Amirkabir University of Technology
(Tehran Polytechnic)
Vol. 44, No. 2, Fall 2012, pp. 33- 45

Amirkabir International Journal of Science& Research
(Electrical & Electronics Engineering)
(AIJ-EEE)

Statistical Variation Analysis of Formant and Pitch Frequencies in Anger and Happiness Emotional Sentences in Farsi Language

D. Gharavian[1*]

1-Assistant Professor, Department of Electrical Engineering, Shahid Beheshti University, Tehran, Iran

ABSTRACT

Setup of an emotion recognition or emotional speech recognition system is directly related to how emotion changes the speech features. In this research, the influence of emotion on the anger and happiness was evaluated and the results were compared with the neutral speech. So the pitch frequency and the first three formant frequencies were used. The experimental results showed that there are logical and reasonable relations between the emotions and variations of these speech features. These results were also used to confirm our previous research about emotion recognition and emotional speech recognition.

KEYWORDS

Emotion, Emotion Recognition, Emotional Speech Recognition, Statistical Analysis, Formant and Pitch Frequencies.

* Corresponding Author, Email: D_Gharavian@Sbu.ac.ir

1- INTORDUCTION

Speech is important for human communication. Emotion is also the way for transferring extra information, besides textual information, to the listener. The change in the speech rate, volume, delay and other similar parameters make the emotions [1].

To carry out the Human-Machine Interaction (HCI), the computer should percept the affective states of human. Speech is a major tool in the HCI and occasionally, it is the only tool. Some examples for the HCI are phone channel [2], Interactive Voice Response (IVR) systems [3], call-center applications, intelligent automobile systems, interactive movie systems and so on. There have so far been many open problems to put into effect the practical HCI [4].

The linguistic and affective contents (statement, question, continuation and command) (happy, angry, sad, surprised, disgust, fear and sarcastic) may be conveyed by the prosody of the speech [5]. The machine learning and data mining techniques may be used for Emotion Recognition (ER) and Emotional Speech Recognitions (ESR). These applications are a way for HCI.

The extracted features from the emotional speech are deteriorated in comparison to the neutral speech. To characterize the effect of these variations, special techniques can be applied. This information might improve the performance of ER and ESR systems [6].

There are some individual features for each speaker that help to recognize different speakers from one another. These are the inter-speaker features, which are dependent on the environment, speaker's health or emotion [7].

One of the main problems for ER and ESR systems is the scarcity of training and testing data [1]. One sentence can be uttered to many emotional states. So recording the database with the most emotional states is impossible. Another way for improving the performance of ER or ESR systems is the recognition of variations due to emotion. This information can be used for designing and implementation of enhanced ER and ESR systems.

In this research, variation of pitch frequency, the first three formant frequencies and their slopes, as well as the direction of slopes for all voiced phoneme and phoneme group were evaluated. The logical relations caused by emotion in happiness and anger emotional states were extracted and the results were compared with those of the neutral state. The database was built by the speech of 24 male speakers.

The extracted statistical results were used to confirm the performance of ER and ESR systems. The results showed an interesting convergence between the statistical analysis of speech features in this study and the performance of these systems in our previous researches.

The reminder of this paper has been organized as follows: Section 2 reviews the background in this field. Section 3 presents the specification of the recorded database and how the statistical information was selected for the analysis. Sections 4 and 5 deal with the effect of emotion on pitch frequency and formant frequencies for all voiced phonemes. Section 6 presents the convergence between the statistical results obtained in Section 4 and 5 and the results of ER and ESR systems in our previous works. Finally, the discussion and conclusion are presented in Section 7.

2- BACKGROUD AND RELATED WORKS

A few dedicated published works about evaluating the influence of emotion on speech features for Farsi language are existed in the literature; some of the related works for other languages are reported.

Hagenaars and Minnen believe that autobiographical speech is slower and has lower pitch frequency than that of the script talking. They also mention that pitch variation is lower in fearfullness than the happy state. Twenty-five native Dutch women speaker were used for recording their database to avoid the probable difference because of gender or culture.

They found the rate of speech increases because of the arousal, and independent of the emotional states. The rate of speech is equal for fear and happy states. They also depict that the influence of fear on pitch frequency is more complicated to interpret [8].

Anton Batliner et. al., comparing the performance of ESR system using acoustic and linguistic features, showed that the performance of ER system using these two sets is comparable [2].

Lakshminarayanan et. al. used various spectral features for the detection of affective prosody such as happy, anger states, and linguistic prosody such as statement, question and continuation. Thirty-six native male American English speakers were used for recording the database.

They showed that higher frequency bands have more influence on stress pattern and lower frequency bands do the same on the intonation pattern [9].

Zhang et. al. depicted that intra-speaker variation is an important problem for the deterioration of performance in speaker identification applications in the forensic situations. In their research, the stability of acoustic features was evaluated for inter-speaker and intra-speaker variations. The results showed that the higher formant frequencies are more valuable than the lower formants for the speaker identification [7].

Steidl et. al. in their children's speech recognition system showed that how specific emotion influences the word accuracy. They reported various results about adapted models using neutral or emotional speech. They found that the performance of recognition system is better for anger or emphatic speech than neutral speech. They also saw noticed the distributing emotion states in mel frequency cepstral coefficients (MFCC) space using sammon transformation. Their database was constructed using the speech of 14 professional actors [10].

Toivanen et. al., applying the optimal combination of prosodic features, found that the performance of ER system is close to human's ER skill [11].

Pell et. al. used audience to recognize the emotion in four languages. Four native speakers uttered six emotional states. The experimental results showed that anger, sadness and fear states were recognized better than other states; however language and linguistic information could not affect the emotion recognition results [12].

Jong evaluated the effect of stress on vowel duration and quality for various positions of vowel in sentence for English language [13]. In the research works carried out on Farsi prosody so far, the effects of stress on the speech parameters and their use in the improvement of speech recognition were evaluated [14-19].

Some research results about ER and ESR systems published [20-27] by this author and his colleagues have been presented in Section 6 to confirm the statistical results of this paper. In this research, the statistical analysis of speech parameters such as formants and pitch frequency were carried out. Our results show that there is conformity between these analysis and ER and ESR of our previous results. In other words, in this research, we show that which features are better for ER and ESR systems and our previous results showed the validity of them.

