Dimension Reduction of the Modulation Spectrogram for Speaker Verification

Tomi Kinnunen
Speech and Image Processing Unit
Department of Computer Science
University of Joensuu, Finland

Kong Aik Lee and Haizhou Li
Human Language Technology Department
Speech and Dialogue Processing Lab
Institute for Infocomm Research (I²R), Singapore

Speaker Recognition
Recognizing persons from their voices

Physiology and anatomy
("Speech hardware")

Manner of speaking
("Speech software")

Cool, hehe, that rocks! Cool, hehe, hehe, hehe cool

I like to use the same tone all the time … the engine broke down blah blah blah blah
## Speaker recognition systems

<table>
<thead>
<tr>
<th>Physical features (physiology)</th>
<th>Stylistic features (manner of speaking)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front-end (Feature extractor)</td>
<td>Short-term spectrum (MFCC, LPCC)</td>
</tr>
<tr>
<td></td>
<td>Tokenizer (HMM recognizer, prosodic accent extractor)</td>
</tr>
<tr>
<td>Back-end (Classifier)</td>
<td>Gaussian mixture model (GMM), support vector machine, neural nets</td>
</tr>
</tbody>
</table>

Is it possible to extract stylistic features 'directly' from the signal, without a complex tokenizer?

- **+ Computationally efficient**
- **- Computationally expensive**
- **+ Simple implementation**
- **- Complex front-end**
- **- Speaking style assumed to be discrete and categorical**
Speech: a low bandwidth process which modulates higher bandwidth carriers

• Lips, jaw and tongue movements are low-frequency processes that modulate the glottal airflow
  - Energy oscillations at syllabic rates
  - Formant transitions
• Syllable rate of continuous speech ~4 Hz


Modulation spectrum in speech technology

• RASTA filtering
• Improving speech recognition by modulation filtering
  [Kingsbury, Morgan & Greenberg, Speech Communication, 1998]
• Speaker separation from a single-channel audio
  [Schimmel, Atlas & Nie, ICASSP 2007]
• Age and gender classification  [Ajmera & Burkhardt, Odyssey 2008]
• Many others: speech enhancement, voice activity detection, audio compression

In speaker recognition :

• Filtering in the modulation domain to improve conventional cepstral systems  [v. Vuuren & Hermansky, ICSLP 1998], RASTA filtering

Our proposal: using joint acoustic and modulation spectrum, or modulation spectrogram, as a feature [ICASSP 2006]
Modulation spectrum

FFT spectrogram

Temporal trajectory of one subband

Another short-term FFT

Magnitude

Modulation spectrum

Modulation frequency ($\eta$)
What is modulation spectrogram?

Spectrogram:
- short-term (~30ms)
- distribution of the energy across different “acoustic” frequencies

A practical problem: high dimensionality! $(10^3 \sim 10^4)$

Modulation spectrogram:
- Longer-term (200~300 ms)
- joint distribution of the energy across different “acoustic” and “modulation” frequencies
Dimensionality reduction

1. “Acoustic” frequency dimension:
   - A bank of triangular shaped mel-frequency filters as usual

2. “Modulation” frequency dimension:
   - Heavy damping of frequencies above 20 Hz
   - Smooth shape, no harmonic structure
   ==> Apply discrete cosine transform (DCT) to approximate the envelope

Summary of the steps

[Diagram showing the steps]

Mel spectrogram computation

Modulation spectral analysis
Dimensionality reduction

Original mod. spectrogram

Mel-filtered mod. spectrogram

Approx. of the mel-filtered mod. spectrogram with DCT

Dimensionality
129 x 65 = 8385

Dimensionality
30 x 65 = 1950

Dimensionality
30 x 4 = 120
Experiments

- NIST 2001 speaker recognition evaluation (SRE) corpus
  - 174 target speakers
  - 22,418 verification trials (90% impostors, 10% genuine)
  - Training data: 2 minutes / speaker
  - Test data: 0~60 sec
- Gaussian mixture model - universal background model (GMM-UBM) recognizer
- Background model trained from the development set of the NIST 2001 corpus
How many mel filters and DCT coefficients?

NIST 2001 corpus, GMM-UBM recognizer
Context length = 27 frames = 225 milliseconds

Numerical problems due to high dimensionality
Context length

Dimensionality fixed to
$30 \times 2 = 20 \times 3 = 12 \times 5 = 60$

Better time resolution, stationarity
Better mod. spectrum resolution
Comparison with our previous result

[ICASSP 2006]: EER = 25.1 %
Classifier: Long-term averaging classifier with Kullback-Leibler distance + T-norm score normalization
Dimensionality = 3200

[This study] EER = 17.4 %
Classifier: GMM-UBM (256 Gaussians), no score normalization
Dimensionality = 60
Comparison with MFCCs

<table>
<thead>
<tr>
<th>Test duration (s)</th>
<th>MFCC</th>
<th>Mod.spec.</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–20</td>
<td>10.5</td>
<td>18.6</td>
<td>10.5</td>
</tr>
<tr>
<td>20–30</td>
<td>8.5</td>
<td>17.6</td>
<td>8.4</td>
</tr>
<tr>
<td>30–40</td>
<td>7.6</td>
<td>16.6</td>
<td>7.3</td>
</tr>
<tr>
<td>40–60</td>
<td>7.7</td>
<td>15.8</td>
<td>7.3</td>
</tr>
</tbody>
</table>

Fusion: linear score fusion with the weights optimized using logistic regression (FoCal toolkit)

... but the improvement is relatively modest

- Would the benefit be better for significantly longer training and test data?
- Fusion too simplistic?
- Phase differences of the subbands should be retained as well?
Summary

Modulation spectrogram as a feature for speaker recognition

Added mel filtering and DCT to reduce dimensionality

Demonstrated accuracy improvement on NIST 2001 compared to our previous result

Fusion gain with MFCCs was minor, cannot be recommended for applications yet

\[ \text{EER} = 25.1\% \quad \Rightarrow \quad \text{EER} = 17.4\% \]

… but we will not give up yet :-(