AUDITORY-BASED FEATURES FOR ROBUST SPEAKER IDENTIFICATION

Qi (Peter) Li, Ph.D.
Yan Huang

Li Creative Technologies (LcT), Inc.
25 B Hanover Road, Suite 140
Florham Park, NJ 07932
www.licreativetech.com
Email: li@licreativetech.com
Tel: 1-973-822-0048; Fax: 1-973-822-0399

Biometric Consortium Conference, 9:40 AM, September 21, 2010

This work was supported by US AFRL under the contract number FA8750-08-C0028
Outline

- Speaker authentication
- Auditory-based transform (AT)
- Auditory-based feature extraction algorithm (named CFCC)
- Experiments
Speaker Authentication

- Speaker Recognition
  - Authentication by speech characteristics
    - Speaker Verification
    - Speaker Identification
- Verbal Information Verification
  - Authentication by verbal content

3
Outline

- Speaker authentication
- **Auditory-based transform (AT)**
- Auditory-based feature extraction
- Experiments
Auditory System

- The cochlea and traveling wave
Forward Auditory Transform

\[ T(a, b) = \int_{-\infty}^{\infty} f(t) \psi_{a,b}(t) \, dt \]

\[ \psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \left( \frac{t-b}{a} \right)^{\alpha} \exp \left[ -2\pi f_L \beta \left( \frac{t-b}{a} \right) \right] \cos \left[ 2\pi f_L \left( \frac{t-b}{a} \right) + \theta \right] u(t) \]

\[ a = \frac{f_L}{f_c}. \]

- The filter bank was modified from the Gammatone filter.
Inverse Auditory Transform

\[ f(t) = \frac{1}{C} \int_{a=0}^{\infty} \int_{b=0}^{\infty} \frac{1}{|a|^2} \psi_{a,b}(t) \, da \, db \]

- Please refer to [2] for mathematic proving.
- Real number transform
- The frequency distribution can be in any scale: Bark, Mel, ERB, log, etc.
- The filter bandwidths are adjustable.
The responses are similar to auditory neural fiber measurement.
The filter width is adjustable:

(A) \( \alpha = 3, \beta = 0.2 \); (B) \( \alpha = 3, \beta = 0.035 \)
AT Traveling Waves

- Filter outputs to speech signals
AT Experiments

- Fig. 1. Simultaneously recorded speech in a moving car: (A) Mic on lapel; (B) Mic on car visor
Robustness Analysis

- FFT Spectrogram for speech in Fig. 1 (A) and (B)
Proposed AT Spectrograms

- Proposed auditory transform for speech in Fig. 1 (A) and (B)
Spectrum Comparison

- (A) Proposed auditory transform; (B) FFT transform
  Read lines from Fig. 1(A) and blue lines from Fig. 1(B)
Outline

• Speaker authentication
• Auditory-based transform
• **Auditory-based feature extraction (CFCC)**
• Experiments
Auditory-Based Features

- Block diagram of proposed cochlear filter cepstral coefficients (CFCC)
Cochlear Filter Cepstral Coefficients

**Hair Cells:**

\[ h(a, b) = T(a, b)^2 \]

**Variable window sizes:**

\[ S(i, j) = \frac{1}{d} \sum_{b=\ell}^{\ell+d-1} h(i, b), \quad \ell = 1, L, 2L, \ldots \quad \forall i, j, \]

where \( d = [3.5 \tau_i, 20\text{ms}] \) is the window length, \( \tau_i \) is the period of the \( i \)th band, and \( L = 10 \text{ ms} \) is the window shift duration.

**Non-Linearity:** This operation implements cubic root non-linearity from the physical energy to the perceived loudness.

\[ y(i, j) = S(i, j)^{1/3} \]

**DCT:** The discrete cosine transform is used to decorrelate features and generate cepstral coefficients.
Outline

- Speaker authentication
- Auditory-based transform
- Auditory-based feature extraction
- Experiments
### Dataset

- **There are five noisy testing conditions:** -12 dB, -6 dB, 0 dB, and 6 dB SNRs.
- **Download the dataset:**
  
  www.dcs.shef.ac.uk/martin/SpeechSeparationChallenge

<table>
<thead>
<tr>
<th>Disjoint Data Set</th>
<th>No. of Speakers</th>
<th>No. of Utteran./Spk</th>
<th>Total Length (Sec/Spk)</th>
<th>Length / Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>34</td>
<td>20</td>
<td>36.8 s</td>
<td>1.8 s</td>
</tr>
<tr>
<td>Develop. Testing</td>
<td>34</td>
<td>20</td>
<td>18.3 s</td>
<td>1.8 s</td>
</tr>
<tr>
<td>Testing</td>
<td>34</td>
<td>10~20</td>
<td>29.6 s</td>
<td>2-3 s</td>
</tr>
</tbody>
</table>
CFCC vs. MFCC in White Noise
CFCC vs. MFCC in Car Noise
CFCC vs. PLP and RASTA-PLP
Conclusions

• We presented a new auditory-based algorithm for speech feature extraction and applied it to robust speaker identification.

• The algorithm was developed based on a recently presented invertible auditory transform plus several components motivated by the human periphery hearing system.

• Our experiments suggested that under mismatched acoustic conditions, the new features perform significantly better than the MFCC, PLP, and RASTA-PLP features in speaker identification.
References

