THE DESIGN OF ACOUSTIC PARAMETERS FOR SPEAKER-INDEPENDENT SPEECH RECOGNITION

Nabil N. Bitar and Carol Y. Espy-Wilson

Boston University, Electrical & Computer Engineering Dept.
44 Cummington St.
Boston, MA. 02215
E-mail: nabil@bu.edu, espy@bu.edu

ABSTRACT
This paper presents a two-stage procedure, based on the Fisher criterion and automatic classification trees, for designing acoustic parameters (APs) that target phonetic features in the speech signal. This procedure and a subset of the TIMIT training set were used to develop acoustic parameters for the phonetic features: sonorant, syllabic, strident, palatal, alveolar, labial and velar. Results on a subset of the TIMIT test set show that the developed parameters achieve correct phonetic-feature classification rates in the 90 % range with the exception of stop-consonant place of articulation (labial, alveolar and velar) where correct classification is about 73 %. Furthermore, it is shown that by basing the acoustic parameters on relative measures (e.g. an acoustic parameter that measures energy in a frequency band relative to energy in the same band at another time instant) the effect of interspeaker variability (e.g. gender) on the parameters is reduced.

1. INTRODUCTION
A speech signal contains phonetic and non-phonetic information. The non-phonetic component reveals, among other things, information about the speaker’s gender, dialect and emotional state. On the other hand, the phonetic component carries information required to decipher the linguistic message in the speech signal. In speech recognition applications, different methodologies are used to extract the phonetic component in the speech signal. Fundamental to all methodologies is an appropriate signal representation that emphasizes the phonetic contrast among the different speech sounds. Since phonetic contrast is described by the distinctive phonetic-feature theory [1] (closely related to the manner and place of speech production), phonetic features are chosen as a basis for deriving a speech signal representation.

The signal representation in this case consists of a set of acoustic parameters (APs) that target these phonetic features.

In previous research [2], several acoustic parameters that target manner-of-articulation phonetic features were developed. These parameters were used in an event-based approach to speech recognition [2] to recognize broad speech classes. The same parameters were slightly modified in [3] and [4] to fit into the frame-based Hidden Markov Model (HMM) framework and were compared to a Mel-Cepstral representation in a broad-class recognition task. The recognition results showed that the APs based on phonetic-features target the relevant information in the speech signal and reduce speaker-dependent effects. These APs were developed using acoustic-phonetic knowledge (e.g. [5] [6]) and histogram analysis to eyeball the data, a time-consuming and subjective process. In contrast, the parameter design procedure presented in this paper is efficient and uses the objective Fisher criterion and classification trees. Place of articulation features are also considered in this paper whereas, in the earlier work, only manner of articulation features were investigated.

In section 2, the AP design procedure is outlined. In section 3, APs that were derived using the developed procedure are discussed. These parameters were tested in a phonetic-feature classification task and the results are presented in section 4. Finally, the main conclusions from this work are discussed in section 5.

2. PARAMETER DESIGN PROCEDURE
The parameter design procedure has its roots in [7]. The differences between the approach undertaken in this research and that in [7] are (1) the APs are based specifically on acoustic cues relevant to phonetic features, (2) the APs are defined in relative terms to reduce the effects of interspeaker variability and (3) classification trees are used to eliminate redundant parameters. The objective is to develop parameters that best characterize a phonetic feature and sepa-
rate it from its antonym(s). To accomplish this task, the Fisher criterion \( (FC) \) \[8\] was chosen along with classification trees. Given data samples from a number of classes \((C)\), the Fisher criterion \( (FC) \) \[8\] computes the ratio of the between class scatter \(SB\) to the within class scatter \(SW\) as

\[
FC = \frac{\text{trace}(SB)}{\text{trace}(SW)}
\]

where,

\[
SB = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - m_{i})(x^{(i)} - m_{i})^T
\]

and,

\[
SW = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{C} (x^{(i)} - m_{i})(x^{(i)} - m_{i})^T
\]

