CLASSIFICATION OF STOP CONSONANT PLACE OF ARTICULATION: COMBINING ACOUSTIC ATTRIBUTES

Atiwong Suchato
Speech Communication Group, Laboratory of Electronics, MIT, Cambridge, MA, USA
atiwong@mit.edu

ABSTRACT
This study evaluates the classification of stop consonant place of articulation in running speech using knowledge-based cues. Acoustic attributes are chosen to capture four categories of cues: amplitude and energy of burst, formant movement of adjacent vowels, spectrum of noise after the release, and some temporal cues. Correlation analysis shows no redundant information among cross-category attributes, although a few attributes in the same category show high correlation. Combinations of non-redundant attributes are used for the place classifications based on Mahalanobis distance. When stops contain release bursts, the classification accuracies are better than 90%. When bursts are absent, the classification can only rely on formant structures, and this results in reduced accuracy. It is also shown that voicing and vowel frontness contexts lead to a better classification accuracy of stops in CV context. However, there is no evidence of significant improvement for stops in VC context. When stops are located between two vowels, information on the formant structures in the vowels on both sides can be used. The probabilities of each place of articulation from two classifiers, each of which uses information on one side of the stops, are combined in order to obtain final decisions. Such combination yields the best classification accuracy of 95.5%. By using appropriate methods for stops in different contexts, an overall classification accuracy of 92.1% is achieved. The overall accuracy could be improved significantly if stops in VC context with no release burst were classified more accurately.

INTRODUCTION
Researchers have been approaching the problem of automatic speech recognition (ASR) in various ways. One of the most prevalent methods is a statistical method in which speech recognizers learn the patterns of some spectral representations of the speech units expected in incoming utterances from some sets of examples and then try to match the incoming patterns with the patterns learned. However, this approach to automatic speech recognition does not use much knowledge of human speech production, and although some of the speech recognizers using this approach yield good recognition accuracies, they are still inferior to human's ability to recognize speech, especially in a noisy environment. Many believe that to improve the performance of a speech recognizer, knowledge about human speech production should be used more extensively. One of the approaches to ASR that relies heavily on speech production knowledge is a distinctive feature-based speech recognition system, in which each of the underlying word segments is represented with a set of discrete phonological units called the distinctive features. Certain combinations of such distinctive features, called feature bundles, can contrast all of the sounds in any human language. There are about 20 such distinctive features in English, and each distinctive feature is specified by a binary value. More details on...
distinctive features can be found in (Stevens, 1998). In a distinctive feature-based system, analog acoustic signals are examined for cues, based on acoustic-phonetic knowledge, which can be used to determine the value of each distinctive feature in the feature bundles.

Stop consonants represent one of the various classes of sounds in human speech. In English, there are three types of stop consonants according to their place of articulation, namely labial stops (‘b’, ‘p’), alveolar stops (‘d’, ‘t’), and velar stops (‘g’, ‘k’). Along with voicing, distinctive features that discriminate among the three places of articulation need to be determined in order to uniquely identify the six English stop consonants. For several decades, different researchers have studied the acoustic attributes that affect human discriminating ability for place of articulation for stop consonants (Delattre et al., 1955, Blumstein & Stevens, 1979). Release bursts and formant transitions of the vowels adjacent to stop consonants were found to be important to the determination of the stop consonant place of articulation. Recent studies have shown that using combinations of acoustic attributes provided good place of articulation classification accuracy (Hasegawa-Johnson, 1996, Stevens et al., 1999, Ali, 2001).

This study proposes a different set of acoustic attributes from the ones used in previous studies. These acoustic attributes are chosen to capture four categories of cues, including amplitude and energy of burst, formant movement of adjacent vowels, spectrum of noise after the release, and some temporal cues. It was shown in (Suchato, 2004) that the values of each individual attribute in this set distributed significantly differently among the three places and could separate them at some level. The purpose of this study is to evaluate how well combinations of these acoustic attributes can be used in the place of articulation classification tasks, as well as to observe the effect of voicing and vowel contexts on the classification results.

