Human Factor Cepstral Coefficients: improving on a generation’s speech feature extraction algorithm

Mark D Skowronski
Computational NeuroEngineering Laboratory
University of Florida
Outline

- Automatic Speech Rec overview
- Mel Freq Cepstral Coefficients: a generation’s standard
- Bio-inspired BW
- Testing and Results
- Conclusions
Automatic Speech Recognition is the extraction of phonetic information from an utterance of speech (Text-to-Speech).

- **Isolated/Continuous** speech
- **Dependent/Independent** speaker operation
- **Word/Phoneme** recognition unit
- Vocabulary **size** and **perplexity**

**Input → Feature Extraction → Classification**
Information: phonetic, gender, age, emotion, pitch, accent, physical state, additive/channel noise
**Feature Extraction**

**Goal:** emphasize phonetic information over other characteristics.

- Acoustic: *formant frequencies, bandwidths*
- Model based: *linear prediction*
- Filter-bank based: *mel freq cepstral coeff*

Provides dimensionality reduction on *quasi-stationary* windows.
MFCC Details

\[ x(t) \rightarrow \mathcal{F} \rightarrow \text{FILTER BANK} \rightarrow \text{LOG ENERGY} \rightarrow \text{DCT} \rightarrow \text{CEPSTRAL DOMAIN} \]
Choose the option from the list: **MFCC Filterbank**

- Design parameters: \textit{FB freq range, number of filters}.
- Center freqs equally-spaced in \textit{mel frequency}.
- Endpoints of triangle (BW) set by center freqs of \textit{adjacent filters}.

\textbf{Ambiguity} in the original FB description has led to \textit{various implementations} used in the ASR community today.
Filter bandwidth

![Plot showing filter bandwidth with different models: D&M, Slaney, HTK, hfcc](image)
Endpoints of triangle set by ERB (Moore and Glasberg, 1983).
Experiments

**Goal:** To determine the performance of hfcc in various noise environments for a common ASR task with a standard classifier as compared to the versions of mfcc.

**Vocabulary:** English digits ‘zero’ through ‘nine’.

**Features:** hfcc, mfcc (3 versions), delta coefficients.

**Classifier:** 1st-order left-right HMM word model.

**Noise:** white, pink, babble.

**Corpus:** TI46 (8 male, 8 female, 26 utterances/speaker/word).
Male & female TRAIN speakers, 4 random TEST speakers, 10 trials.
White noise: Relative

Male & female TRAIN speakers, 10 trials, relative to D & M, 95 % confidence interval.
Nearest-neighbor vowel phoneme experiment:

Male $\rightarrow$ Male: 97.3 %, Female $\rightarrow$ Female: 92.3 %.
Male only, 4 random TEST speakers, 10 trials.
White noise: Male only, Relative

Male only, 4 random TEST speakers, 10 trials, relative to D & M.
White noise: Female only

Female only, 4 random TEST speakers, 10 trials.
Female only, 4 random TEST speakers, 10 trials, relative to D & M.
Male only, 4 random TEST speakers, 10 trials.
Male only, 4 random TEST speakers, 10 trials, relative to D & M.
Pink noise: Male only, Relative

Male only, 4 random TEST speakers, 10 trials, relative to D & M.
Noise summary
<table>
<thead>
<tr>
<th></th>
<th>hfcc</th>
<th>DM</th>
<th>Slaney</th>
<th>HTK</th>
</tr>
</thead>
<tbody>
<tr>
<td>M&amp;F</td>
<td>60.5</td>
<td>55</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>relative</td>
<td>5.1 ± 2</td>
<td>−</td>
<td>−0.2 ± 3</td>
<td>−0.1 ± 2.5</td>
</tr>
<tr>
<td>Male</td>
<td>74</td>
<td>62.5</td>
<td>58</td>
<td>62.5</td>
</tr>
<tr>
<td>relative</td>
<td>11.2 ± 4</td>
<td>−</td>
<td>−5.0 ± 2</td>
<td>0 ± 2</td>
</tr>
<tr>
<td>Female</td>
<td>61</td>
<td>57.5</td>
<td>51.5</td>
<td>56</td>
</tr>
<tr>
<td>relative</td>
<td>3.4 ± 2</td>
<td>−</td>
<td>−5.8 ± 2</td>
<td>−1.7 ± 2.5</td>
</tr>
<tr>
<td>M,Babble</td>
<td>89</td>
<td>89.5</td>
<td>90</td>
<td>87</td>
</tr>
<tr>
<td>relative</td>
<td>−0.8 ± 1</td>
<td>−</td>
<td>−0.4 ± 1.5</td>
<td>−2 ± 0.8</td>
</tr>
<tr>
<td>M,Pink</td>
<td>65</td>
<td>70</td>
<td>68</td>
<td>66.5</td>
</tr>
<tr>
<td>relative</td>
<td>−4.5 ± 4</td>
<td>−</td>
<td>−1 ± 1.5</td>
<td>−2.5 ± 2</td>
</tr>
</tbody>
</table>
Male only, white noise, 10 trials.
Male only, pink noise, 10 trials.
Conclusions

- We introduce a novel scheme, human factor cepstral coefficients (hfcc), for **decoupling** filter bandwidth from other filter bank design parameters in the mel freq cepstral coefficient algorithm.

- We use Moore and Glasberg’s expression for critical bandwidth (ERB), a function only of center frequency, to determine filter bandwidth.

- ASR experiments show hfcc to be more robust to **white noise** than mfcc, yet performance across different noise sources is mixed.
Conclusions

- **Gender** information is present in extracted mfcc/hfcc features. Recognition rates for male-only tests were higher than that for female-only tests across all experiments.

- Varying the number of filters used in hfcc does **not** significantly affect experimental results.