In equations 2 and 3, \(m_{i}\) is the sample mean for class \(i\) computed from \(n_{i}\) samples that consist of the observation vectors \(x^{(i)}\) whereas \(m\) is the pooled data mean computed from \(n\) samples. The parameter that maximizes the \(FC\), i.e. maximizes the ratio of between-class scatter to within-class scatter, from a pool of parameters is the one that is best for classification. The procedure for deriving a set of APs consists of the following steps:

1. Group the set of all sounds that have a phonetic feature in one group and the sounds that do not have that feature in another group. This grouping is guided by the phonetic-feature hierarchy of Figure 1. For instance, in APs for the syllabic phonetic feature, only vowels (syllabic) and semivowels and nasals (nonsyllabic) are considered.

2. Based on acoustic phonetics, define a set of generic APs, with free parameters that are intended to separate the formed groups from each other. For instance, the generic AP, \(E[f_{1} : f_{2}]\), is an energy measure between frequencies \(f_{1}\) and \(f_{2}\) (free parameters) with the condition \(f_{1} < f_{2}\). The objective is to determine optimum values for the free parameters. The possible set of values that the free parameters (e.g. \(f_{1}\) and \(f_{2}\)) can take may be restricted using acoustic phonetic knowledge.

3. For each generic AP, determine the free parameter values that result in local maxima in the \(FC\) surface. The resulting APs, obtained by fixing the free parameters to these values, are considered to be optimum.

4. Use the APs generated from the previous step and any additional APs\(^2\), if needed, to grow a classification tree that best distinguishes between the groups of sounds considered. The APs selected by the classification tree constitute the final set of APs that target the considered phonetic feature(s). In this research, classification trees are greedily grown using the minimum deviance criterion and then pruned back using cross validation.

As an example of acoustic parameter design, consider the place-of-articulation feature that differentiates the alveolar stridents (nonpalatals): /s/, /z/ from the palatal stridents: /sh/, /zh/, /ch/, /jh/\(^3\). The parameter design procedure in this case was based on the following steps:

1. All /s/, /z/ samples from the training set were placed in one group (the nonpalatal group) while the /sh/, /zh/, /ch/ and /jh/ samples were placed in the palatal group.

2. The energy of the palatal sounds is concentrated in the third formant region while that of the nonpalatal sounds is concentrated in the fifth formant region. Based on this acoustic-phonetic knowledge, generic parameters were chosen so that energy in a mid frequency band (around the third formant) is measured relative to (a) energy in a higher frequency band, (b) energy in a lower frequency band and (c) total energy. Some of these parameters are computed within the obstruent boundaries relative to the maximum, minimum and average values of the same parameters across the utterance.

\(^2\)The additional APs [e.g. zero-crossing rate, formant values] do not need the Fisher-criterion optimization stage since they do not have any free parameters to be fixed.

\(^3\) /sh/ and /jh/ are noncontinuant stridents but are lumped here with the continuant stridents /sh/ and /zh/ because they are very similar to /sh/ and /zh/, respectively, sharing the same palatal place of articulation.

Figure 1: The phonetic feature hierarchy that guided the development of acoustic parameters to target phonetic features.
3. The mid frequency band was allowed to vary by 3000 Hz around the third formant (F3) value with a minimum bandwidth of 300 Hz. The F3 value was estimated for each TIMIT utterance in the training set separately using the Waves\textsuperscript{4} formant tracker.

4. For each of the generic parameters, a Fisher surface was computed. As an example, the Fisher surface obtained by computing the energy in the band \([f_{st} : f_{end}]\) relative to the overall energy at the same time frame and then averaged across the obstreant is shown in Figure 2.

5. From each of the Fisher surfaces, the local maxima are picked specifying candidate parameter variables. The final parameter set is obtained by feeding all candidate parameters to a classification tree and selecting the ones that significantly contribute to the intended discrimination. As a result, the parameter that measures the energy in the frequency band \([F3-187 \text{ Hz}, F3 + 594 \text{ Hz}]\) relative to the overall energy within the obstreant was chosen as the best parameter yielding 91\% correct classification. Two additional parameters were chosen by the tree that increase the overall classification rate to about 93\%.

