum of \( \rho \) values

- pass \( g(\rho) \) through a filterbank to yield \( G \in \mathbb{R}^7 \)
- decorrelate \( G \)
Computation, for Each (32 ms) Analysis Frame

- estimate the FFV spectrum $g[\rho]$
  - estimate the power spectra $F_L$ and $F_R$
  - dilate $F_R$ by a factor $2^\rho$, $\rho > 0$
  - dot product with undilated $F_L$
  - repeat for a continuum of $\rho$ values

- pass $g(\rho)$ through a filterbank to yield $G \in \mathbb{R}^7$
- decorrelate $G$
Computation, for Each (32 ms) Analysis Frame

- estimate the FFV spectrum \( g[\rho] \)
  - estimate the power spectra \( F_L \) and \( F_R \)
  - dilate \( F_R \) by a factor \( 2^\rho, \rho > 0 \)
  - dot product with undilated \( F_L \)
  - repeat for a continuum of \( \rho \) values

- pass \( g(\rho) \) through a filterbank to yield \( G \in \mathbb{R}^7 \)
- decorrelate \( G \)
Computation, for Each (32 ms) Analysis Frame

- estimate the FFV spectrum $g[\rho]$
  - estimate the power spectra $F_L$ and $F_R$
  - dilate $F_R$ by a factor $2^{\rho}$, $\rho > 0$
  - dot product with undilated $F_L$
  - repeat for a continuum of $\rho$ values

- pass $g(\rho)$ through a filterbank to yield $G \in \mathbb{R}^7$
- decorrelate $G$
estimate the FFV spectrum $g[\rho]$
- estimate the power spectra $F_L$ and $F_R$
- dilate $F_R$ by a factor $2^\rho$, $\rho > 0$
- dot product with undilated $F_L$
- repeat for a continuum of $\rho$ values

pass $g(\rho)$ through a filterbank to yield $G \in \mathbb{R}^7$
- decorrelate $G$
Computation, for Each (32 ms) Analysis Frame

- estimate the FFV spectrum $g[\rho]$
  - estimate the power spectra $F_L$ and $F_R$
  - dilate $F_R$ by a factor $2^\rho$, $\rho > 0$
  - dot product with undilated $F_L$
  - repeat for a continuum of $\rho$ values

- pass $g(\rho)$ through a filterbank to yield $G \in \mathbb{R}^7$
- decorrelate $G$
Computation, for Each (32 ms) Analysis Frame

- estimate the FFV spectrum $g[\rho]$
  - estimate the power spectra $F_L$ and $F_R$
  - dilate $F_R$ by a factor $2^\rho$, $\rho > 0$
  - dot product with undilated $F_L$
  - repeat for a continuum of $\rho$ values

- pass $g(\rho)$ through a filterbank to yield $G \in \mathbb{R}^7$
- decorrelate $G$
Computation, for Each (32 ms) Analysis Frame

- estimate the FFV spectrum $g[\rho]$
  - estimate the power spectra $F_L$ and $F_R$
  - dilate $F_R$ by a factor $2^\rho$, $\rho > 0$
  - dot product with undilated $F_L$
  - repeat for a continuum of $\rho$ values

- pass $g(\rho)$ through a filterbank to yield $G \in \mathbb{R}^7$
- decorrelate $G$
Computation, for Each (32 ms) Analysis Frame

- estimate the FFV spectrum $g[\rho]
- estimate the power spectra $F_L$ and $F_R$
- dilate $F_R$ by a factor $2^\rho$, $\rho > 0$
- dot product with undilated $F_L$
- repeat for a continuum of $\rho$ values

- pass $g(\rho)$ through a filterbank to yield $G \in \mathbb{R}^7$
- decorrelate $G
Computation, for Each (32 ms) Analysis Frame

- estimate the FFV spectrum $g[\rho]$
  - estimate the power spectra $F_L$ and $F_R$
  - dilate $F_R$ by a factor $2^\rho$, $\rho > 0$
  - dot product with undilated $F_L$
  - repeat for a continuum of $\rho$ values

