Auditorily Motivated Structuring Element for Morphological Filtering of Speech Spectrograms Applied to Automatic Speech Recognition

F. de-la-Calle-Silos, A. Gallardo-Antolin, C. Pelaez-Moreno
Universidad Carlos III de Madrid

Abstract
- Morphology operations are applied to noisy spectrograms using an auditorily motivated structuring element based in the masking properties of the human auditory system.
- This method was tested in automatic speech recognition tasks, improving the recognition rates in a noise-contaminated version of the ISOLET database.

ASR System Overview

Morphological Filtering

Closing operator: \( S \odot M = (S \oplus M) \ominus M \)

The closing operator preserve the regions which have a similar shape to the structuring element, removes small holes and is the basic tool of morphological noise removal.

Masking
- Auditory masking occurs when the perception of one sound is affected by the presence of another sound.
- Auditory masking in the frequency domain is known as simultaneous masking, and in the time domain is known as temporal masking.

Simultaneous Masking Model
- The simultaneous masking is better represented in logarithmic scales where the spacing and the masker slope are regularly extended [3].
- The simultaneous masking is modeled with a slope of +30 dB per band for the lower bands and −8 db per band for the upper bands.

Temporal Masking Model
- The premasking effect is modeled as a constant slope of +25 dB/mseg.
- A fitted model for single masker-induced postmasking presented in [2] is used to model the postmasking effect:

\[ M(t_d, t_m) = a (b - \log t_d) (t_m - c) \]

Results
Recognition results in terms of WER[%] for 80-band filterbanks in noisy INSOLET database:

<table>
<thead>
<tr>
<th>Features</th>
<th>Mismatched</th>
<th>Matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>53.80 ± 1.23 17.06 ± 0.93</td>
<td></td>
</tr>
<tr>
<td>MFCC + SS</td>
<td>39.85 ± 1.20 16.70 ± 0.92</td>
<td></td>
</tr>
<tr>
<td>MFCC + MF</td>
<td>51.80 ± 1.23 16.77 ± 0.92</td>
<td></td>
</tr>
<tr>
<td>MFCC + MF [1]</td>
<td>52.08 ± 1.24 16.93 ± 0.93</td>
<td></td>
</tr>
<tr>
<td>MFCC + SS + MF</td>
<td>35.60 ± 1.18 16.73 ± 0.92</td>
<td></td>
</tr>
<tr>
<td>MFCC + SS + MF [1]</td>
<td>36.53 ± 1.19 16.83 ± 0.92</td>
<td></td>
</tr>
</tbody>
</table>

Conclusions
- Auditorily motivated structuring elements applied to spectrograms for automatic speech recognition.
- An alternative shape and weight scheme for the structural element that emulates more precisely the masking effects of the human peripheral system.
- The combination of spectral subtraction and morphological filtering for automatic speech recognition improves the recognition rates.

References