Objective. The objective of this brief report is to list the results of my research work in the area of feature-based speech recognition. This includes the results of the segmentation and coarse categorization, the fricative-consonant classification and the stop-consonant classification.

1. Introduction

In spite of the wealth of information that exists in the literature about the recognition and perception of the obstruents (namely the fricatives, affricates and stop consonants), considerable research is still needed to fully understand this intriguing class of phonemes. Their noisy, dynamic, relatively short, weak, speaker- and context-dependent nature made them one of the most challenging phonemes to automatically recognize in continuous speaker-independent speech. In our work, we investigated the acoustic-phonetic characteristics of the obstruents in the framework of a front-end feature-based speaker-independent phoneme-based continuous speech recognition. Using the wealth of information that exists in the literature and using our own acoustic, spectrogram and statistical analysis experiments on continuous speech from the TIMIT database, we studied several acoustic features for their information content and their possible role in the recognition. The features that proved to be vital and rich in their information were
extracted. New algorithms were developed for manipulating these information-rich features. Both hard-decision and soft-decision algorithms were devised to form articulatory-based features such as voicing, manner of articulation and place of articulation. These formed features could be used as additional inputs to the traditional classifiers (like Hidden Markov Models (HMMs) or Neural Networks), or they could be used directly to perform a completely knowledge-based (rule-based) recognition. Very good performance was obtained from these recognition experiments which proved the usefulness of these features in performing speaker-independent continuous speech recognition.

2. Brief Description of Results

Using the auditory-based system given in Fig. (1) as a front-end, acoustic-phonetic features were investigated for the automatic recognition of the fricative and stop consonants. The task was divided into 3 main parts:

1. Segmentation and categorization of continuous speech, in order to break down the speech signal into sonorants, fricatives (and affricates), stops and silences.

2. Classification of the fricative consonants.

3. Classification of the stop consonants.
A segmentation and categorization algorithm was developed to extract the stops and fricatives (and affricates) from continuous speech. The algorithm was tested on continuous speech of 30 speakers (not used in the design (training) process) from 6 different dialects from the TIMIT database. An overall accuracy of 92% was achieved, with 4% substitution error, 3% insertion error and 1% deletion error. The confusion matrix of the detected categories is given in table (1).

Table (1) Confusion matrix for the segmentation and categorization system. This table shows the substitution errors only. The accuracy is 96%.

<table>
<thead>
<tr>
<th></th>
<th>Stops</th>
<th>Fricatives</th>
<th>Sonorants</th>
<th>Silences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stops</td>
<td>92%</td>
<td>6%</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td>Fricatives</td>
<td>10%</td>
<td>86%</td>
<td>4%</td>
<td>0%</td>
</tr>
<tr>
<td>Sonorants</td>
<td>0.5%</td>
<td>0.5%</td>
<td>97%</td>
<td>2%</td>
</tr>
<tr>
<td>Silences</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>94%</td>
</tr>
</tbody>
</table>

After the extraction of the fricatives and the stops, a classification system is used to classify the different phonemes. The classification of the fricatives is divided into voicing detection and place of articulation detection. The results are shown in tables (2), (3) and (4).

Table (2). Confusion matrix for the voicing detection of fricatives. Accuracy is 95%.

<table>
<thead>
<tr>
<th></th>
<th>Detected as voiced</th>
<th>Detected as unvoiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced</td>
<td>186 (92%)</td>
<td>17 (8%)</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>9 (3%)</td>
<td>288 (97%)</td>
</tr>
</tbody>
</table>
Table (3). Confusion matrix for the place of articulation detection of fricatives. Correct recognition rate is 93%. (Alveolars: 92%, dentals: 95% and palatals: 93%).

<table>
<thead>
<tr>
<th></th>
<th>Detected as alveolar</th>
<th>Detected as dental</th>
<th>Detected as palatal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alveolars: /s/ and /z/</td>
<td>188 (92%)</td>
<td>11 (5%)</td>
<td>5 (3%)</td>
</tr>
<tr>
<td>Dentals: /f/, /v/, /th/, /dh/</td>
<td>2 (1%)</td>
<td>144 (95%)</td>
<td>6 (4%)</td>
</tr>
<tr>
<td>Palatals: /sh/ and /zh/</td>
<td>8 (6%)</td>
<td>2 (1%)</td>
<td>134 (93%)</td>
</tr>
</tbody>
</table>

Table (4). Confusion matrix for fricatives’ classification (in percentages). Overall correct recognition rate is 90%. Rows are inputs, columns are outputs.

<table>
<thead>
<tr>
<th></th>
<th>/s/</th>
<th>/f/, /th/</th>
<th>/sh/</th>
<th>/z/</th>
<th>/v/, /dh/</th>
<th>/zh/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/s/</td>
<td>90</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>/f/, /th/</td>
<td>4</td>
<td>87</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>/sh/</td>
<td>4</td>
<td>1</td>
<td>92</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>/z/</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>85</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>/v/, /dh/</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>/zh/</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>7</td>
<td>1</td>
<td>83</td>
</tr>
</tbody>
</table>

The classification of the stop consonants is also divided into voicing detection and place of articulation detection. The results are given in tables (5), (6) and (7).

Table (5). Confusion matrix for the voicing detection of stops. Accuracy is 97%.

<table>
<thead>
<tr>
<th></th>
<th>Detected as voiced</th>
<th>Detected as unvoiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced</td>
<td>428 (95%)</td>
<td>22 (5%)</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>15 (2%)</td>
<td>735 (98%)</td>
</tr>
</tbody>
</table>

Table (6) Confusion matrix for the place of articulation detection of stops. Accuracy is 90%.