The rest of paper is organized as follows: Section 2 reviews the related work. The stressed corpus and feature set are introduced in Section 3. The prosodic stress recognition method is presented in Section 4. The simulation and experimental results of using different investigated models for stress recognition are detailed in Section 5. The proposed combined classification method and its recognition results are presented in Section 6. Finally, Section 7 concludes the paper.

of variations of pitch and formant in adjacent frames were used. The slope of pitch and formants were shown with dF0 and dF1-dF3, respectively.

The speech emotion changes significantly the pitch and first three formants (F1-F3). In this study, the F0 and F1-F3 and their slopes values were determined at the phoneme level for each speaker to monitor the variation of pitch and formant frequencies variations in each emotional state compared with the neutral speech. The slope parameters contain extra information such as speech parameter variation rate and rate of the speech between adjacent frames.

The Hidden Markov Model (HMM) [29] was used for alignment of the phoneme boundaries.

To improve the time alignment accuracy, 15 mixtures were used in the HMMs. The models were generally five states left-right HMMs with no skip transitions. In labeling the neutral speech of a speaker, the baseline model was adapted by the neutral speech. The baseline model was trained using FARSDAT. Likewise, for the emotional speech alignment, the adapted model was readapted by using the emotional speech of the same speaker. This was performed for every emotional speech of all the speakers in our experiments.

After the time alignment, the time of start and end of each voiced phoneme in each recorded sentence for each speaker was extracted. The values of pitch and formant frequencies related to this time section contain the

3- SPEECH CORPUS AND TOOLS

The only available corpus of Farsi continuous speech is FARSDAT [28]. This corpus consists of 6000 sentences from 300 speakers (male and female), who have randomly uttered some of the 392 predefined available sentences. The sentences are normally uttered and do not contain any stressed parts. In this work, we used about 1800 of sentences for training the speech recognition system. For this purpose, the texts of 99 sentences from FARSDAT were selected as the base sentences. The emotional speech corpus was recorded using 24 speakers.

Each speaker uttered 252 sentences in three emotional states: neutral (N), happiness (H), and anger (A). The speakers were amateur and uttered each sentence several times from the template corpus. These two states were uttered by amateur speakers more naturally in comparison to other emotional states such as fear and surprise. Therefore, in this research, all statistical analysis was carried out for anger and happiness states. The emotional sentences with better quality were selected from the recorded sentences. The pitch frequency (F0) and three formant frequencies (F1-F3) were extracted from all the sentences in the database separately.

Besides the pitch and formants, their slopes were also used. Each slope contains the information of the envelope of pitch and formants. To calculate the slopes, the values statistical information. After this phase, for each voiced phoneme, the statistical information was used to evaluate the influence of emotional states on the speech parameters.

It is obvious that this information might be related to each phoneme in sentence that was uttered by different speakers. This variety of speech by position in sentence causes evaluation of this statistical information valuable. Comparison of the results of statistical study of this information in anger and happiness states with neutral state gives some new information about how emotion influences the speech features. The results were reported as the effect of emotional states on speech features for all of the voiced phonemes and phoneme groups.

Study the distribution of phonemes in FARSDAT database showed that about 80% of them are voiced phoneme and 20% unvoiced phoneme. Figure 1 explains how the phoneme groups spread in FARSDAT. In this figure, only the voiced phoneme groups are showed because the pitch and formant frequencies have only been defined for them.

There are 30 phonemes in Farsi language that were shown here according to the IPA standard. Twenty out of these phonemes are voiced phonemes and the rest are unvoiced.

In our statistical analysis, the data outside the mean ± standard deviation were deleted as irrelevant data.

All formants and pitch parameters were considered to have Gaussian distributions. In order to evaluate the appropriateness of this assumption, a χ^2 test of significance was carried out. All parameters resulted in χ^2 test level of 0.01. These results indicate that the Gaussian assumption is almost acceptable.

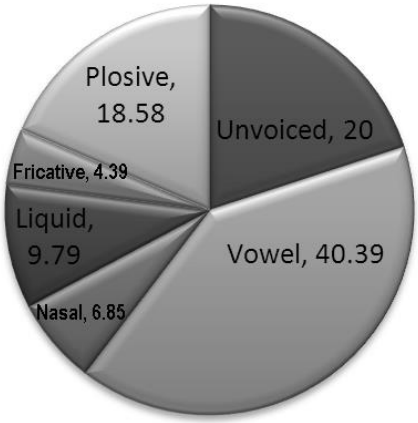

Figure 1: Distribution of phonemes in FARSDAT

4- PITCH FREQUENCY ANALYSIS

Figure 2 shows the mean of the pitch frequency of 20 voiced phonemes for neutral, anger and happiness states.

This figure shows some specific results:

In anger state, almost for all voiced phonemes, the mean of pitch frequency raises compared with the neutral state. Only for two voiced phonemes • and ○↗ (about 2% of the database), this relation does not meet.

The mean of pitch frequency for anger state is greater than the happiness state. There is one violation for /φ/ (about 2.5% of the database).

In most cases, the mean of pitch frequency for happiness state is greater than the neutral state. There are some violations for ○‗ • ‗ ℓ and /○↗/ (about 7% of the database).

Figure 3 shows the mean of pitch frequency for each phoneme group in various emotional states:

The maximum and minimum variations between neutral and anger states belong to the nasal and plosive states.

The maximum and minimum variations between the neutral and happiness states belong to the liquid and fricative state.

As shown in Figure 3, the maximum rate of variations in the anger and happiness states, compared with the neutral state, is about 15% and 5%, respectively.

Table 1 depicts the average direction of slope for pitch frequency in different emotional states. The average direction might be positive 'P' or negative 'N'. This table shows that:

For the first three emotional groups in the neutral state, the average of slope is negative and positive for the next two groups, respectively.

As shown in Figure 1, averagely, 60% of the slopes in neutral state are negative and 20% are positive.

In anger and happiness states, the average of slopes is negative.

