SPEECH SEGMENTATION USING PROBABILISTIC PHONETIC FEATURE HIERARCHY AND SUPPORT VECTOR MACHINES

Amit Juneja and Carol Espy-Wilson

Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742, USA
juneja@glue.umd.edu

ABSTRACT

We propose a method that combines acoustic-phonetic knowledge with support vector machines for segmentation of continuous speech into five classes: vowel, sonorant consonant, fricative, stop, and silence. We show that using a probabilistic phonetic feature hierarchy, only four classifiers are required to recognize the five classes. Due to the probabilistic nature of the hierarchy, the method overcomes the disadvantage of the traditional acoustic-phonetic methods where the error is carried down the hierarchy. On the other hand, the hierarchical approach allows the use of comparable amount of training data of two classes that each classifier is designed to discriminate. The segmentation method with 13 knowledge-based parameters performs considerably better than a context-independent Hidden Markov Model (HMM) based approach that uses mel-cepstrum based parameters. The probabilistic nature of the algorithm allows the method to be extended to phoneme and word recognition with a small number of classifiers.

1. INTRODUCTION

Support Vector Machines (SVMs) [1] have been shown to be very effective in feature detection [2, 3] and classification of segmented phonemes [4, 5] in continuous speech. SVMs have attractive properties for pattern classification, for example, capability of learning from small amount of data, capacity control, and elegant handling of high dimensional data. But the use of SVMs as a unified statistical framework in automatic speech recognition remains limited because of inferior modeling of time-varying dynamics and coarticulation effects. We propose a segmentation method for continuous speech using acoustic-phonetic knowledge and SVMs that can be extended to phoneme recognition and higher recognition tasks.

In our event-based system (EBS), speech is first segmented into broad classes: vowels, sonorant consonants (nasals and semi-vowels), fricatives and silence using the phonetic feature hierarchy shown in Figure 1. Then the parameters for place and voicing are extracted to decide upon the phonemes. We show that for segmentation into five broad classes we do not need five classifiers because the hierarchy in Figure 1 can be exploited in a probabilistic manner. The idea can be extended to phoneme recognition because all phonemes can be represented by presence or absence of 20 phonetic features [6]. At present, EBS gives out a single segmentation but the method can be extended for producing multiple hypothesis segmentations in a manner similar to [8]. Such a probabilistic segmentation would enable the system to be used in word and sentence recognition by using the research that has already gone into phoneme classification using SVMs [5, 4] as well as in acoustic phonetics.

It has been shown before [9] that knowledge-based acoustic parameters (APs) are more speaker independent and give comparable performance for digit recognition task than the mel-cepstrum based coefficients. We have used 13 APs that are acoustic correlates of manner phonetic features - sonorant, sonorant consonant, plosive and obstruent, in addition to silence for the segmentation task. Classifiers for different phonetic features use parameters that are correlates of the corresponding phonetic features. The method has three added advantages: (1) Not all APs are used for all decisions, (2) Since APs carry strong physical interpretation, it is easy to tell whether the pattern matcher has failed or the knowledge based APs have failed to perform their designated task. (3) The method can take advantage of years of research that has gone into acoustic phonetics.

2. DATABASE

The TIMIT database [10] was used for the experiments presented in this paper. The phonetically rich ‘si’ and ‘sx’ sentences of all dialect regions from the training set were used for training and the ‘si’ sentences of all dialect regions from the test set were used for testing.

3. SUPPORT VECTOR MACHINES

SVMs are learning machines for pattern classification and regression tasks based on statistical learning theory [1]. Given a set of training vectors \( \{x_i\}_{i=1}^N \) and the corresponding class labels \( \{y_i\}_{i=1}^N \) such that

\[
y_i \in \{-1, +1\} \quad \text{and} \quad x_i \in \mathbb{R}^n,
\]

SVMs select a set of support vectors \( \{x_i^{SV}\}_{i=1}^{N_{SV}} \) that is a subset of the training set \( \{x_i\}_{i=1}^N \) and find an optimal decision function

\[
f(x) = \text{sign}\left( \sum_{i=1}^{N_{SV}} y_i \alpha_i K(x_i^{SV}, x) - b \right)
\]

where \( K \) is an a priori chosen kernel function. The weights \( \alpha_i \), the set of support vectors \( \{x_i^{SV}\}_{i=1}^{N_{SV}} \) and the bias term \( b \) are found from the training data using quadratic optimization methods. Many methods have been suggested to convert SVM outputs
to probabilities, but in our project we currently clip the output between -1 and +1 and map the result to [0,1]. We have used radial base function (RBF) kernels and linear kernels in these experiments. For RBF kernel,

$$K(x_t, x) = \exp(-\gamma|x_t - x|^2)$$

where the parameter $\gamma$ is chosen empirically by cross-validation from the training data. For the linear kernel,

$$K(x_t, x) = x_t, x + 1$$

The experiments in this project were carried using the SVM Light toolkit [11], which provides very fast training of SVMs.