THE DATABASE
The speech data used in this study were extracted from utterances naturally read by 2 male speakers and 2 female speakers. Speakers were asked to read a list of 110 meaningful and grammatically correct sentences. The sentences were selected so that the number of stop consonants with the three places and the two voicing properties were fairly well balanced. Each utterance was digitized at 16kHz, and stored directly to a computer. The stop consonants to be included in this study were restricted to only the stop consonants that were located next to at least one vowel. The vowel can be on either side of the stop consonant segment as long as it is adjacent to that stop segment regardless of any word or syllable boundaries. The underlying segments were determined as appeared in the transcriptions of the sentences. Here, we will define a CV token as a stop consonant segment that has a vowel segment immediately to the right and a VC token as a stop consonant segment that has a vowel segment immediately to the left. Thus, a stop consonant that was included in this study must create either a CV or VC token, or both types of token. The measurements involving any CV or VC tokens were made from the original acoustic signal at the place where the tokens of interest were located. The vowel in each CV or VC token must not be reduced. Flaps and other stops that contained irregular behaviors due to high level of gestural overlap were omitted from the study.

ACOUSTIC ATTRIBUTES
The acoustic attributes used are categorized into 4 groups as shown in Table1. Each of the acoustic attributes describing the spectral shape of the release burst is in dB and in the form of a relative amount between two measurements (e.g. X-Y is the value of X relative to Y). ‘Av’ is
the amplitude of the first formant prominence measured at either the voicing onset or the voicing offset of the adjacent vowel. ‘Ahi’ and ‘Amax23’ are the amplitudes of the biggest peaks of the burst spectra in the range from 3.5kHz to 8kHz and from 1.25kHz to 3kHz. ‘A23’ is the average peak amplitude of the burst spectrum in the range from 1.25kHz to 3kHz. ‘Ehi’ is the total energy of the burst spectrum in the range from 3.5kHz to 8kHz. ‘E23’ is the total energy of the burst spectrum in the range from 1.25kHz to 3kHz. ‘Avhi’, ‘Av3’ and ‘Av2’ are the amplitudes of the biggest peaks of the vowel spectra measured at either the voicing onset or the voicing offset of the adjacent vowels in the range from 3.5kHz to 8kHz, from 1.5kHz to 3kHz, and from 1.25kHz to 2.5kHz respectively. ‘A3’ and ‘A2’ are the amplitudes of the biggest peaks of the burst spectra in the range from 1.5kHz to 3kHz and from 1.25kHz to 2.5kHz respectively.

Table 1. Acoustic attributes used and their corresponding categories

<table>
<thead>
<tr>
<th>Spectral shape of the release burst</th>
<th>Formant structure</th>
<th>Temporal cues</th>
<th>Noise frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Av-Ahi,Av-Amax23,Ahi-A23,</td>
<td>F1o,F2o,F3o,F2b,F3b,</td>
<td>VOT</td>
<td>CgF10a</td>
</tr>
<tr>
<td>Ehi-E23,Avhi-Ahi,Av3-A3,</td>
<td>F3o-F2o, F3b-F2b,</td>
<td>Cls_Dur</td>
<td>CgF20a</td>
</tr>
<tr>
<td>Av2-A2</td>
<td>dF2,dF3,dF2b,dF3b</td>
<td></td>
<td>CgFa</td>
</tr>
</tbody>
</table>

‘F1o’, ‘F2o’, and ‘F3o’ are the frequencies of the first, second and third formant at either the voicing onset or the voicing offset of the adjacent vowel respectively. ‘F2b’ and ‘F3b’ are the frequencies of the second and the third formant frequencies at the time of the release burst. ‘dF2’ is the difference between F2o and the second formant frequency at 20 ms after the voicing onset of the following vowel or 20 ms prior to the voicing offset of the preceding vowel, while ‘dF3’ is similar to dF2 but for the third formant frequency. ‘dF2b’ is the difference between F2b and the second formant frequency at 20 ms after the release burst for the CV case or 20 ms prior to the release for the VC case. ‘dF3b’ is similar to dF2b but for the third formant frequency. ‘F3o-F2o’ and ‘F3b-F2b’ are the differences in frequency between the third and the second formant frequencies at the voicing onset of the following vowel or the voicing offset of the preceding vowel and the time marked as release burst, respectively. ‘VOT’ is the voicing onset time, which is the duration of the interval between the release burst and the time marked as the voicing onset of the following vowel. ‘Cls_Dur’ is the closure duration, which is the duration of the interval between the time marked as the voicing offset of the preceding vowel and the release burst. The acoustic attributes in the last category include cgF10a, cgF20a, and cgFa. The value of each attribute is the center of gravity in the frequency scale of the power spectrum obtained from a specific portion of the speech signal. The portion that corresponds to cgF10a is from the time marked as the release burst to the point 10 ms after that. For cgF20a, it is the portion from the time marked as the release burst to the point 20 ms after that. For cgFa, the corresponding portion is from the time marked as the release burst to the time marked as the voicing onset of the following vowel.