Emotion decreases the mean of pitch frequency slope for anger and happiness states.

TABLE 1
THE PITCH FREQUENCY SLOPE VARIATIONS IN DIFFERENT PHONEME GROUPS FOR DIFFERENT EMOTIONAL STATES

Phoneme group / State	Vowel	Nasal	Liquid	Fricative	Plosive
Neutral	N	N	N	P	P
Anger	N↓	N↓	N↓	N↓	N↓
Happiness	N↓	N↓	N↓	N↓	N↓

P: Positive Slope N: Negative Slope

5- FORMANT FAEQUENCIESA ANALYSIS

Figures 4 to 9, depict the formant frequencies variations caused by emotion for each voiced phoneme and each phoneme group. It is obvious that the formant variations, because of emotion, are more complicated than the pitch frequency.

Studying the variation of formant frequencies caused by emotion in Figures 4 and 5 shows that:

Anger state raises the mean of the first formant for all voiced phonemes.

In anger state, the mean of the first formant frequency for each phoneme group raises in comparison with the neutral state.

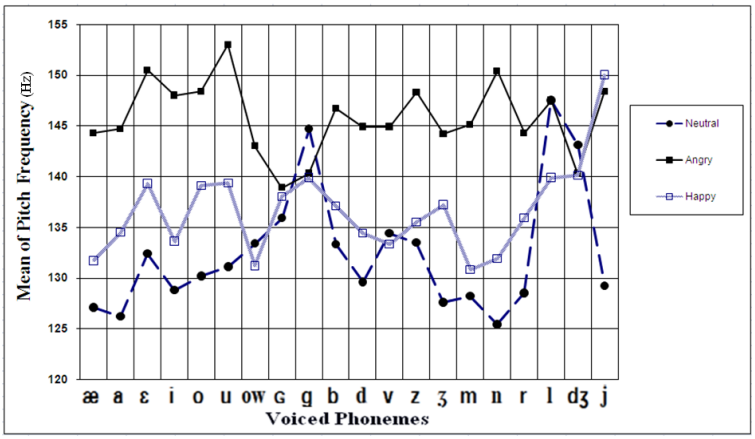

Figure 2: Pitch variations for each voiced phoneme in various emotional states

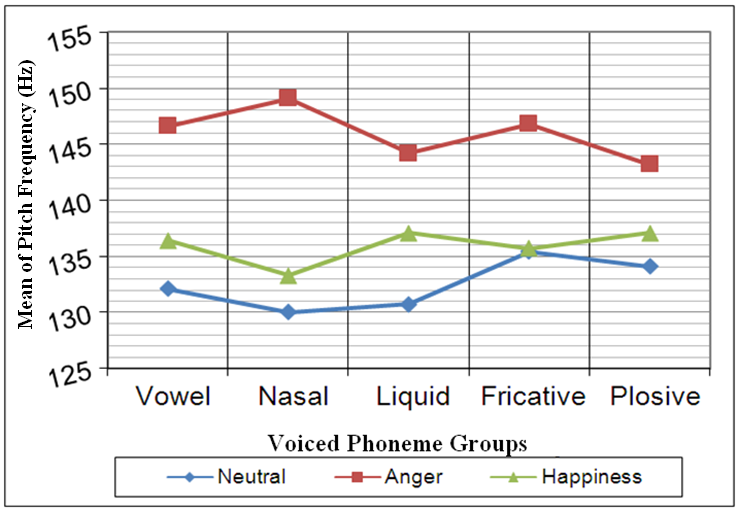

Figure 3: The pitch frequency variations in different phoneme groups and different emotional states

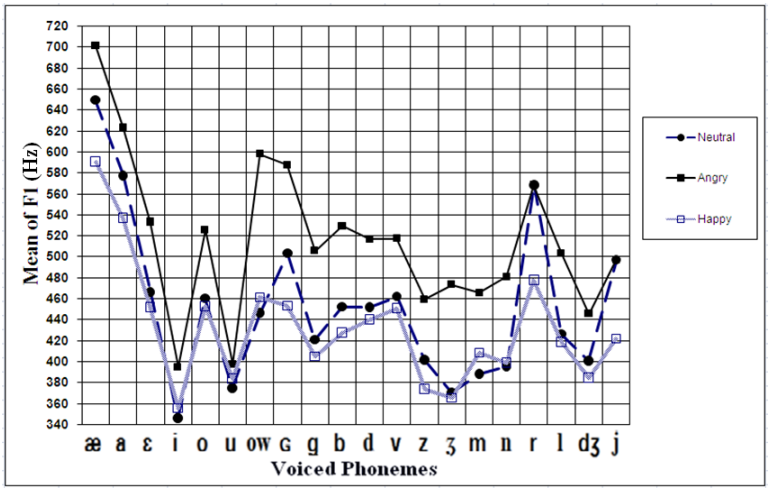

Figure 4: The first formant variations for different voiced phonemes in various emotional states

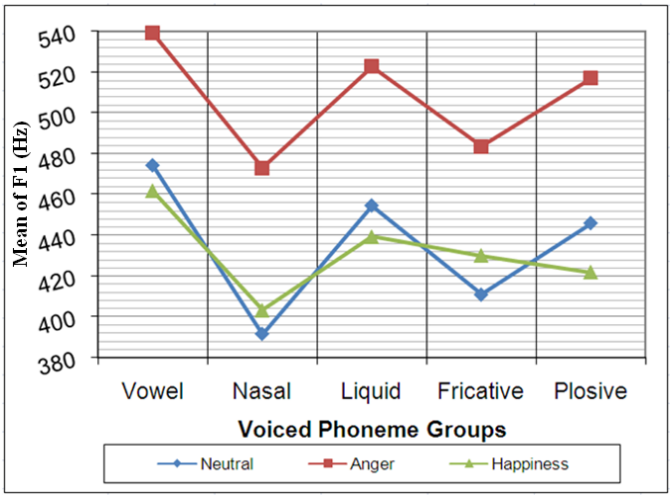

Figure 5: The first formant variations for different phoneme groups in various emotional states

