Landmark-based speech recognition

Amit Juneja and Carol Espy-Wilson
Speech Communication Lab, Institute for Systems Research (ISR), Department of Electrical and Computer Engineering
http://www.isr.umd.edu/Labs/SCL

1. INTRODUCTION
We present a probabilistic framework for our automatic speech recognition (ASR) system based on acoustic phonetic knowledge. In this event-based system (EBS), knowledge-based acoustic parameters (APs) are first used to segment the speech signal into the broad manner classes – vowel, sonorant consonant, fricative, stop and silence. Landmarks (articulatory speech events) obtained from the broad class segmentation are then used to extract APs for place of articulation recognition.

2. PROBABLISTIC PHONETIC FEATURE HIERARCHY
Figure 1: Phonetic Feature Hierarchy
Figure 1 shows the phonetic feature hierarchy [1] used for recognition. In a probabilistic representation of this hierarchy, each bifurcation is assigned a probability. For example, the probability of the phoneme /n/ can be written as: P(n) = P(speech) P(sonorant/speech) (1-P(syllabic/sonorant)) P(consonantal/syllabic) P(nasal/consonantal) P(alveolar/nasal)

3. PROBABLISTIC BROAD CLASS SEGMENTATION
A Support Vector Machine (SVM) [2] classifier is trained for each of the features – sonorant, continuant, and syllabic, in addition to silence. SVM outputs are converted to probabilities. The posterior probabilities are found for the broad classes fricative (Fr), vowel (V), sonorant consonant (SC), stop (ST) and silence (SIL). Multiple segmentations, ranked in probabilities, are found using a beam search algorithm. Part (c) of Figure 2 shows the most probable unconstrained segmentation obtained for the digit “zero”.

4. CONSTRAINED SEGMENTATION
A broad class level finite state automata (FSA) is used to constrain the broad class paths in accordance with the vocabulary. For example, broad class level representation of the digit ‘zero’ is SIL-Fr-V-SC-V-SC-SIL. This can be represented by the FSA shown in Figure 3. The constrained segmentation for the digit “zero” is shown in part (d) of Figure 2.

4. RESULTS
Results on broad class segmentation are shown in Table 1. EBS performs better than a state-of-the-art Hidden Markov Model (HMM) based system using the same set of APs but using only relevant and minimum information. Although the HMM-MFCC system performs better on TIMIT, its performance drops markedly on cross-database testing. EBS performs close to HMM-AP even though it uses 10 times less training data than the HMM system.

Table 1: All results in percentage. All models were trained on TIMIT

<table>
<thead>
<tr>
<th>System: Front-end</th>
<th>Database</th>
<th>EBS APs</th>
<th>HMM APs</th>
<th>HMM MFCCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT (Broad class)</td>
<td>87.1</td>
<td>85.2</td>
<td>87.9</td>
<td></td>
</tr>
<tr>
<td>TIDIGITS (Word accuracy)</td>
<td>70.3</td>
<td>72.3</td>
<td>63.1</td>
<td></td>
</tr>
</tbody>
</table>

5. REFERENCES