- pass $g(\rho)$ through a filterbank to yield $G \in \mathbb{R}^7$
- decorrelate $G$
Comparison with MFCC Computation

**MFCC features**

1. **Audio Pre-emphasis**
2. **Power Spectrum Estimation**
3. **Filterbank (MEL)**
4. **Decorrelate (Inv. Cos-II)**
5. **Modeling**

**FFV features**

1. **Audio Pre-emphasis**
2. **FFV Spectrum Estimation**
3. **Perceptual Filterbank**
4. **Decorrelate (KLT)**
5. **Modeling**
FFV versus Pitch Tracking, Conceptually

Formant Tracking

→

Pitch Tracking

→

FFV Peak Tracking

→

FFT Spectrum

→

Autocorr Spectrum

→

FFV Spectrum

K. Laskowski, M. Heldner & J. Edlund
Interspeech 2009, Brighton, UK
Related Work

- well-established pitch processing & modeling techniques
  - e.g., Shriberg & Stolcke, “Direct modeling of prosody: An overview of applications in automatic speech processing”, *Speech Prosody 2004*.

- similar in purpose to $\Delta \log F_0$ from *cepstra*

- algorithmically similar, for different task
FFV for Speaker Change Prediction (ICASSP 2008)

- **CONTEXT:** Swedish Map Task (GIVER and FOLLOWER)
- **AUDIO:** 16kHz, close-talk, anechoic-chamber
- **GIVEN:** 500ms of speech at end-of-talkspurt, by GIVER
- **TASK:** predict whether GIVER will be next to speak

- **FINDINGS:**
  1. expected: GIVER appears to employ flat pitch to hold floor
  2. significantly outperm state-of-the-art pause-only system
  3. ML classifier outperforms a manually constructed state-of-the-art decision tree, which additionally uses pitch range information
FFV for Speaker Classification (ICASSP 2009)

- **CONTEXT:** read WSJ + some spontaneous utterances
- **AUDIO:** 16kHz, close-talk
- **GIVEN:** 1 minute interval of speech
- **TASK:** classify which of 100 people is the speaker

**FINDINGS:**
1. modeling speaker-dependent “intonation contour bias”
2. model-space combination with MFCC features reduces error rates by > 40%rel
FFV for Floor Mechanism Classification (EUSIPCO 2009)

- **CONTEXT**: naturally occurring meetings, American English
- **AUDIO**: 16kHz, close-talk, lots of crosstalk
- **GIVEN**: manually determined 500ms at beg-/end-of-talkspurt
- **TASK**: classify whether a floor mechanism or other DA type

**FINDINGS:**
1. flat pitch at the beginning of floor mechanism talk
2. slower speech at the end of floor mechanism talk
Basic but Relevant Questions

1. What are the effects of feature transformation?
2. What are the effects of feature regularization?
3. What are the effects of feature combination?
4. Can we reduce the computation time?
5. What are the effects of higher model complexity?

Will try to answer using classification accuracy and ROC area

- HMM log-likelihood-ratio classifier, 4 states, 1 Gaussian each
- task: floor DA type versus other DA type, balanced
- averaged over 4 individual context subtasks
Feature Transformation

- compare raw features with
  - global-Z-transformed features
  - global-PCA-rotated features

<table>
<thead>
<tr>
<th>System</th>
<th>Acc</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Baseline</td>
<td>64.8</td>
<td>71.5</td>
</tr>
<tr>
<td>2a Z-Transform</td>
<td>65.7</td>
<td>73.8</td>
</tr>
<tr>
<td>2b PCA Rotation</td>
<td>67.8</td>
<td>74.8</td>
</tr>
</tbody>
</table>

- feature transformation can be beneficial
- an optimal transform may be task- and audio- dependent
Feature Regularization

- replace response of 5 center filters with best parabolic fit

<table>
<thead>
<tr>
<th>System</th>
<th>Acc</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2b New Baseline</td>
<td>67.8</td>
<td>74.8</td>
</tr>
<tr>
<td>3 Quadratic Fit</td>
<td>68.9</td>
<td>76.7</td>
</tr>
</tbody>
</table>