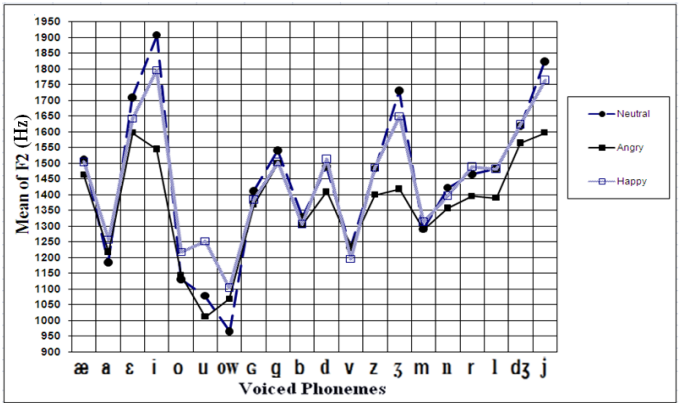

Figure 6. The second formant variations for voiced phonemes in different emotional states

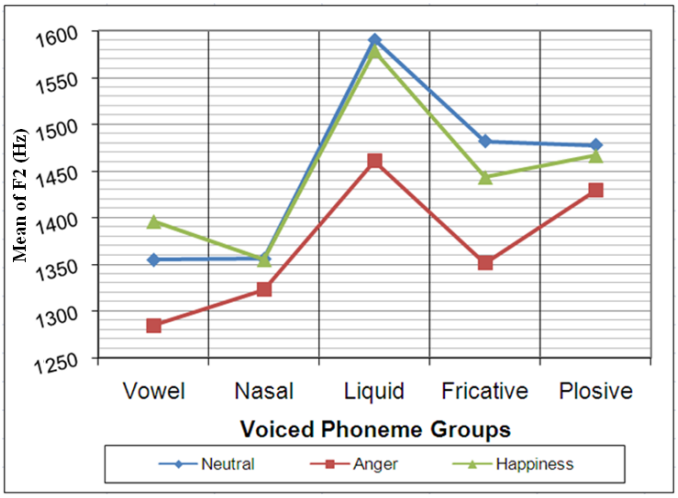

Figure 7: The second formant variations for different phoneme groups in various emotional states

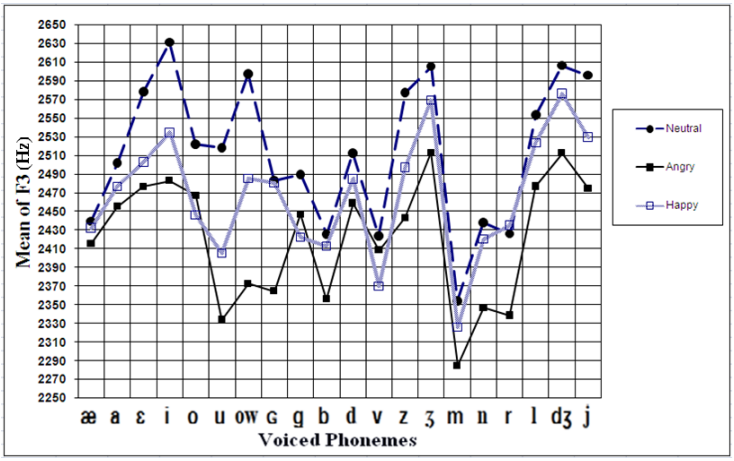

Figure 8: The third formant variations for different voiced phonemes in various emotional states

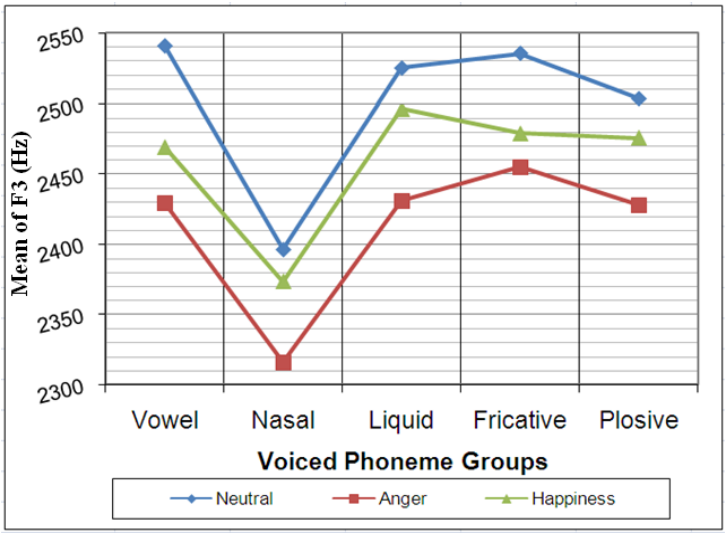

Figure 9: The third formant variations for different phoneme groups in various emotional states

In happiness state, the mean of the first formant frequency decreases compared with the neutral state. This is not true for /◌/,/◌̲/, /○◌̲/, /∓/ and /□/. These phonemes occupy only about 12% of the database.

In anger state, the mean value of the first formant is greater than the mean value of the happiness state.

For the nasal and fricative groups, the mean value of the first formant frequency in happiness state is greater than neutral state.

According to Figures 6 and 7:

In anger state, the average value of the second formant decreases compared to neutral state. There are consistencies for ↦ ○ ○◌̲ ↦□ ◌̲ These phonemes are about 14% of the database.

• In happiness states, the mean value of the second formant decreases compared to neutral state. There are consistencies for ↦ ○ ◌̲ ○◌̲ ◯ ∓ and /◌̲/. These phonemes occupy about 38% of the database.

In anger state, the mean value of the second formant is lower than that of happiness states. This is true for all voiced phonemes except for /◌̲/ (about 1.6% of the database).

In happiness state, the mean value of the second formant for vowels is greater than neutral state.

According to Figures 8 and 9, we can conclude that:

In anger state, the mean value of the third formant for all voiced phonemes decreases.

Regarding Figure 1, decreasing of the mean value of the third formant in anger state for all phoneme groups is

seen.