4. METHOD

Figure 1 shows the phonetic feature hierarchy suggested in [6]. We have made the hierarchy probabilistic by giving probabilities to each branch. The hierarchy is shown only to the level of five broad classes - vowel (V), sonorant consonant (SC), fricative (Fr), stop (ST) and silence, in this picture. Consider a sequence of parameter vectors $x_t = \{\alpha_{t-s}, \alpha_{t-s+1}, ..., \alpha_{t+e}\}$ at a given instant $t$ where $s$ previous frames and $e$ following frames are used along with the current frame $\alpha_t$ for analysis. Assume that the frame at time $t$ lies in the region of one of the broad classes. We can write the posterior probability of vowel at time $t$ as (with the absence of an event denoted by a bar)

$$P(V|x_t) = P(speech, sonorant, SC|x_t)$$

$$= P(speech|x_t)P(sonorant|speech, x_t)$$

$$= (1 - P(SC|sonorant, x_t))$$

$$= (1 - p_1)p_2(1 - p_3)$$

and similarly for the other broad classes. Note that we have used the fact that the presence of the event sonorant implies the presence of the event speech, that is,

$$P(SC|sonorant, x_t) = P(SC|sonorant, speech, x_t)$$

In EBS, we train one SVM for each branching in the hierarchy. The probabilities $p_t$ are posterior probabilities and we calculate them from the output of SVMs as described Section 3. Therefore, for the recognition of five broad classes, only four binary SVMs are needed. In Table 1 we show the classes that are trained against each other for building these four SVMs. A clear advantage of this system is that each class to be recognized does not have to be trained against all the other classes, for example, the samples of V do not have to be trained against the samples of the classes - SC, Fr, ST and silence. For each binary classifier in Table 1, a comparable amount of training data for the two classes is available. Moreover, since all the decisions are binary, the method overcomes the need to find good multi-class SVMs. Although non-probabilistic hierarchy can be used to limit the number of classifiers to four, such an approach will not allow probabilistic segmentation, therefore, the errors at phonetic feature level will not be corrected by language and duration constraints.

A stop burst is characterized by a period of closure of about 30ms and an energy burst. At a frame step size of 5ms, we use $s = 6$ and $e = 3$ for the stop burst classifier. For all other classifiers, a single frame of speech is used. That is, $s = 0$ and $e = 0$. With no duration and language constraints, and by making the independence assumption across frames, the class label at time $t$ is hypothesized by

$$\hat{w}_t = \arg \max_w P(w|x_t)$$

with $w \in \{V, SC, ST, Fr, SILENCE\}$

The segmentation of the test signal is then found by collapsing the consequent identical class labels. Note that this is not a strictly frame-based system because (1) a certain number of adjoining frames are used for stop burst detection, (2) acoustic-phonetic knowledge is used to normalize some of the APs with their values in certain locations, for example, syllabic peaks and dips [7].

Table 2 shows the acoustic parameters used by each SVM classifier. Unlike the HMM based approach, each classifier uses only the parameters that are required for the corresponding phonetic feature.

5. EXPERIMENTS AND RESULTS

The training of SVMs was performed with 5000 samples for each of the classes in Table 1 selected randomly from the TIMIT training files. Less than 1000 files were used for the training of SVMs. RBF kernels were used for the features silence, sonorant and plosive. Linear SVM was used for sonorant consonant detection.

HMM experiments [13] were carried out for comparison purposes using HTK [14]. A 39 parameter set consisting of 12 mel-frequency cepstral coefficients (MFCCs) and energy with their delta and acceleration coefficients were used in the HMM broad classifier. All the manner class models were context-independent 3-state (excluding entry and exit states) left-to-right HMMs with diagonal covariance matrices and 8-mixture observation densities for each state. A skip transition was allowed from the first state to the third state in each model. All the ‘sx’ and ‘si’ files, that is, a total of 3696 utterances were used for training the HMM broad classifier.

