IMPROVEMENT IN ROBUSTNESS OF SPEECH FEATURE EXTRACTION METHOD USING SUBBAND BASED PERIODICITY AND APERIODICITY DECOMPOSITION

Kentaro Ishizuka*, Noboru Miyazaki+, Tomohiro Nakatani* and Yasuhiro Minami*

*NTT Communication Science Laboratories
+NTT Cyber Space Laboratories

NTT Corporation

ishizuka@cslab.kecl.ntt.co.jp, miyazaki.noboru@lab.ntt.co.jp,
{nak, minami}@cslab.kecl.ntt.co.jp
Abstract

This posters show improvements in the robustness of a speech feature extraction method using Sub-band based Periodicity and Aperiodicity DEcomposition (SPADE). With SPADE, the speech signal is divided into sub-band signals through bandpass filter banks, after which the sub-band signals are decomposed into their periodic and aperiodic features by comb filters. The comb filters are designed individually based on the estimated periodicities of each sub-band signal. Both the periodic and aperiodic features are used as speech feature parameters. An evaluation experiment conducted with AURORA-2J (Japanese AURORA-2) showed that SPADE certainly reduces the average word error rate (WER) under open noise condition. However, SPADE degrades the performance under open channel condition. To cope with this problem, we apply cepstral mean normalization (CMN) to SPADE. The result shows that CMN greatly improves the performance not only for test data under the open-channel condition but also for data under the closed-channel condition. SPADE with CMN achieves an average word accuracy of 89.96 %, and an average WER reduction of 28.61 %. This word accuracy is better than that achieved when using MFCC with CMN.
Background

- Purpose: Robust speech feature extraction for ASR
- Speech extraction method ‘SPADE’ [Ishizuka, 2004]
  (Subband based Periodicity and Aperiodicity DEcomposition)
  - Motivated by the auditory comb filtering hypothesis [de Cheveigné, 1997]
  - Decomposition of periodicity and aperiodicity features of speech signals by using comb filters
  - Both features are used as speech feature parameters for automatic speech recognition (ASR)

References:
**SPADE (1):**

Bandpass filter bank analysis

Input speech signal

Bandpass filter banks

Output waveform
**SPADE (2):**
Feature decomposition using comb filters

**Framed waveforms in subbands**

*Comb filters are designed adaptively for each subband and time frame*

Log powers suppressed by comb filters:
*Periodic features*

Log powers passed through comb filters:
*Aperiodic features*
**SPADE (3):**
Feature representation after decomposition

Output from bandpass filter banks (before decomposition)

- **Decomposition**
- **Periodic feature**
- **Aperiodic feature**

Discrete cosine transform

Speech feature parameter
AURORA-2J

- Noisy digit recognition task (Japanese AURORA-2)
- Two training data sets (Clean / Multicondition)
- Three kinds of test data set (Set A / B / C)

Noise and channel conditions between training and test data sets

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Noise (Additive)</th>
<th>Channel (Convolutitive)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clean</td>
<td>Multi</td>
</tr>
<tr>
<td>A</td>
<td>Open</td>
<td>Closed</td>
</tr>
<tr>
<td>B</td>
<td>Open</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Open</td>
<td>Open/Closed</td>
</tr>
</tbody>
</table>
Characteristics of SPADE

- Performance is:
  - consistently improved under open noise condition
  - degraded under open channel condition
- Therefore, SPADE is:
  - robust as regards additive distortion
  - insufficiently robust as regards convolutive distortion

→ Apply CMN to SPADE
Experimental result  
(Clean training)

<table>
<thead>
<tr>
<th>Test Set</th>
<th>MFCC</th>
<th>MFCC+CMN</th>
<th>SPADE</th>
<th>SPADE+CMN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SetA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SetB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SetC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**MFCC**: 24-filter banks, 12 coefficients and log power, and their Δ and ΔΔ (39 dim in total)

**SPADE**: 24-filter banks, 24 coefficients and log power, and their Δ and ΔΔ (75 dim in total)
Experimental result (Multicondition training)

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Word Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set A</td>
<td>70</td>
</tr>
<tr>
<td>Set B</td>
<td>80</td>
</tr>
<tr>
<td>Set C</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>

- **MFCC**: 24-filter banks, 12 coefficients and log power, and their $\Delta$ and $\Delta\Delta$ (39 dim in total)
- **SPADE**: 24-filter banks, 24 coefficients and log power, and their $\Delta$ and $\Delta\Delta$ (75 dim in total)
Channel distortion intrinsically included in SPADE

Periodic and aperiodic features
(S: Speech signal, C: Comb filter, t: time, s: subband)

\[ A_t(s) = C_t(s)S_t(s) \]  \quad \text{Aperiodic features}

\[ P_t(s) = (1 - C_t(s))S_t(s) \]  \quad \text{Periodic features}

\( C_t \) changes in time, and differs between subbands; depends on F0 tendency of speakers

Improvement in Sets A and B under clean training might be due to the elimination of the effect of \( C_t \) by CMN.
Elimination of channel distortions by CMN

Periodic and aperiodic features
(S: Speech signal, C: Comb filter, H: Channel distortion)

\[ A_t(s) = C'_t(s)H_t(s)S_t(s) \quad \text{Aperiodic features} \]
\[ P_t(s) = (1 - C'_t(s))H_t(s)S_t(s) \quad \text{Periodic features} \]

\( C_t \) changes in time, and differs between subbands; depends on F0 tendency of speakers and \( H_t \)

The effect of \( H_t \) is eliminated by CMN (same as MFCC). The effect of \( C'_t \) might be eliminated by CMN as above. Improvement in Set C under multicondition training might be due to these two effects of CMN.
Conclusion

• Speech feature extraction method ‘SPADE’
  (Subband based Periodicity and Aperiodicity DEcomposition)
  – Robust as regards additive noise
  – Insufficiently robust as regards convolutive distortion

• CMN can improve performance of SPADE
  – Performance is better than MFCC+CMN

• The effect of CMN
  – Elimination of effect of channel distortion
  – Elimination of effect of comb filtering