The mean value of the third formant frequency in happiness state is lower than neutral state. This is true for all voiced phonemes except for /_/ (about 7.5% of the database)

The mean value of the third formant frequency in anger state is lower than happiness state. This is true for all voiced phonemes except for ○ • ↦◯ → (about 7% of the database).

Tables 2, 3 and 4 show the mean values of the formants slopes caused by emotion.

Table 2 shows that:

In neutral state, the mean value of slope is positive only for the vowels.

Anger state changes the slope more positively for the vowels and more negative for the rest of voiced phonemes. Regarding Figure 1, roughly for 40% of the database, the mean value of slope is increased and for 40%, the slope decreases in anger state.

Happiness state makes the slope more complicated than anger state does.

In the happiness state, the mean value of the first formant slope is positive for the vowels and negative for the rest of voiced phonemes.

The mean values of slopes in happiness state are decreased for the vowels and liquids, and increased for the rest of voiced phonemes. Regarding Figure 1, for 50% of the voiced phonemes, the average value of slope is decreased, and for 30% of them, it is increased.

Anger and happiness states have the similar influence on liquid phonemes group.

TABLE 2
THE MEAN VALUE OF THE FIRST FORMANT SLOPE VARIATIONS IN VARIOUS EMOTIONAL STATES

Phoneme group / State	Vowel	Nasal	Liquid	Fricative	Plosive
Neutral	P	N	N	N	N
Anger	P↑	N↓	N↓	N↓	N↓
Happiness	P↓	N↑	N↓	N↑	N↑

P: Positive Slope N: Negative Slope

According to results of Table 3:

The mean value of slope for the second formant is negative for most of the phoneme groups except for the vowels and liquids. They occupy about 50% of the database.

Anger state decreases the mean value of the second formant slope for all phoneme groups.

Happiness state causes more complicated change in the second formant slope compared to the anger state.

The mean value of slope is positive for vowels and negative for the rest of voiced phonemes in the happiness state.

The mean value of slope decreases for vowels and liquids and increases for the rest of voiced phonemes compared to the neutral state. Happiness state changes the direction of slope for liquids.

The results given in Table 4 show that:

The mean value of slope for the third formant is negative for all phoneme groups except for vowels and fricatives. As shown in Figure 1, the slope is negative for 35% of phoneme groups and positive for 45% of them.

Anger state changes the average value of slope more negative for the voiced phonemes as compared to other emotional states. For all phoneme groups, the slope decreases.

For vowels and fricatives, the change in direction of slope happens for anger state only.

The behavior of slope for happiness state is similar to the anger state.

TABLE 3
THE MEAN VALUE OF THE SECOND FORMANT SLOPE VARIATIONS IN VARIOUS EMOTIONAL STATES

Phoneme group / State	Vowel	Nasal	Liquid	Fricative	Plosive
Neutral	P	N	P	N	N
Anger	P↓	N↓	P↓	N↓	N↓
Happiness	P↓	N↑	N↓	N↑	N↑

P: Positive Slope N: Negative Slope

Some general results from Table 2, 3 and 4 are:

The slope value of three formants in vowels, nasals and plosives is positive, negative and negative, respectively for the neutral state.

The slope value of three formants slope in nasals, fricatives and plosives for anger and happiness states is always negative.

TABLE 4
THE MEAN VALUE OF THE THIRD FORMANT SLOPE VARIATIONS IN VARIOUS EMOTIONAL STATES

Phoneme group / State	Vowel	Nasal	Liquid	Fricative	Plosive
Neutral	P	N	N	P	N
Anger	N↓	N↓	N↓	N↓	N↓
Happiness	N↓	N↓	N↓	N↓	N↓

P: Positive Slope N: Negative Slope

6- EMOTION RECOGNITION AND EMOTIONAL SPEECH RECOGNITION

One of the major objectives of this research was to show the convergence of the results of this research with our previous publications about emotion recognition and ESR systems. Some reports have been published for ESR system [20-22, 25-26] and for ER system [23-25, 27] by the author and his colleagues.

Our experimental study shows there is an interesting convergence between statistical analysis in this research and the performance of ER and ESR systems in our previous researches.

In our previous works, we used various features and methods to improve the performance of these systems. But to evaluate the mentioned convergence, in this work, we surveyed those results of the previous works that used only the pitch and formant features in the recognition system.

A. Analyzing the ER system's performance

Figure 10 shows the results of ER system. In this research, Gaussian Mixture Model (GMM) was used as a classifier. 5 The GMM's base models were trained using 12 MFCCs, energy and their first and second derivatives.

We named this model M0. Next, we added one of the first three formants as 40th feature to the feature vector and named these models M1 to M3. For the models M10 to M12, the 40th feature is the slope of one of the first three formants. These models can be used to estimate the influence of formants or their slopes on ER system's performance.

In the following, the converging results of the ER system are evaluated with the statistical results of this research. Generally, the features might be useful in ER that causes the maximum difference between emotional states. Figure 10 shows the results of ER system using M0, M1-M3 and M10-M12 models for neutral, anger and happiness emotional states that were reported in our previous work [23-25, 27].

Figure 10 shows the results of ER system using M10 to M12 models. We conclude some notes from this figure as:

The intense variation of the slope of formants in anger state causes these features to have the less effect on anger state recognition. The Std/Mean value for these features in anger state is great and, therefore, this information probably is not suitable to train GMM models.

The slope of the first formant is the most effective extra feature for ER system compared to the slopes of other formants. Tables 2, 3 and 4 show that the slope of the first formant has the highest difference in value and direction of variation for anger and happiness states.

The slope of the third formant has less influence in recognition of happiness state in comparison with the first formant because the variation of the third formant, due to emotion, is smaller than that of the first formant; therefore, the useful information for emotion recognition system in the third formant is less than the first formant.

The second formant's slope has the least effect on recognizing the happiness state. As depicted in Table 3, for vowels, the average value of the second formant's slope and its direction for anger and happiness states are similar. So using these features for ER raises confusion between happiness and anger states. According to Figure 1, the vowels are about 40% of the database; therefore, this problem might influence the ER system's accuracy.

