The ICSI Speaker Recognition System

Kofi Boakye, Barbara Peskin

International Computer Science Institute
Berkeley, CA USA
Motivation

I. Text-dependent systems have high performance, but limited flexibility when compared to text-independent systems

   Capitalize on advantages of text-dependent systems in this text-independent domain by limiting words of interest to a select group:

   Backchannels (*yeah, uhhuh*), filled pauses (*um, uh*), discourse markers (*like, well, now…*)

   => high frequency and high speaker-characteristic quality

II. GMMs assume frames are independent and fail to take advantage of sequential information

   => Use HMMs instead to model the evolution of speech in time
System Overview

Approach

Model each speaker using a collection of keyword HMMs

Speaker models generated via adaptation of background models trained from a development data set

Use standard likelihood ratio approach:

- Compute log likelihood ratio scores using accumulated log probabilities from keyword HMMs

Use a speech recognizer to:

1) Locate words in the speech stream
2) Align speech frames to the HMM
3) Generate acoustic likelihood scores
Implementation

Keywords

Discourse markers: \{actually, anyway, like, see, well, now, you\_know, you\_see, i\_think, i\_mean\}

Filled pauses: \{um, uh\}

Backchannels: \{yeah, yep, okay, uhhuh, right, i\_see, i\_know \}

Keyword Models

Simple left-to-right (whole word) HMMs with self-loops and no skips

4 Gaussian components per state

Number of states related to number of phones and median number of frames for word

HMMs trained and scored using HTK

Acoustic features: 19 mel-cepstra, zeroth cepstrum, and their first differences
Implementation

Scoring

\[ \text{LLR}(X) = \log(p(X|\text{Targ})) - \log(p(X|\text{UBM})) \]

**Target score**: output from scoring adapted HMM against frame sequence corresponding to keyword token (token determined using SRI recognizer)

**UBM score**: output from scoring unadapted (background) HMM against same frame sequence

**Normalization** (by total number of frames):

\[ \text{score} = \frac{\sum_i \text{target}_i - \sum_i \text{UBM}_i}{\sum_j \#\text{frames}_j} \]

*For more implementation details refer to Odyssey'04 paper, “Text-Constrained Speaker Recognition on a Text-Independent Task” (K. Boakye and B. Peskin)*

4 June 2004

NIST SRE Workshop
Development Set: SWB1

Data partitioned into 6 splits
Tests use jack-knifing procedure:

Test on splits 1 - 3 using background model trained on splits 4 – 6 (and vice versa)

Results:
EER = 1.25%, DCF = 0.0060 (8-side train)
EER = 5.33%, DCF = 0.0215 (1-side train)
System Performance: Dev

Development Set: SWB2

Testing procedure similar to SWB1, but now data partitioned into 10 splits

Background model generated from data of 5 splits

Results:

EER = 3.11%, DCF = 0.0183 (8-side train)
EER = 9.17%, DCF = 0.0347 (1-side train)
Evaluation on SRE’04

System run for 8-side and 1-side training conditions

Run on English-language trials only
  • Non-English trials given arbitrary score, since no ASR output available

Background models generated from splits 6 -10 of SWB2’s Extended Data set
  • Note mismatch (e.g., no cell data)

Results:

EER = 8.85%, DCF = 0.0382 (8-side train)

EER = 13.06%, DCF = 0.0526 (1-side train)

*Note: DET curves show performance on all and only English-language trials*
System Performance: Eval

Evaluation: Fusion results

Fusion of scores with GMM system (provided by SRI: SWB2-trained)

Combination weights determined using SWB2 development data

Results:

<table>
<thead>
<tr>
<th>System</th>
<th>Train Condition</th>
<th>EER (%)</th>
<th>DCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword</td>
<td>1 side</td>
<td>13.06</td>
<td>0.0526</td>
</tr>
<tr>
<td>GMM</td>
<td>1 side</td>
<td>11.99</td>
<td>0.0414</td>
</tr>
<tr>
<td>Keyword+GMM</td>
<td>1 side</td>
<td>10.52</td>
<td>0.0364</td>
</tr>
<tr>
<td>Keyword</td>
<td>8 sides</td>
<td>8.85</td>
<td>0.0382</td>
</tr>
<tr>
<td>GMM</td>
<td>8 sides</td>
<td>8.76</td>
<td>0.0334</td>
</tr>
<tr>
<td>Keyword +GMM</td>
<td>8 sides</td>
<td>6.57</td>
<td>0.0263</td>
</tr>
</tbody>
</table>

Note: DET curves show performance on all and only English-language trials
Post-eval System Contrast

Fisher-trained models

Background models trained using data obtained from Fisher collection

=> Presumed better match to evaluation data

Results:

<table>
<thead>
<tr>
<th>Background Model</th>
<th>Train Condition</th>
<th>EER (%)</th>
<th>DCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWB2 Keyword</td>
<td>1 side</td>
<td>13.06</td>
<td>0.0526</td>
</tr>
<tr>
<td>Fisher Keyword</td>
<td>1 side</td>
<td>12.98</td>
<td>0.0445</td>
</tr>
<tr>
<td>SWB2 Keyword</td>
<td>8 sides</td>
<td>8.85</td>
<td>0.0382</td>
</tr>
<tr>
<td>Fisher Keyword</td>
<td>8 sides</td>
<td>7.06</td>
<td>0.0306</td>
</tr>
</tbody>
</table>

*Note: DET curves show performance on all and only English-language trials*
Next Steps

Normalizations: So far, the system does not yet employ any of the standard normalizations (H-norm, T-norm, …)

Additional Words: Expand list of current types; try highest-frequency words (cf. Sturim et al. 2002); consider longer expressions

Filter keywords: Filter occurrences of words based on usage (e.g. “like” as discourse marker vs. as verb)

Explore / tune model parameters: Number of states and Gaussians, adaptation weights, etc.

LVCSR Contrast System: Use phone-level HMMs covering all of the data, rather than word-specific HMMs
  => Less tightly-focused models, but can score all the words

Other System Fusions: Combine systems incorporating additional knowledge sources (e.g. LM capturing preference among back-channels, not just how speaker sounds when says them)
Concluding Remarks

First time participation for ICSI – still ramping up

Encouraged by results, though system has known shortcomings we are eager to address (e.g., no normalizations yet)

System serves as a contrast to a submission by SRI, which contained a fusion of our Keyword and their GMM scores

System demonstrates power of using tightly focused models for a very small subset of speaker-characteristic words

Acknowledgements: Special thanks to SRI for ASR output, GMM score fusion, and general technical assistance