<table>
<thead>
<tr>
<th></th>
<th>Detected as alveolar /t,d/</th>
<th>Detected as velar /k,g/</th>
<th>Detected as labial /p,b/</th>
<th>Detected as flap /dx/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alveolar (300)</td>
<td>91%</td>
<td>6%</td>
<td>3%</td>
<td>0%</td>
</tr>
<tr>
<td>Velar (300)</td>
<td>3%</td>
<td>88%</td>
<td>9%</td>
<td>0%</td>
</tr>
<tr>
<td>Labial (300)</td>
<td>6%</td>
<td>6%</td>
<td>86%</td>
<td>2%</td>
</tr>
<tr>
<td>Flap (300)</td>
<td>X</td>
<td>2%</td>
<td>4%</td>
<td>94%</td>
</tr>
</tbody>
</table>
Table (7) Confusion matrix for the overall classification of stops. Overall accuracy is 86%.

<table>
<thead>
<tr>
<th></th>
<th>Detected as /t/</th>
<th>Detected as /d/</th>
<th>Detected as /k/</th>
<th>Detected as /g/</th>
<th>Detected as /p/</th>
<th>Detected as /b/</th>
<th>Detected as /dx/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/t/</td>
<td>88%</td>
<td>3.5%</td>
<td>5%</td>
<td>0%</td>
<td>3.5%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>/d/</td>
<td>3%</td>
<td>89%</td>
<td>0%</td>
<td>7%</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>/k/</td>
<td>2.5%</td>
<td>0%</td>
<td>86.5%</td>
<td>1%</td>
<td>10%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>/g/</td>
<td>2%</td>
<td>2%</td>
<td>10%</td>
<td>76%</td>
<td>0</td>
<td>10%</td>
<td>0%</td>
</tr>
<tr>
<td>/p/</td>
<td>7%</td>
<td>0%</td>
<td>7%</td>
<td>0%</td>
<td>85%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>/b/</td>
<td>0%</td>
<td>5%</td>
<td>0%</td>
<td>5%</td>
<td>2.5%</td>
<td>86%</td>
<td>1.5%</td>
</tr>
<tr>
<td>/dx/</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
<td>0%</td>
<td>4%</td>
<td>94%</td>
</tr>
</tbody>
</table>

3. Contribution

Despite the recent successes in the Automatic Speech Recognition (ASR) field, the acoustic-phonetic characteristics of speech and their variability with context and speaker are not fully understood yet. More research is still needed to achieve a good understanding of this topic in order to build improved front-end processing systems that are able to extract the useful, information-rich, acoustic features. This knowledge is expected to have a profound effect on the automatic speech recognition systems whose performance can significantly improve by integrating more knowledge into their design.

Our work is concerned with this problem. We studied the acoustic-phonetic characteristics of fricative and stop consonants for continuous speech from multiple speakers with different dialects from the TIMIT database. The acoustic features described in our work and the algorithms developed to extract them are expected to contribute constructively to the acoustic characterization and the automatic recognition of continuous speech. Our work builds on the previous work in this area and introduces modifications and additions that resulted in performance improvements.
This work could be exploited in Hidden Markov Models (HMMs) or Artificial Neural Network (ANN) speech recognition systems. These systems could make use of the designed features in order to improve the front-end processing and create additional inputs to the data-driven (training-based) classifiers. These inputs are rich in information and incorporate considerable speech knowledge in their design. This is especially true with the soft-decision algorithms whose output is in the form of posterior probability estimates (certainty factors) which are compatible with HMMs and could be used as additional inputs that are rich in information and independent of speaker or context. They can also be used in a knowledge-based acoustic-phonetic speech recognition system by making a hard decision directly from the probability estimates.

The major contributions of this work could be summarized as follows:

1. A new method is proposed for integrating acoustic-phonetic knowledge in statistical speech recognition.

2. New techniques and design strategies are designed for the automatic knowledge-based segmentation, categorization and classification of continuous speech.

3. Acoustic-phonetic features which exist in the literature are tested and evaluated for their information content and their role in the automatic recognition of fricative and stop consonants. New features are proposed and tested.

4. New algorithms for the extraction, manipulation and combination of the acoustic features are designed and implemented. These algorithms combine several features in the decision making process.
5. An acoustic-phonetic recognition system, for the fricative and stop consonants, is developed and tested on the TIMIT database continuous speech of 30 speakers from 6 different dialects of American English. The segmentation and categorization has an accuracy of 92%. For fricatives, a recognition accuracy of 93% is achieved for the place of articulation and 95% for voicing detection. While for the stops, 90% accuracy is achieved for the place detection and 97% for the voicing detection. These results compare favorably with previous results as shown briefly in table (8). A more detailed comparison is given in [4,5,6]. Such improvement is mainly due to the use of auditory-based front-end processing, and new feature extraction techniques and manipulation algorithms, which integrate several acoustic properties in the decision making process.

<table>
<thead>
<tr>
<th></th>
<th>Best Previous</th>
<th>Our Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentation and Categorization</td>
<td>90% [67]</td>
<td>92%</td>
</tr>
<tr>
<td>Fricative Voicing</td>
<td>83% [95]</td>
<td>95%</td>
</tr>
<tr>
<td>Fricative Place</td>
<td>80% [55]</td>
<td>93%</td>
</tr>
<tr>
<td>Stops Classification</td>
<td>77%-82% [28,84]</td>
<td>86%</td>
</tr>
</tbody>
</table>

4. Publications and Reports

Publications:


**Technical Reports:**


5. References