Figure 10 shows that the slopes are not effective features for recognition of the neutral state. Regarding this figure, the first, third and second slopes of the formants have the most effective results, respectively. Our statistical evaluation results showed that, the Std/Mean value for the slopes of formants are greater than those for the formants' features. These variations in slope features caused by emotion make the neutral state to have higher miss-classification compared to the other emotional states.

B. Analyzing ESR system's performance

Figure 11 shows the results of emotional speech recognition system for anger and happiness states [25]. The HMM was used as a recognition tool. The base model was trained using 12 MFCCs features, energy and their first and second derivatives. The M1 to M6 models have the 40th extra feature that is one of the first three formant frequencies and their slopes, respectively.

Instead of GMM, usually HMM is used for speech recognition. Those features might be used for ESR systems that contain more information about emotional states to recognize each phoneme. Although slope features contain more information but GMM cannot model them significantly. In the first section of [25], the formant frequencies and their slopes were used for ESR. Here, the convergence between the results of ESR and statistical results of this paper are evaluated.

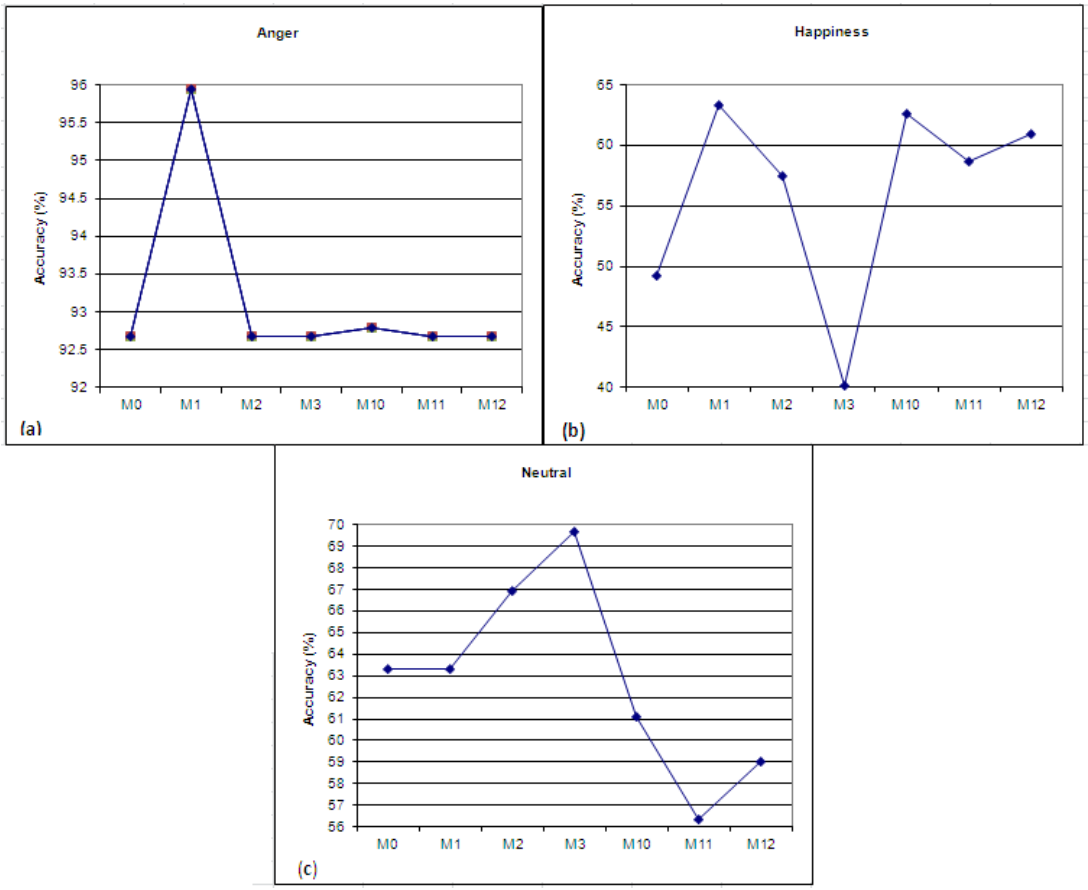

Figure 10: The effect of different formant features in emotion-state recognition; a) Anger, b) Happiness, c) Neutral

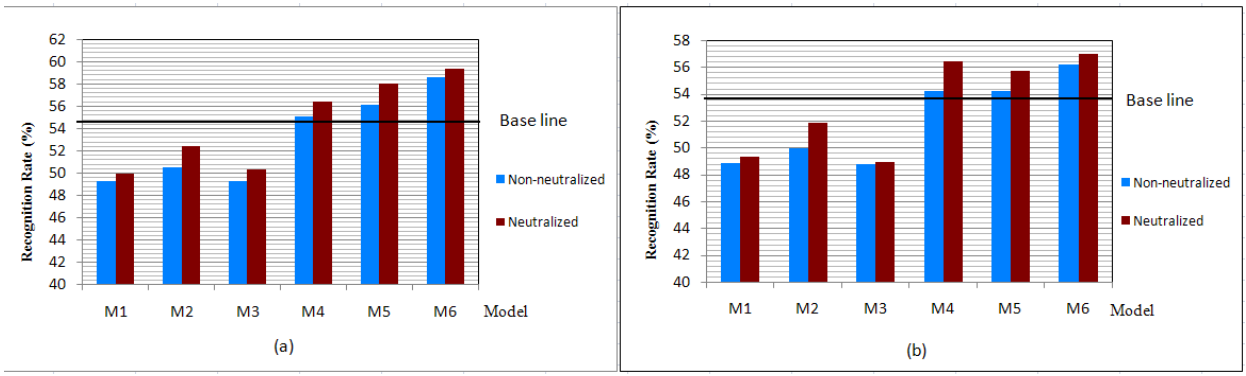

Figure 11: ESR system's performance for M_1-M_6 models using neutralized and non-neutralized formants: (a) Anger state, and (b) Happiness state.

