Frequency-centroid features for word recognition of non-native English speakers

Pierre Berjon\textsuperscript{1}
Rajib Sharma\textsuperscript{2}
Avishek Nag\textsuperscript{3}
Soumyabrata Dev\textsuperscript{4,5}

\textsuperscript{1}Department de Sciences du Numérique, INP-ENSEEIHT, Toulouse, France
\textsuperscript{2}Department of DSIS, Indian Institute of Information Technology Dharwad, India
\textsuperscript{3}School of Electrical and Electronic Engineering, University College Dublin, Ireland
\textsuperscript{4}ADAPT SFI Research Centre, Dublin, Ireland
\textsuperscript{5}School of Computer Science, University College Dublin, Ireland
1. **Automatic word recognition (AWR): relevance and challenges.**

2. **Feature extraction: MFCCs and FCs.**

3. **Feature modelling: Two-stage CNN.**

4. **Experimental setup: The Speech Accent Archive.**

5. **Results.**
   - **Clean Speech.**
   - **White noise, HFchannel noise, Babble noise.**

6. **Conclusion and future work.**
Why word recognition?

- Proliferation of low-cost smart digital assistants.
Keyword spotting (KWS)

- KWS models must be **smaller** and involve **less computation**.

- Most of the input will be **silence or background noise**, not speech, so false positives/alarms due to those must be minimized.

- The important unit of recognition is a **single word or short phrase**, not an entire sentence.

- Speech commands (Google): Objective is to standardize the evaluation of the important task of keyword spotting (KWS), and thereby promote collaboration and progress in the field, just like ImageNet.

Automatic Word/Speech Recognition: Challenges

- **Accents of non-native speakers.**
  - **Non-native speakers:** A speaker with Spanish, for example, as its primary language/mothertongue is a native Spanish speaker. It is, then, a non-native speaker of any other language, such as English.

- **Noisy audio signals (low signal-to-noise ratio).**

- Different dialects.

- Speakers with voice pathology.

- Words unknown by the system.

---


The Mel filterbank

- Mimics the spectral analysis process of the basilar membrane of the inner ear.

- The Mel frequency scale represents the perceptual frequency scale of the ear, wherein certain “critical bands” significantly influence our perception of sound.

\[ f_{\text{mel}} = 2595 \log_{10}(1 + \frac{f}{700}) \]

- The representation of the critical bands in the analog frequency (Hz) scale results in a sequence of non-uniform and overlapping bands of triangular filters, well known as the Mel filterbank.

Mel frequency cepstral coefficients (MFCCs)

Pre-emphasised speech

\[
s'(n) = s(n) - 0.98s(n - 1) \leftrightarrow S'(f),
\]

\[
\hat{S}'_k = \int f |S'(f)|^2 W_k(f) \, df, \quad k = 1, 2, \ldots, K,
\]

\[
\hat{C}_n = \sum_{k=1}^{K} (\log(\hat{S}'_k)) \cos[n(k - \frac{1}{2}) \frac{\pi}{K}], \quad n = 1, 2, \ldots, K.
\]

In the above equations, \( K \) is the total number of filters in the Mel filterbank. \( \hat{C}_n \) is the \( n^{\text{th}} \) cepstral coefficient: the compressed feature obtained by Discrete Cosine Transform.

- Most widely accepted and used features for speech/voice applications. Based purely on energy.
- They may not always be the optimal features, and other complementary features may be needed to boost performance in real-life conditions.


Complementary features: Frequency-centroids (FCs)

We attempt to compute the dominant frequency in the filter range.

\[ s(n) \leftrightarrow S(f), \quad (4) \]

\[ S_k(f) = \begin{cases} S(f), & f_{k-1} < f < f_{k+1} \\ 0, & \text{otherwise} \end{cases}, \quad (5) \]

\[ F_k = \int_f f \frac{|S_k(f)|}{\int_f |S_k(f)| \, df} \, df, \quad k = 1, 2, \ldots, K. \quad (6) \]

- MFCCs are **based purely on energy**.

- Frequency centroids (FCs) utilize the **frequency information of the critical bands** of the Mel filterbank.
Two-stage Convolution Neural Network (CNN)

- Inspired by **small-footprint** KWS systems. **Only acoustic modelling**, no language modelling unlike in speech recognition.

- Multi-stage (upto 5) CNNs were experimented with, but **without much gain in performance beyond the second stage**.

- The input layer, or feature set, consists of a two-dimensional matrix \((D = 1)\) if either the MFCCs or the FCs are used as features. If both are used, the input layer is a three-dimensional matrix with \(D = 2\).


Speech Accent Archive database

**Table:** Number of speech signals used for training and testing.

<table>
<thead>
<tr>
<th>Accent</th>
<th>Data</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>kids</td>
</tr>
<tr>
<td>Arabic</td>
<td>Train</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>28</td>
</tr>
<tr>
<td>French</td>
<td>Train</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>15</td>
</tr>
<tr>
<td>Spanish</td>
<td>Train</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>15</td>
</tr>
</tbody>
</table>

- Speech Accent Archive database consists of speakers from **177 countries belonging to 214 different mother tongues**, speaking a particular English passage.

- Arbitrarily selected a speech subset of only **three different accents (French, Spanish and Arabic)**.
  - Manually isolated **five specific words** for the isolated word recognition task. On an average, the utterances in the dataset are of 0.5 s duration.
  - The speakers with these accents consider their ‘accents’ as their primary language or mothertongue, however **they may be from different regions of the world**. For example, the speakers with Spanish accents are constituted from not only Spain but different regions of Latin America and other former Spanish colonies.

Results: clean speech

Table: Word recognition accuracy for clean speech.

<table>
<thead>
<tr>
<th>Accent</th>
<th>MFCCs</th>
<th>FCs</th>
<th>MFCCs + FCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>0.78</td>
<td>0.48</td>
<td>0.80</td>
</tr>
<tr>
<td>French</td>
<td>0.80</td>
<td>0.49</td>
<td>0.82</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.75</td>
<td>0.52</td>
<td>0.80</td>
</tr>
</tbody>
</table>

- The FCs are not competitive as standalone features.
- In combination with the MFCCs, FCs provide decent performance gain, for each of the three accents.
Results: white noise

Figure: Accuracy of the word recognition system under white noise.
Results: hfchannel noise

Figure: Accuracy of the word recognition system under hfchannel noise.
Results: babble noise

Figure: Accuracy of the word recognition system under babble noise.
Conclusion and Future work

- The results observed for both clean and noisy speech show that the FCs are versatile features which aid the MFCCs in accurate classification of the words.

- Three different CNNs have been utilized for the three different accents. The ability of the proposed features in incorporating inter-accent variation has not been studied.
  - As such, this work is only a part of a larger project in which the accents could be identified before word recognition or an end-to-end model is developed which caters to the inter-accent variations.

- It is possible that for other words the accuracy might be substantially better or worse. As such, using a larger set of words would give a better sense of the overall results.

- The effectiveness of the FCs in the experiments conducted encourages the exploration of other kinds of spectral representation of the speech signal, such as the Hilbert Spectrum and the HNGD spectrum for extracting complementary features.

---


End of the Presentation

Thank You