CORRELATION ANALYSIS

Redundancy in the information contained in all of the acoustic attributes could degrade the classification performance and cause unnecessary computational cost. Therefore, we wish to identify the acoustic attributes that are highly correlated with other acoustic attributes and use the findings to guide the selection of acoustic attribute subsets for the classification.
experiments. The squares of the correlation coefficients ($\rho^2$) between every pair of acoustic attributes were calculated from the tokens for which both acoustic attributes are applicable.

The attributes Av3-A3 and Av2-A2 are highly correlated. Their $\rho^2$ are always the highest among any pairs of attributes. Both Av3-A3 and Av2-A2 are intended to capture the amplitude of the release burst in different, but overlapping, frequency ranges. The correlation result indicates that the difference in the frequency ranges for the two attributes does not add much additional information. Therefore, using only one of them in the classification should suffice. F2o and F2b are highly correlated, as well as F3o and F3b, for CV tokens with voiced stops. However, they are not strongly correlated in the voiceless cases. This is not surprising since the VOTs of voiced stops are significantly shorter than the ones of voiceless stops. So, the time points where the voicings of the vowels start are closer to the release burst and this results in more similar formant frequencies between the two time points. We have found that there are no highly correlated acoustic attributes across the four categories of acoustic attributes used in this study.

SOME RESULTS FROM CLASSIFICATION EXPERIMENTS

Combinations of not-highly-correlated acoustic attributes were used as classification feature vectors in various classification experiments. The classification criterion used was the minimum Mahalanobis distance criterion. The classification accuracies were estimated by using the Leave-One-Out Cross Validation (LOOCV) technique. When stops contain release bursts, the classification accuracies are 93.9% and 90.4% for 2521 CV tokens and 1526 VC tokens respectively. When bursts are absent, the classifier can only rely on formant structures. The classification accuracies in this case are 73.2% and 81.1% for 2734 CV tokens 2271 VC tokens respectively.

To examine the benefit of knowing the voicing and vowel frontness contexts, classification accuracies obtained from two different classifiers were compared: (1) classifiers trained on the training data whose contexts match with the contexts of the test data and (2) classifiers trained on training data with the same size, but of random context. The result shows that, for CV tokens, knowing the frontness of the adjacent vowels improves the classification accuracy by 1.6% and knowing both the frontness of the adjacent vowels and the voicing of the stops improves the accuracy by 2.9%. However, knowing only the voicing does not significantly improves the accuracies. Also, neither type of information significantly improve the classification accuracy of the VC tokens.

Stop consonants in some of the CV and VC tokens were located between two vowels (regardless of any word or syllable boundaries), and then there were CV and VC tokens that shared these stop consonants. In predicting the place of articulation of these stop consonants, it should intuitively be better to take into account the information on the vowels on both sides of each stop consonant than to use only the information on the vowel on either side of the stop consonant. In this study, we proposed the combination of the information from both sides at two levels: 1) the attribute-level combination, and 2) the classifier level combination.

In the attribute-level combination, one classification was performed for each of the stop consonants that had a vowel on each side. The acoustic attribute that was used as the classification feature vector of the classifier was constructed from the acoustic attributes related to both of the vowels and the burst release of that stop, if these attributes existed. This resulted
in a classification feature vector that had more dimensions than the vectors for the classification of its corresponding CV and VC tokens. When the information of the release bursts is used, the classification accuracy for 807 stops in the VCV form is 94.3%. When only the formant information is used, the classification accuracy is 87.1% for 933 stops in the VCV form. In both cases, the classification accuracies are significantly better than the ones when only the information on either side of the stops is used separately. In the classifier-level combination, the information obtained from the vowels on both sides of a stop consonant was not used together at first. The corresponding CV and VC tokens were classified separately using their own set of acoustic attributes. Posterior probabilities of the three places of articulation provided by the CV and VC classifiers were combined using the rule in equation 1.

