Use of Formants in Stressed and Unstressed Continuous Speech Recognition

D. Gharavian$^{1,2}$ & S.M. Ahadi$^1$

$^1$Electrical Engineering Department, Amirkabir University of Technology, Tehran, Iran
$^2$Power & Water Institute of Technology, Tehran, Iran

gharavian@pwit.ac.ir, sma@cic.aut.ac.ir

Abstract

Stress plays a crucial role in the understanding of speech by human listeners. However, automatic speech recognition results deteriorate in the presence of stress due to the change it causes in the speech parameters. Meanwhile, due to the vast diversity of the presence of stress in speech, a speech corpus that contains the majority of different stress conditions is difficult to obtain in real world. Therefore, other ways to improve stressed speech recognition performance have to be taken into account.

In previous works, we have evaluated the effects of stress on several speech parameters such as phone durations, pitch and formant frequencies. In this paper, the use of formants in stressed speech recognition will be discussed. We have found that formants and their dynamics (slopes) are useful in improving speech recognition rates both in stressed and unstressed conditions.

1. Introduction

Continuous speech recognition always faces several issues and problems. These problems get even worse when irregularities, such as stress and emotion appear in speech production. In this research, we tackle the stressed speech recognition problem.

Stress is known as an important prosodic feature. It usually points to some part of the speech that is pronounced with a different perceived loudness and causes an outstanding change in some prosodic parameters such as pitch or duration. Although, the effect of stress is spread over the whole word, phrase or even sentence, usually it is contained in a single syllable. Normally, the stressed syllable is easily distinguishable from others [1].

Stress plays an important role in directing the human listener to understand the speech [2]. Meanwhile, disregarding it can lead to deterioration in speech recognition system performance [3]. Some researchers have paid attention to this point and tried to improve the recognition rate using prosodic features. In some cases, the prosodic parameters are added to the speech feature vectors used in recognition [4], used in a post-processing phase [5], used in conjunction with the recognition system (to eliminate some paths in the Viterbi search) [6] or utilized in hybrid models or multi-level recognition [7]. All the above implementations have been proposed for the case of unstressed speech. As these approaches make use of prosodic parameters in speech recognition, they can also be considered in stressed speech recognition.

In our previous works, we have evaluated the phone durations and the effects of stress on duration, pitch and formant frequencies [8][9][10]. According to our previous statistical results on the formants, in this paper, we will discuss the effect of the use of formant frequencies in stressed and unstressed speech recognition. We have added formant frequencies in several different ways to the speech feature vectors and evaluated their effects on the recognition performance. The McCandless technique has been used for formant frequency extraction [11] throughout this work.

2. Recognition systems

The only commercially available speech corpus for Persian (Farsi) language is Farsdat [12]. This corpus has been used in this work to build our recognition systems. This corpus consists of 6000 sentences uttered by 300 speakers from both genders with different Iranian regional accents. Each speaker has uttered 20 sentences randomly selected from a set of 392 total available sentences. There is no particular stress on the utterances. 1800 sentences were chosen from speakers with standard Farsi accent (Tehran) for HMM model training purposes. 980 sentences from similar speakers were used for testing. The recognition systems were built using the HTK tools [13] for all standard tasks including MFCC feature extraction, model training and recognition. The systems consisted of monophone models and the tests were carried out without applying any language model.

In order to be able to add a stressed section to the corpus, all the 392 sentences of the corpus were read three times by a male speaker with no particular stress. Then, the same speaker has uttered 154 of those sentences applying stress on different points of these sentences. Altogether, 468 sentences with different stress conditions were generated. The data from this speaker have been used to test the models created using Farsdat data.

Table 1 lists the model sets created using different feature parameter combinations.

Table 1: List of model sets created using different feature parameter combinations.

<table>
<thead>
<tr>
<th>Models</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>C+LE+∆(C+LE)+∆(C+LE)</td>
</tr>
<tr>
<td>M2</td>
<td>C+LE+∆(C+LE)+∆(C+LE)+F1</td>
</tr>
<tr>
<td>M3</td>
<td>C+LE+∆(C+LE)+∆(C+LE)+F2</td>
</tr>
<tr>
<td>M4</td>
<td>C+LE+∆(C+LE)+∆(C+LE)+F3</td>
</tr>
<tr>
<td>M5</td>
<td>C+LE+∆(C+LE)+∆(C+LE)+F2/F1</td>
</tr>
<tr>
<td>M6</td>
<td>C+LE+∆(C+LE)+∆(C+LE)+F1+F2+F3</td>
</tr>
<tr>
<td>M7</td>
<td>C+LE+∆(C+LE)+∆(C+LE)+AF1</td>
</tr>
<tr>
<td>M8</td>
<td>C+LE+∆(C+LE)+∆(C+LE)+AF2</td>
</tr>
<tr>
<td>M9</td>
<td>C+LE+∆(C+LE)+∆(C+LE)+AF3</td>
</tr>
<tr>
<td>M10</td>
<td>C+LE+∆(C+LE)+∆(C+LE)+AF2+AF3</td>
</tr>
<tr>
<td>M11</td>
<td>C+LE+∆(C+LE)+∆(C+LE)+AF1+AF2+AF3</td>
</tr>
</tbody>
</table>