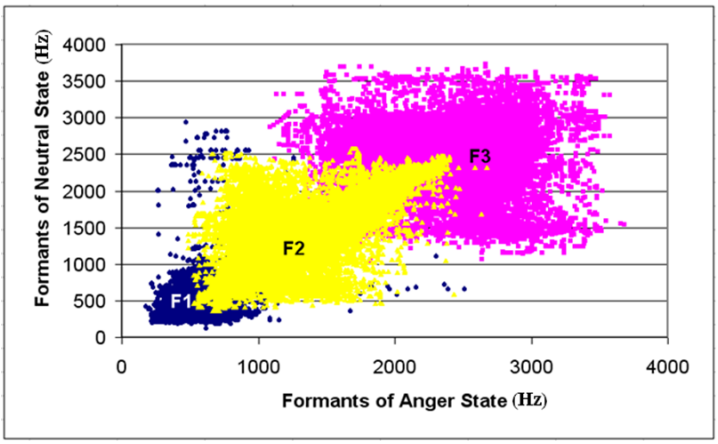

Figure 12: Variations of F_1-F_3 in neutral state compared with anger speech

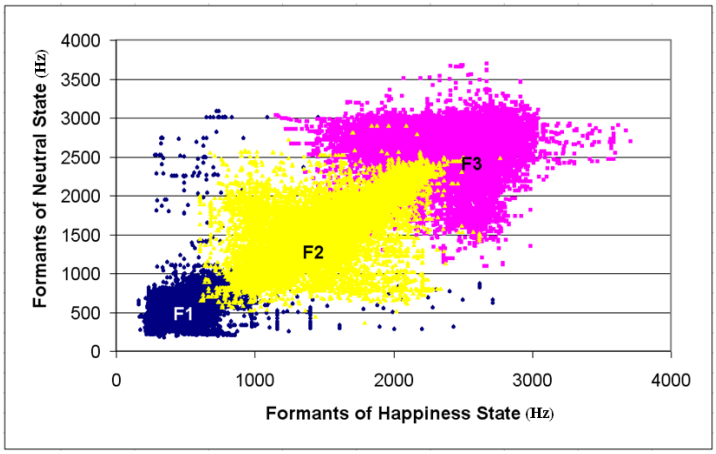

Figure 13: Variations of F_1-F_3 in neutral state compared with happiness speech

Slope features contain the information about the adjacent frames. So as shown in Figure 11, the slopes might improve the recognition performance. Figure 11 presents the ESR system's accuracy for anger and happiness states. The slope features have better performance for anger and happiness states. In this research, to improve the performance of ESR, the formant frequencies were neutralized.

The Second formant has more influence on improving the performance of ESR system. This result is just opposite to the results of the second formant for ER in the preceding section. The results showed that the second formant's variation caused by emotion was more regular than that of the other formants. So the HMM modeled this feature better than the other formants. Figures 12 and 13 show the formants variations caused by anger and happiness states compared to the neutral state. The regular variation of the second formant can be seen in these figures.

Figure 11 shows that the slopes of the first three formants are more efficient features for improving ESR accuracy than themselves.

For anger state, the ER results using M4 to M6 models showed that the third, second and first formants are the most effective features for ER, respectively. Tables 2-4 show that the third and second formants' slopes have the most regular variation (because of emotion) compared to the slopes of other formants', respectively. The order of the most effective features for ESR is opposite to those of the ER system. In other words, for ESR, the features with regular variations are more important than the features that have vast variations.

As depicted in the related tables for the happiness states, the third formant's slope has the most regular variation (caused by emotion) and, therefore, has the best results for ESR. Tables 3 and 4 show that the first and second formants' slopes have the similar variations and, therefore, their influence on ESR for happiness state is almost equal.

Table 5 shows the results of ESR using pitch frequency and its slope as an extra feature. M0 to M2 are the HMM recognition models. M0 is the baseline model that was trained using MFCC and energy and their first and second derivative features. M1 and M2 were trained using 39 base features and pitch or its slope as extra features, respectively.

TABLE 5
RECOGNITION PERFORMANCE IN DIFFERENT EMOTIONAL STATES USING M_0 TO M_3 MODELS

Models	Neutral	Emotional states	
		Angry	Happiness
M_0	52.10	6.04	20.57
M_1	55.26	18.03	23.70
M_2	61.43	13.15	28.40

According to the data given in this table, the following conclusion can be made:

The pitch frequency and its slope are effective features for ESR as extra features. Statistical analysis showed that these features have regular variations caused by emotion as depicted in Figures 2 and 3 and Table 3.

The pitch slope is more effective than pitch to improve ESR results. The only violation case is for the anger state. Our preceding statistical results showed that the pitch slope's variations in anger state are complicated and, therefore, it is difficult to train HMM models due to the small size of the database.

7- CONCLUSION

In this research, the influence of various emotional states on pitch frequency, the first three formants and their slope were evaluated. This evaluation was done for all of the voiced phonemes and phoneme groups, separately.

The results showed that, in most cases, it is possible to extract a logical relation between emotion and how it influences these features. These findings were used to confirm our previous results about the performance of ER and ESR systems.

8- REFRENCENS

[1] Rong, J., Li, G. and Chen, Y. P., "Acoustic Feature Selection for Automatic Emotion Recognition from Speech", Information Processing and Management, 45 (3), pp. 315- 328, doi:10.1016/j.ipm.2008.09.003, 2009.

[2] Batliner, A., Steidi, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L. and Amir, N., "Whodunnit- Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech", Computer Speech and Language, 25(1), pp. 4- 28, doi:10.1016/j.csl.2009.12.003, 2010.

[3] Polzehl, T., Schmitt, A., Metze F., and Wagner, M., "Anger Recognition in Speech Using Acoustic and Linguistic Cues", Speech Communication, 53 (9-10), pp. 1198- 1209, doi: 10.1016/j.specom2011.05.002, 2011.

[4] Bozkurt, E., Erdem, C. E., Erdem, A. T. and Erzin, E., "Formant Position Based Weighted Spectral Features for Emotion Recognition", Speech Communication Journal, 53(9-10), pp. 1186- 1197, 2011.

[5] Petridis, S., and Pantic, M., "Audiovisual Discrimination between Laughter and Speech", in Proc Int. Conf. on Acoustic, Speech and Signal Processing, pp. 5117- 5120, 2008.

