Semi-Supervised Keyword Spotting in Arabic Speech Using Self-Training Ensembles

Mohamed Mahmoud (elgeish@stanford.edu)

Abstract

The Arabic speech recognition field suffers from the scarcity of properly labeled data. We introduce a pipeline that performs semi-supervised segmentation of audio then—after hand-labeling a small dataset—feeds labeled segments to a supervised learning framework to select, through many rounds, an ensemble of models to infer labels for a larger dataset; using which we improved the F1 score of KWS from 75.85\% (using a baseline model) to 90.91\% on a ground-truth test set. We picked the keyword na’am (yes) to spot. We define the system’s input as an audio file of an utterance and the output is a binary label: keyword or filler.

Corpora

Selected utterances from West Point’s and King Saud’s (unlabeled) corpora were segmented and normalized. We hand-labeled the former and a small subset of the latter; both sets took turns as either a training or a test set for the other but were never mixed. White noise was added to enrich the former set and generalize better.

Features

Each file is divided into short-term frames with sliding windows; we extract 34 features for each frame, from which we derive mid-term windows’ features (\(\mu\) and \(\sigma\)). We use zero crossing rate; energy; entropy of energy; spectral centroid, spread, entropy, flux, and roll-off; 13 MFCCs; and a 12-element Chroma vector and its \(\sigma\).

Models Pipeline

West Point’s Corpus

Hand-tuned

SVM

King Saud’s Corpus

Analyze & Segment

Semi-supervised SVM

Ensemble Analysis: Uniform ensembles outperformed weighted ones as ensemble members improved. They also won when matched against the ground-truth manually labeled datasets: 90% F1 score vs. 79% for the top ensemble of each category.

Final Round’s Results

<table>
<thead>
<tr>
<th>Train (KS Labels)</th>
<th>Test</th>
<th>Predictor</th>
<th>Train F1</th>
<th>Test F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>15680 (predicted)</td>
<td>1762 (WP+Noise)</td>
<td>GB</td>
<td>93.8</td>
<td>90.91</td>
</tr>
<tr>
<td>15680 (predicted)</td>
<td>881 (WP)</td>
<td>GB</td>
<td>93.8</td>
<td>90.79</td>
</tr>
<tr>
<td>15680 (predicted)</td>
<td>881 (WP)</td>
<td>Ensemble</td>
<td>N/A</td>
<td>90.70</td>
</tr>
<tr>
<td>15680 (predicted)</td>
<td>881 (WP)</td>
<td>SVM</td>
<td>94.6</td>
<td>89.05</td>
</tr>
<tr>
<td>15680 (predicted)</td>
<td>1762 (WP+Noise)</td>
<td>Ensemble</td>
<td>N/A</td>
<td>85.87</td>
</tr>
<tr>
<td>15680 (predicted)</td>
<td>1762 (WP+Noise)</td>
<td>SVM</td>
<td>94.6</td>
<td>84.70</td>
</tr>
<tr>
<td>15680 (predicted)</td>
<td>881 (WP)</td>
<td>Ensemble</td>
<td>N/A</td>
<td>71.99</td>
</tr>
<tr>
<td>15680 (predicted)</td>
<td>1762 (WP+Noise)</td>
<td>SVM</td>
<td>82.7</td>
<td>61.82</td>
</tr>
<tr>
<td>15680 (predicted)</td>
<td>1762 (WP+Noise)</td>
<td>GB</td>
<td>84.1</td>
<td>60.98</td>
</tr>
<tr>
<td>15680 (predicted)</td>
<td>1762 (WP+Noise)</td>
<td>Ensemble</td>
<td>N/A</td>
<td>55.17</td>
</tr>
</tbody>
</table>

Discussion & Future Work

- Ensemble learning has a boosting effect for weak learners; self-training helped bootstrap the supervised learning framework; combining them was a big plus.
- KNNs tend to overfit; Extra Trees are slow at scoring.
- Downsampling the negative examples had the biggest impact on improving the results.
- The best performing frame and window parameters are within the literature’s recommended range (even for Arabic): 25ms with a step size of 10ms.
- Future Work: Attempts to train neural networks yielded an F1 score of 41\% – various setups to test; trying random, instead of grid, search for parameters.

Ensembles

- Uniform Ensemble Prediction = sign(\(\sum \text{pred}_i\))
- Weighted Ensemble Prediction = sign(\(\sum \text{pred}_i \times \text{F}_i\))
REFERENCES


