Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition

Zhizheng Wu¹, Eng Siong Chng¹, Haizhou Li¹,²,³

¹School of Computer Engineering, Nanyang Technological University, Singapore
²Human Language Technology Department, Institute for Infocomm Research, Singapore
³School of EE & Telecom, University of New South Wales, Australia

12-Sep-2012
Outline

- Motivation
- Voice conversion overview
- Phase feature extraction
- Experiments
- Conclusions
Motivation

• We would like to detect converted speech (synthetic speech) to prevent spoofing attack against speaker verification system

• Phase artifacts in synthetic speech is an informative cue. We study the ways of phase feature extraction


Overview of Voice Conversion (1/3)

• GMM-based voice conversion
Overview of Voice Conversion (2/3)

• Unit-selection based voice conversion

Source frame sequence → Analysis → Target Speech Inventory → Synthesis → Target

Phase artifacts created between analysis and synthesis!
Overview of Voice Conversion (3/3)

• An analysis-synthesis *pass-through* without transformation

![Diagram showing the process of analysis and synthesis with phase artifacts](image-url)

---

6
Phase Artifacts

• Voice conversion techniques focus on spectral conversion
  – Magnitude spectrum contains more information
  – Many vocoders usually use random phase, not the original phase to reconstruct the speech

Phase feature extraction

- Short-time Fourier transform of signal $x(n)$

\[ X(\omega) = |X(\omega)| e^{j\phi(\omega)} \]

$|X(\omega)|$ is the magnitude spectrum

$\phi(\omega)$ is the phase spectrum

MFCC

This study
Cosine Normalized Phase Feature (Cos-phase)

Apply discrete cosine function (DCT) and keep 12 coefficients as the feature.
Modified group delay phase (MGD-phase)

Apply DCT and keep 12 coefficients as the feature
Synthetic speech detector

• GMM-based detector

\[ \Lambda(C) = \log p(C|\lambda_{converted}) - \log p(C|\lambda_{natural}) \]

- \(C\) is the feature vector sequence of a speech signal
- \(\lambda_{converted}\) is GMM model for converted speech
- \(\lambda_{natural}\) is GMM model for natural speech

We use 512 Gaussian components in this study.
Experimental setups

• Corpus: a subset of NIST SRE 2006

<table>
<thead>
<tr>
<th>Training set (number of sessions)</th>
<th>Natural model</th>
<th>Converted model</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

– The duration of each session is 5 minutes
– Three training situations for converted model
  • GMM-based converted speech for training
  • Unit-selection based converted speech for training
  • Pass-through speech for training

We will conduct three experiments under the three training situations
Experimental setups

- Testing set: in total 3500 sessions.
- Evaluation metric: Equal error rate
  - Natural to converted
  - Converted to natural
Experimental setups

• Spoofing attack corpus construction
  – SPTK: http://sp-tk.sourceforge.net/
    • Analysis: Mel-cepstral analysis
    • Synthesis: MLSA filter

**Results:**

- 3 speech models vs 3 features for synthetic speech detection

<table>
<thead>
<tr>
<th>Feature</th>
<th>GMM-based</th>
<th>Unit-selection based</th>
<th>Pass-through</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>16.80</td>
<td>15.35</td>
<td>20.20</td>
</tr>
<tr>
<td>cos-phase</td>
<td>6.60</td>
<td>3.93</td>
<td>5.95</td>
</tr>
<tr>
<td>MGD-phase</td>
<td>9.13</td>
<td>4.60</td>
<td>2.35</td>
</tr>
</tbody>
</table>
Conclusions

• Phase artifacts are useful in detecting the synthetic speech
• When transformation technique is unknown, we may use analysis-synthesis pass-through method to simulate converted data
Thank you!