[6] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C., "Automatic Speech Recognition and Speech Variability, A Review", Speech Communication, vol.49, pp. 763- 786, doi:10.1016/j.specom, 02, 006, 2007.

[7] Zhang, C., Weijer, J. V. D., Cui, J., "Intra- and Inter-Speaker Variations of Formant Patter for Lateral Syllables in Standard Chinese", Journal of Forensic Science International, 158 (2-3), pp. 117- 124, doi:101016/j.forsciint.2005.04.043, 2005.

[8] Hagenaars, M. A., and Minnen, A. V., "The Effect of Fear on Paralinguistic Aspects of Speech in Patients with Panic Disorder with Agoraphobia", In journal of Anxiety Disorder, 19(5), pp. 521- 537, 2005.

[9] Lakshminarayanan, K., Shalom, D. B., Wassenhove, V. V., Orbelo, D., Houde, J., and Poeppel, D., "The Effect of Spectral Manipulations on the Identification of Affective and Linguistic Prosody", Brain and Language, 84 (2), pp. 250- 263, 2003.

[10] Steidl, S., Batliner, A., Seppi, D., and Schuller, B., "On the Impact of Children's Emotional Speech on Acoustic and Language Models", EURASIP Journal on Audio, Speech and Music Processing, doi:10.1155/2010/783954, 2010.

[11] Toivanen, J., Vayrynen, E., and Seppanen, T. "Automatic Discrimination of Emotion from Spoke Finnish", Language and Speech Journal, 47 (4), pp 383- 412, doi:10.1177/00238309040470040301 2004.

[12] Pell , M. D., Paulmann, S., Dara, C., Alasseri, A., and Kotz, S. A., "Factors in the Recognition of Vocally Expressed Emotions: A Comparison of Four Languages", Journal of Phonetics, 37 (4), pp. 417- 436, 2009.

[13] Jong, K. D., "Stress, Lexical focus, and segmental focus in English: Patterns of Variation in Vowel Duration", Journal of Phonetics, 32 (4), pp. 493-

516, 2004.

[14] Gharavian, D., Sheikhzadeh, H. and Ahadi, S. M., "An Experimental Multi-Speaker Study on Farsi Phoneme Duration Rules Using Automatic Alignment", in Proc. 8th Australian International Conference on Speech Science and Technology, pp. 186-191, 2000.

[15] Gharavian, D. and Ahadi, S. M., "Statistical Evaluation of the Influence of Stress on Pitch Frequency and Phoneme Durations in Farsi Language", in Proc 8th European Conference on Speech Communication and Technology, pp. 1- 4, 2003.

[16] Gharavian, D. and Ahadi, S. M., "Evaluation of the Effect of Stress on Formants in Farsi Vowels", in Proc. 2004 International Conference on Acoustics, Speech, and Signal Processing, pp. 661- 664, 2004.

[17] Gharavian, D., "Prosody in Farsi Language and Its Use in Recognition of Intonation and Speech", PhD Thesis, Elec. Eng. Dept., Amirkabir University, Tehran, 2004.

[18] Gharavian, D. and Ahadi, S. M., "Use of Formants in Stressed and Unstressed Continuous Speech Recognition", in Proc. 8th International Conference on Spoken Language Processing, pp. 1- 4, 2004.

[19] Gharavian, D. and Ahadi, S.M., "Statistical Evaluation of Stress in Farsi and Its Effect on Vowel Pitch Frequencies, Durations and Energies", Amirkabir Scientific Research Journal, 15 (58-A), pp. 258- 268, Spring, 2004.

[20] Gharavian, D., Sheikhan, M. and Janipour, M., "Pitch in Emotional Speech and Emotional Speech Recognition Using Pitch Frequency", Majlesi Journal of Electrical Engineering, 4(1), pp. 19- 24, 2010.

[21] Gharavian, D. and Sheikhan, M., "Emotion Recognition and Emotion Spotting Improvement Using Formant-Related Features", Majlesi Journal of Electrical Engineering, 4(1), pp. 1- 8, 2010.

[22] Sheikhan, M., Gharavian, D. and Ashoftedel, F., "Using DTW-Neural Based MFCC Warping to Improve Emotional Speech Recognition", Neural Computing and Applications Journal, 21 (7), pp. 1765- 1773, doi: 10.1007/s00521- 011- 0620- 8, 2012.

[23] Gharavian, D., Sheikhan, M., Nazerieh, A. and Garoucy, S., "Speech Emotion Recognition Using FCBF Feature Selection Method and GA- Optimized Fuzzy ARTMAP Neural Network", Neural Computing and Applications Journal, 21(8), pp. 1-12, doi:10.1007/s00521- 011- 0643- 1, 2011.

[24] Gharavian, D. and Sheikhan, M., "GMM-Based Emotion Recognition in Farsi Language Using Feature Selection Algorithms", World Applied Science Journal, 14(4), pp. 626- 638, 2011.

[25] Gharavian, D., Sheikhan, M. and Ashoftedel, F., "Using Neutralized Formant Frequencies to Improve Emotional Speech Recognition", IEICE Electronic Express, 8(14), pp. 1155- 1160, 2011.

[26] Sheikhan, M., Bejani, M. and Gharavian, D., "Modular Neural-SVM Scheme for Speech Recognition Using ANOVA Feature Selection Method", Neural Computing and Applications Journal, pp. 1-13. doi:10.1007/s00521- 012- 0814-8, 2012.

[27] Gharavian, D., Sheikhan, M. and Ashoftedel, F., "Emotion Recognition Improvement Using Normalized Formant Supplementary Features by Hybrid of DTW-MLP-GMM", Neural Computing and Applications Journal, pp. 1-11, doi: 10.1007/s00521- 012- 0884- 7, 2012.

[28] Bijankhan, M., Sheikhzadegan, J., Roohani, M. R., Samareh, Y., Lucas, C. and Tebiani, M., "The Speech Database of Farsi Spoken Language", in Proc. 1994 5th Australian Int. Conf. on Speech Science and Technology, pp. 826- 83, 1994.

[29] Young, S. J., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P., The HTK Book (ver 3.2), Cambridge University Eng. Dept, 2002.