![Figure 2: Fisher criterion for the parameter which computes the energy between \(f_{st}\) and \(f_{end}\) relative to the overall energy within the duration of considered sound. In the figure, the origin is \(F3-1000\) (Hz).](image)

In order to emphasize the importance of reducing interspeaker variability and specifically gender differences, the density of the best classification parameter obtained without F3 normalization and that obtained with F3 normalization for the nonpalatal sounds are plotted in Figure 3 (a) and (b), respectively. In comparing these two distributions, it is clear that the normalized parameter is better able to reduce gender differences.

### Table 1: Phonetic Features and APs obtained from the optimization process.

<table>
<thead>
<tr>
<th>Phonetic Feature</th>
<th>APs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sonorant</td>
<td>E[0:688] / E[4000:8000]</td>
</tr>
<tr>
<td>Nonsyllabic</td>
<td>dip-to-peak of E[2750:3562] / E[1250:2562]</td>
</tr>
<tr>
<td>Strident (Noncont.)</td>
<td>/ [E[31:318000] / Emin[F3 + 318000]</td>
</tr>
<tr>
<td>Stop place</td>
<td>(labial, velar, alveolar)</td>
</tr>
</tbody>
</table>

\textsuperscript{4}Waves is a signal processing and analysis tool developed by Entropic Research Laboratories Inc.

\textsuperscript{5}Computed from the Entropic Waves software based on energy and maximum cross correlation.
4. CLASSIFICATION RESULTS

Using classification trees obtained from the development stage, the developed APs were evaluated in classifying the phonetic features on both the development set and an independent test set that consists of all 504 phonetically compact TIMIT test sentences (SI sentences). The results on the development set and independent test set are in most cases comparable indicating that the developed parameters do target the relevant phonetic information in the speech signal. The largest error occurs for the stop place of articulation. At present, only information in the stop burst is used. However, formant transition information will be incorporated as error analysis showed that most of the errors may be eliminated using such information.

5. CONCLUSION

In this paper, a procedure for designing acoustic parameters that target phonetic features was defined and tested. Furthermore, it was shown that acoustic parameters defined in a relative fashion reduce speaker variability in the parameter space while emphasizing the phonetic information. These results have great indications to speaker-independent speech recognition in terms of reducing the speech-model complexity and lessening the demand on the training data while improving recognition results. This is supported in [4] and in work on vocal tract normalization [9].

REFERENCES


Table 2: Features and classification results on the training and test sets obtained from the classification trees grown in the development process.

<table>
<thead>
<tr>
<th>Phonetic Feature</th>
<th>% Error on Training Set</th>
<th>% Error on Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sonorant</td>
<td>4.9</td>
<td>5.4</td>
</tr>
<tr>
<td>Nonsyllabic (intervocalic)</td>
<td>1.6</td>
<td>2.2</td>
</tr>
<tr>
<td>Nonsyllabic (pre/post-vocalic)</td>
<td>12.7</td>
<td>12.6</td>
</tr>
<tr>
<td>Strident (for fricatives)</td>
<td>5.2</td>
<td>5.0</td>
</tr>
<tr>
<td>Palatal</td>
<td>7.4</td>
<td>8.0</td>
</tr>
<tr>
<td>Strident (for Noncontiguous)</td>
<td>4.5</td>
<td>5.3</td>
</tr>
<tr>
<td>Stop place (labial, alveolar, velar)</td>
<td>23.6</td>
<td>27.0</td>
</tr>
</tbody>
</table>

Figure 3: Distribution of best parameter computed using the nonpalatal fricative (/s/, /z/) samples for males (m) and females (f). In (a) bands were chosen relative to third formant (F3) location. In (b) parameter was independent of F3.