Table 1 lists the model sets created using different feature sets used in recognition experiments. C+LE points to cepstral plus log energy parameters and ∆ and ∆' refer to the first and second order dynamic parameters respectively. F1 to F3 are the first three formant frequencies. All these model sets were created using the above-mentioned Farsdat training data. In the following text, for
the sake of simplicity, the test section of the Farsdat corpus, the natural (unstressed) test data of the new speaker and the stressed data of the new speaker will be called D1N, D2N and D2S respectively.

3. Speech recognition using cepstral and energy parameters

The set of feature parameters usually used in speech recognition include cepstral and log energy parameters and their delta and acceleration over time. The results for model set M1 with D1N, D2N and D2S test data are reported in Table 2. These results will be used as a basis for evaluating the effectiveness of the models M2 to M11 throughout this paper.

<table>
<thead>
<tr>
<th>Test Data</th>
<th>No. of mixture components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>D1N</td>
<td>36.62</td>
</tr>
<tr>
<td>D2N</td>
<td>36.97</td>
</tr>
<tr>
<td>D2S</td>
<td>19.89</td>
</tr>
</tbody>
</table>

According to Table 2:
- The recognition rate for the new speaker is generally lower than that of the test section of the Farsdat speech corpus. This can be attributed to the difference in the environmental and recording conditions of the new speaker’s speech and that of the Farsdat corpus.
- The recognition rate of the system is significantly lower for the stressed speech in comparison to that of the unstressed (natural) speech. This is an indication of a radical change in cepstral parameters due to stress.

4. Speech recognition using cepstral features, energy and formant frequencies

Investigating the formant values in Farsdat corpus, we have found that the range of change in the formant frequencies is between 300-800Hz for F1, 1000-2000Hz for F2 and 2400-3500Hz for F3. Furthermore, the distribution of the values in this range have been found fairly regular. Therefore, this has stimulated the belief that the use of formant frequencies as extensions of the feature vector can help in improving system performance in recognizing stressed speech.

Another investigation has also shown that the formant frequency contours for different people, in similar sentences, follow similar patterns leading to valuable information in unstressed speech recognition. The similarities in formant contours in people with completely different formant frequencies indicate that formant slopes (rates of change in time) are even more useful, compared to the formant frequency values, in unstressed speech recognition.

In [10] the effects of stress on the formant frequencies and formant slopes were investigated. We found that the stress causes fairly orderly changes in the means of both formants and their slopes. Furthermore, for the formant slopes, the variance is changed almost in all cases. However, as the amount of change, in comparison to the changes in cepstral parameters, is small, we believe that these parameters, as more robust features to stress, can help in improving system performance in recognizing stressed speech.

In speech recognition, the ultimate goal is to extract a sequence of words (phones) from the speech data. Therefore, the phone type and its boundary are not known before recognition. Consequently, in order to be able to use formants as feature vector parameters, one must initially extract the formants for all the phones in the sentence, and use them in the whole sentence together with the other parameters. For the unvoiced consonants, where the formant values cannot be determined, a constant value can be used.

In the following experiments, the extra features (formants and their slopes) were added to the basic 39 parameter feature vectors used in M1. Due to space limitations, only the results of the first six Gaussians will be reported. However, in some special cases, changes due to the increase in the mixture component count will also be pointed out.

4.1. Formant frequencies

Model sets M2 to M6, as shown in Table 1, were created by adding the formant parameters to the basic feature vectors and using them in training the model sets. The results are reported in Tables 3, 4 and 5.

Table 2: Recognition results of the model set M1.

<table>
<thead>
<tr>
<th>Models</th>
<th>No. of mixture components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>M2</td>
<td>38.20</td>
</tr>
<tr>
<td>M3</td>
<td>37.27</td>
</tr>
<tr>
<td>M4</td>
<td>37.52</td>
</tr>
<tr>
<td>M5</td>
<td>37.35</td>
</tr>
<tr>
<td>M6</td>
<td>35.83</td>
</tr>
</tbody>
</table>

Table 3: Recognition results for D1N data using cepstral and formant parameters in the feature vector.