<table>
<thead>
<tr>
<th>Branch in hierarchy</th>
<th>class +1</th>
<th>class -1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>silence</td>
<td>speech</td>
</tr>
<tr>
<td>$p_2$</td>
<td>sonorant</td>
<td>non-sonorant</td>
</tr>
<tr>
<td>$p_3$</td>
<td>sonorant consonant</td>
<td>vowel</td>
</tr>
<tr>
<td>$p_4$</td>
<td>stop burst</td>
<td>frication noise</td>
</tr>
</tbody>
</table>

Table 1. Training of phonetic feature SVMs
A manner class segmentation system may not separate out two consecutive phonemes having the same manner representation. Therefore, for the purpose of scoring, the reference phoneme labels from the TIMIT database were mapped to manner class labels, and consecutive identical broad class labels were mapped into one. The resulting manner class labels were used as the reference labels for scoring EBS as well as the HMM broad classifier. The results are shown in Table 3. EBS shows a better segmentation performance than the HMM based system with three times less parameters and four times less training data. Figure 2 shows the spectrogram of a TIMIT test sentence and the broad class labels generated by EBS as well as HMM along with the TIMIT phoneme labels. It can be easily seen from the examples shown in the figure that EBS does a finer analysis of the spectrum. In particular, the HMM broad classifier is not able to separate out a stop when it is followed by a fricative, while EBS can make this distinction.

### 6. DISCUSSION AND FUTURE WORK

We have presented a method for broad class segmentation, but the method can be extended to phoneme recognition and bigger recognition tasks in at least three different ways. All the three methods will use the complete phonetic feature hierarchy. An example of representation of the phoneme /n/ with the phonetic feature hierarchy appears below

\[
P(|n|/x_t) = P(speech, sonorant, SC, nasal, alveolar/x_t) = P(speech/x_t)P(sonorant/speech, x_t)P(SC/sonorant, x_t)P(nasal/SC, x_t)P(alveolar/nasal, x_t)
\]

The three ways we propose to use EBS for bigger recognition tasks are

1. **Frame based approach.** The method shown here for segmentation is simply extended to take into account the complete phonetic feature hierarchy. Duration and language constraints can be integrated into the system to carry out N-best segmentation in a manner similar to [8]. The drawback of such a frame based approach is that coarticulation and dynamic variations within phonemes will not be appropriately modeled.

2. **Acoustic-phonetic landmark based approach.** In this method, landmarks like the locations of stop burst, locations of the syllabic dips in the sonorant consonant regions and the syllabic peaks in the vowel regions can be analyzed for place and voicing phonetic features. A very high classification accuracy of stop consonants and fricatives has been obtained [15, 16] using knowledge based measurements on segmented broad classes. The parameters enlisted in similar research and upcoming acoustic-phonetic research can be used with SVMs at appropriate landmarks to obtain the posterior probabilities of different phonetic features in the phonetic feature hierarchy.

3. **Segment-based approach.** This approach can combine the knowledge of acoustic phonetics as well as model diphones or triphones by using dynamic time-aligned kernels (D-TAK) [5] for SVMs. Starting with different (N-best) hypotheses of an utterance, D-TAK SVM can operate on hypothesized segments or pairs of segments to give out the probabilities of phonemes or diphones respectively. The hierarchy will restrict the search space of the phones or diphones, for example, diphones containing stops and fricatives may not be searched in segments containing vowels and sonorant consonants in a particular hypothesis segmentation of the utterance.

### 7. CONCLUSION

We have shown that SVMs can be combined effectively with acoustic-phonetic knowledge both in terms of knowledge based APs and a phonetic feature hierarchy to provide a segmentation method for continuous speech. EBS does a very fine analysis of the speech spectrum and has a good capability of separating out transient sounds like stops. A probabilistic hierarchy allows the reduction of the number of support vector classifiers and avoids the need of using multi-class SVMs.

### 8. REFERENCES

Fig. 2. (a) Spectrogram of TIMIT sentence "to many experts this trend was inevitable", (b) Labels generated by HMM method, (c) TIMIT phoneme labels, (d) Labels generated by EBS. In region 1, HMM broad classifier misses a short nasal segment. In region 2 and 3, EBS could separate out a stop from a following fricative while the HMM recognizer misses the stop. In the EBS labels, the aspiration noise following the stop burst of \( \text{t} \) is recognized as Fr. This is because SVM model is built for the stop burst and not the complete stop region. Aspiration can be distinguished from frication as the hierarchical system is completed. It can also be seen that EBS gets the off-glide \( \text{y} \) of the vowel \( \text{iy} \) in the word 'many'.