\[
P(\text{Place}_i \mid \bar{X}) = \frac{\prod_{j=1}^{N} P_j(\text{Place}_i \mid \bar{X}_j)^{\gamma_j}}{\sum_{k=1}^{N} \prod_{j=1}^{N} P_j(\text{Place}_i \mid \bar{X}_j)^{\gamma_j}}
\]

\(P(\text{Place}_i \mid \bar{X})\) is the combined posterior probability of the \(i^{th}\) place of articulation. \(P(\text{Place}_i \mid \bar{X}_j)\) is the probability that the place of articulation of the stop consonant is labial \((i=1)\), alveolar \((i=2)\), or velar \((i=3)\), based on the observation. \(\gamma_j\) is the weight of the probability obtained from the \(j^{th}\) classifier. And, \(N\) is the number of the classifiers used. The best classification accuracies obtained are 95.5% for 800 stops with release bursts. The classification accuracy when only the formant information is used is 87.5% for 930 stops. In both cases, the weight used for CV classifier is 0.6 and the one used for VC classifier is 0.4. The classification accuracies are again significantly better than the ones when only the information on either side of the stops is used separately. The improvements in the classification accuracy between both levels of combination are not statistically significant.

By taking the presence of the release burst, the voicing and vowel frontness contexts, and the possible presence of the VCV form into consideration, the overall classification accuracy evaluated on the entire database is 92.1% for 4007 stops. The largest portion of the error, which is approximately 40%, is caused from stops in VC tokens that lack the release bursts.

**DISCUSSION AND CONCLUSIONS**

This paper shows the results of using a set of acoustic attributes in the stop consonant place of articulation classification task. Although the stops involved in this classification are somewhat restricted, the achieved classification accuracy of 92.1% is still encouraging, given that additional processing could be developed in the future to handle the left out cases. Compared to the place classification accuracy of 85% for syllable-initial stops obtained by Stevens et al. (1999), we achieve an approximately 8% higher classification accuracy for stops regardless of syllable structure. In Stevens et al. (1999), six acoustic attributes were used with a statistical classification method similar to the one used in this study to classify syllable-initial stops. In this study, we used approximately seventeen acoustic attributes, including the six attributes used in Stevens, et al. (1999), to classify stop consonants in broader contexts. The improvement gained was mainly due to the introduction of additional acoustic attributes, which allowed more of the relevant information to be captured. In Hasegawa-Johnson (1996), ten acoustic attributes in
context-dependent place classification were used. A classification accuracy of 84% was obtained. Direct comparison should not be made on the classification accuracies due to the difference in the database used. Still, it is worth noting the difference in utilizing context information. Hasegawa-Johnson used 36 different context classes, including all possible combination of 2 genders and 18 right contexts, while we used contexts in a broader sense, i.e. voicing of the stops and vowel frontness. The fact that Hasegawa-Johnson used a large number of context classes leads to a lack of generalization of the training data, and the need for considerably more training materials hampers the classification accuracy. Also, both in Hasegawa-Johnson (1996) and in Stevens et al. (1999), the acoustic attributes used for each context class were fixed, while in our study here, we have found that using different acoustic attribute subsets led to different accuracies, although not all of the acoustic attribute combinations gave significantly different classification accuracies. In Ali (2001), the place classification accuracy obtained is 90%. There are both different and similar aspects between Ali (2001) and this study. Although the exact measurements from the speech signal are different, both studies generally used rather similar information on the spectral shape of the release burst, formant transitions, and some temporal cues, as well as the frequency of the noise in the release burst region. In this study, this information is contained in the acoustic attributes that were used as classification vector of a statistical classifier, while in Ali (2001), the information is used for making decisions in a decision tree with hard thresholds, also learned from the training data. Despite the difference, both methods of classification can be considered knowledge-based methods and the resulting parameters in both classifier models, i.e. the thresholds and the positions of decision nodes in the decision tree, and the canonical weights used in the corresponding analyses in our study, should help improve our understanding of the acoustic-phonetic characteristics of the stop consonants. Such a benefit is hard to obtain, if at all possible, from a spectral-based data-driven approach.

ACKNOWLEDGEMENTS
This research was supported in part by Grant DC02978 from the National Institutes of Health.

REFERENCES