<table>
<thead>
<tr>
<th>Models</th>
<th>No. of mixture components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>M2</td>
<td>36.49</td>
</tr>
<tr>
<td>M3</td>
<td>36.51</td>
</tr>
<tr>
<td>M4</td>
<td>36.79</td>
</tr>
<tr>
<td>M5</td>
<td>36.95</td>
</tr>
<tr>
<td>M6</td>
<td>35.34</td>
</tr>
</tbody>
</table>

Table 4: Recognition results for D2N data using cepstral and formant parameters in the feature vector.

<table>
<thead>
<tr>
<th>Models</th>
<th>No. of mixture components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>M2</td>
<td>23.53</td>
</tr>
<tr>
<td>M3</td>
<td>22.40</td>
</tr>
<tr>
<td>M4</td>
<td>23.23</td>
</tr>
<tr>
<td>M5</td>
<td>25.33</td>
</tr>
<tr>
<td>M6</td>
<td>21.13</td>
</tr>
</tbody>
</table>

Table 5: Recognition results for D2S data using cepstral and formant parameters in the feature vector.

<table>
<thead>
<tr>
<th>Models</th>
<th>No. of mixture components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>M2</td>
<td>23.53</td>
</tr>
<tr>
<td>M3</td>
<td>22.40</td>
</tr>
<tr>
<td>M4</td>
<td>23.23</td>
</tr>
<tr>
<td>M5</td>
<td>25.33</td>
</tr>
<tr>
<td>M6</td>
<td>21.13</td>
</tr>
</tbody>
</table>

Comparing the results reported in Tables 3 and 4 to those reported in Table 2 for unstressed speech recognition:
- Due to the differences in recording conditions between Farsdat and the new speaker’s data, there are major differences in the features and parameters of the formants in the recognition of the data of the latter case is smaller. Meanwhile, the M5 model set has performed the best in this case. This means that F2/F1 is a robust feature to changes to the corpus.
- Looking at the statistical distributions of the formant parameters, we have found that they follow a close to normal distribution.
Gaussian distribution. Therefore, their effect on the recognition performance is noticeable mostly on the lower number of Gaussians.

- The use of several formant features, as in M6, has not led to improvements in either D1N or D2N.
- Investigating the results of larger numbers of Gaussians (not reported here), we have reached to an approximate conclusion that the highest impacts on recognition accuracy in unstressed speech recognition have been obtained by the formants F1, F2 and F3 in order. This can be justified according to this fact that the largest variance of the formant frequencies (or better said “standard deviation to mean ratio”) - SD/ Mean for the unstressed speech belongs to F1, F2 and F3 in order [10]. Therefore, larger numbers of Gaussians are more suitable for representing F1.

In stressed speech recognition, the comparison of the results of Table 5 to those of Table 2 clarifies that:

- All augmented-feature model sets (M2 to M6) positively affect the stressed speech recognition rates.
- Investigating the results of larger numbers of Gaussians (not reported here), we have found that the highest impacts on recognition accuracy in stressed speech recognition have been obtained by the formants F2, F3 and F1 in order. The same order can be seen in the amount of SD/ Mean for the stressed to unstressed case in [10]. Therefore, once again, we can conclude that larger numbers of Gaussians are more suitable for representing F2 in this case.
- Second to M3 is M5 in improving the stressed speech recognition rate. Therefore, as stated before, F2/F1 looks more robust in the situation of changing corpus.
- Another investigation on larger numbers of Gaussians (up to 15) uncovered the fact that, in larger mixture components, the positive effect of formant frequencies on stressed speech recognition, compared to M1, fades. A possible justification for this effect could be the close to Gaussian distributions of formant features.
- Once again, the use of several formant features, as in M6, has not only not led to any improvements in stressed speech recognition, but has also led to relative deterioration of the results.

An overall look at the Tables 4 and 5 tells us that the formant frequencies have larger impacts on stressed speech recognition compared to unstressed speech recognition. However, to be able to combine the benefits of all these features, a better approach, compared to the one used for M6, is required.

### 4.2. Formant slopes

Another set of parameters that can be used in improving stressed and unstressed speech recognition is the formant slopes. M7 to M11, as shown in Table 1, are the systems used for this purpose. Tables 6 to 8 summarize the results of both stressed and unstressed speech recognition using these feature sets.

According to Tables 6 and 7:

- The formant slopes are helpful in unstressed speech recognition.
- Due to the differences in recording conditions between Farsadat and the new speaker’s data, the effect of the formant slopes in the recognition of the unstressed speech of the latter case is smaller. The difference in the recognition accuracy increases with an increase in the number of Gaussians.
- As depicted in Figure 1, an overall increasing tilt is seen in the amount of improvement due to the use of formant slopes, for D2N (and D1N). In other words, formant slopes are better modeled with larger Gaussians.