- the parabolic projection and application of the filterbank are linear operations
- tantamount to applying a modified filterbank
- an optimal filterbank may be task- and audio- dependent
Feature Combination

- concatenate with 5 “auxiliary” correlates of prosodic features:
  - log energy (loudness)
  - delta log energy (change in loudness)
  - normalized autocorrelation maximum (probability of voicing)
  - Mel-spectral flux (speaking rate)
  - log-Mel-spectral flux (speaking rate)

<table>
<thead>
<tr>
<th>System</th>
<th>Acc</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>68.9</td>
<td>76.7</td>
</tr>
<tr>
<td>4a</td>
<td>64.8</td>
<td>71.3</td>
</tr>
<tr>
<td>4b</td>
<td>69.6</td>
<td>77.6</td>
</tr>
</tbody>
</table>

- FFV features are complimentary to non-pitch prosodic features
Runtime Improvements

- most costly: extremity filter computation
- reduce support of extremity filters by a factor of $>5$

<table>
<thead>
<tr>
<th>System</th>
<th>Acc</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>4b New Baseline</td>
<td>69.6</td>
<td>77.6</td>
</tr>
<tr>
<td>5a Exclusion of Extremity Filters</td>
<td>69.1</td>
<td>77.0</td>
</tr>
<tr>
<td>5b Improvement of Extremity Filters</td>
<td>70.5</td>
<td>78.9</td>
</tr>
</tbody>
</table>

- runtime decreased by a factor of $\approx 5$, to $0.27 \times$ at 2.5GHz
- unexpectedly: accuracy improvement also
GMM Model Complexity

- increase number of HMM states from 4
- increase number of Gaussians per HMM state from 1

<table>
<thead>
<tr>
<th>System</th>
<th>Acc</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>5b 6</td>
<td>70.5</td>
<td>78.9</td>
</tr>
<tr>
<td>7a 7b 7c</td>
<td>71.3</td>
<td>79.4</td>
</tr>
<tr>
<td>New Baseline</td>
<td>71.7</td>
<td>80.3</td>
</tr>
<tr>
<td>8 states, 2 Gaussians</td>
<td>73.0</td>
<td>81.5</td>
</tr>
<tr>
<td>8 states, 3 Gaussians</td>
<td>73.3</td>
<td>82.4</td>
</tr>
</tbody>
</table>

- increasing model complexity helps
- appears to be a function of amount of data
### Overview of Numerical Results

<table>
<thead>
<tr>
<th>System</th>
<th>Acc</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Baseline</td>
<td>64.8</td>
<td>71.5</td>
</tr>
<tr>
<td>2a Z-Transform</td>
<td>65.7</td>
<td>73.8</td>
</tr>
<tr>
<td>2b PCA Rotation</td>
<td>67.8</td>
<td>74.8</td>
</tr>
<tr>
<td>3 Quadratic Fit</td>
<td>68.9</td>
<td>76.7</td>
</tr>
<tr>
<td>4b Combination with Auxiliary Features</td>
<td>69.6</td>
<td>77.6</td>
</tr>
<tr>
<td>5b Improvement of Extremity Filters</td>
<td>70.5</td>
<td>78.9</td>
</tr>
<tr>
<td>6 4 states, 2 Gaussians</td>
<td>71.7</td>
<td>80.3</td>
</tr>
<tr>
<td>7b 8 states, 2 Gaussians</td>
<td>73.0</td>
<td>81.5</td>
</tr>
<tr>
<td>7c 8 states, 3 Gaussians</td>
<td>73.3</td>
<td>82.4</td>
</tr>
</tbody>
</table>

Relative reduction of error, % 24.2 38.3
A General-Purpose 32 ms Prosodic Vector for Hidden Markov Modeling

- “32 ms Prosodic Vector”
  - a short-time continuous representation of variation in $F_0$
  - can be combined with standard short-time features

- “General-Purpose”
  - applicability across multiple tasks

- “for Hidden Markov Modeling”
  - feature transformation and regularization are helpful
  - performance improves as model complexity increases
  - real-time factor is 0.27

- normative implementation
  - www.cs.cmu.edu/~kornel/software.html