### Table 6: Recognition results for D1N data using cepstral and formant slope parameters in the feature vector.

<table>
<thead>
<tr>
<th>Models</th>
<th>No. of mixture components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>M7</td>
<td>38.22</td>
</tr>
<tr>
<td>M8</td>
<td>38.00</td>
</tr>
<tr>
<td>M9</td>
<td>37.92</td>
</tr>
<tr>
<td>M10</td>
<td>37.92</td>
</tr>
<tr>
<td>M11</td>
<td>38.47</td>
</tr>
</tbody>
</table>

### Table 7: Recognition results for D2N data using cepstral and formant slope parameters in the feature vector.

<table>
<thead>
<tr>
<th>Models</th>
<th>No. of mixture components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>M7</td>
<td>36.84</td>
</tr>
<tr>
<td>M8</td>
<td>36.82</td>
</tr>
<tr>
<td>M9</td>
<td>36.94</td>
</tr>
<tr>
<td>M10</td>
<td>36.81</td>
</tr>
<tr>
<td>M11</td>
<td>37.33</td>
</tr>
</tbody>
</table>

### Table 8: Recognition results for D2S data using cepstral and formant slope parameters in the feature vector.

<table>
<thead>
<tr>
<th>Models</th>
<th>No. of mixture components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>M7</td>
<td>23.75</td>
</tr>
<tr>
<td>M8</td>
<td>23.93</td>
</tr>
<tr>
<td>M9</td>
<td>23.85</td>
</tr>
<tr>
<td>M10</td>
<td>23.83</td>
</tr>
<tr>
<td>M11</td>
<td>23.83</td>
</tr>
</tbody>
</table>

- Looking at the SD/ Mean ratio and the shape of the formant slope distributions, we have concluded that due to the larger SD/ Mean values for F1 and F2 slopes, in comparison to F3’s, and also the shape of their distribution functions, the first two formant slopes are better modeled in larger numbers of Gaussians (up to 15, results not included). Therefore, the results of their use with Larger Gaussians are better that that of F3’s.
- The improvements in the recognition results for M10 and M11, taking into account the enlargement of the feature vectors, are small. As stated for M6 in the case of using formant frequencies, the combined use of different formant-based features, as is done here, does not seem to be very useful and further research is needed in this case.

In Table 8, the following can be pointed out:

- The formant slopes lead to a noticeable improvement in stressed speech recognition.
- The increase in the size of feature vectors in M10 and M11 has smaller effect on the recognition rate, while M7 to M9 perform better in this case.
- All three formant slopes perform fairly similar in stressed speech recognition.
- According to Figure 2, the improvement caused by formant slopes in stressed speech recognition, for D2S, tends to decrease steadily with an increase in the number of Gaussians beyond 7. As shown in [10], the deviations in the distribution function (and also SD/ Mean ratio) in formant
slopes reduce after the application of stress. Therefore, smaller numbers of Gaussians can easily model the changes in the formant slopes, while for the case of unstressed speech, larger numbers of Gaussians worked better.

- Further investigations have shown that the combined use of formants and their slopes in the feature vector can deteriorate the system recognition performance in comparison to that of M2-M6 (results not reported).

\[ \text{Figure 1: The effect of increase in the number of Gaussians on the improvement in the recognition accuracy for D2N data.} \]

The results of Tables 3 to 8 indicate that the inclusion of formant slopes leads to more improvements on the recognition performance in comparison to the formant parameters with the highest amount of improvement witnessed in stressed speech recognition case.

While every formant frequency or every formant slope can lead to certain amount of improvement in speech recognition, the combined use of them has not even led to an amount of improvement equal to that of one of them. It appears that further research is needed to determine how the formant frequencies and their slopes should be combined to make better use of all the information available in these parameters.

Altogether, taking into account the effect of stress on formant parameters and their slopes, emphasized in [10], and also the nature of models built using unstressed data, these parameters seem to be the appropriate candidates for the normalization of stressed data toward unstressed data, or in other words, de-stressing the stressed speech. It appears that such an approach could lead to even better stressed speech recognition performance.

\[ \text{Figure 2: The effect of increase in the number of Gaussians on the improvement in the recognition accuracy for D2S data.} \]

5. Conclusions

In this research, the formant values and their slopes were added to the speech feature vectors applied to HMM and their positive impact on the recognition rate was evaluated. According to the research work done on the formant frequencies, we can conclude that the first to third formants and their slopes are somewhat effective in unstressed speech recognition and noticeably effective in stressed speech recognition. Among these features, the effect of formant slopes is more important in the recognition of both unstressed and stressed speech, while they are of higher importance in the stressed speech recognition. These results agree with previous results concerning the effect of stress on formant slopes. Obviously, further research is needed to achieve better stressed speech recognition results.

6. References