Pitch in Emotional Speech and Emotional Speech Recognition Using Pitch Frequency

D. Gharavian[1, 2], M. Sheikhan[2], M. Janipour[3]
1- Department of Electrical Engineering, Shahid Abbaspour University, Tehran, Iran
Email: gharavian@pwut.ac.ir
2- Department of Electrical Engineering, Islamic Azad University, South Tehran Branch, Tehran, Iran
Email: msheikhn@azad.ac.ir
3- Department of Electrical Engineering, University of Science and Technology, Tehran, Iran

Received: September 2009 Revised: November 2009 Accepted: January 2010

ABSTRACT:
The variations of speech parameters due to emotion or stress are noticeable. In the presence of such variations, if a neutral model is used for the system, the speech recognition accuracy deteriorates. The evaluation of how emotion influences speech parameters is the first step towards emotional speech recognition. Pitch frequency is an important parameter in speech processing systems. Therefore in this research the effect of pitch frequency and its slope due to emotion is explored for voiced phonemes. On the other hand, the influence of emotional state on continuous speech recognition performance is evaluated. The results show that the recognition performance of sentences with angry and happy states and also interrogative sentences has the most deterioration. This deterioration is more than 68% when compared to neutral speech recognition accuracy. To improve recognition results, we add the pitch frequency information to the end of speech recognizer feature vector. The amount of improvement depends on the type of emotion and also added pitch information. The results show that, pitch frequency slope has a significant affect on the improvement of speech recognition accuracy even for neutral speech.

KEYWORDS: Pitch frequency, Emotional States, Speech Recognition, Emotional Corpus, Slope.

1. INTRODUCTION

Speech is the main tool for human communication. Some factors such as the gender of the speaker, dialect, age, language, emotion, stress and many other factors can influence the features of speech [1]. All of the mentioned factors give additional information to the listener. But one of these factors, emotion, can deteriorate speech recognition accuracy very noticeably.

By using emotional states, speakers convey their intention to listeners. Obviously this information is more than textual information that exists in a sentence. Usually it is possible to use different emotional states in a sentence. It is well-known that a sentence without any emotional state can not transfer extra information between the speaker and the listener, although using emotion in speech leads to some problems for automatic speech recognition [1, 2].

In other words, the accuracy of speech recognizer, which uses baseline models (those trained using ordinary and non-emotional speech) deteriorates, because emotional states cause significant changes in speech parameters.

Variation of emotional speech is the largest problem for emotional speech recognition. On the other hand, constructing speech recognition models that contain various emotional states for each sentence is not trivial. Therefore, evaluation of how emotion influences speech parameters is the first step towards emotional speech recognition [3, 4].

Regular variations that can be found due to emotion are useful for enhancing speech recognition performance [5, 6]. Farsi is a pitch accent language; therefore, pitch frequency is one of the most important features in this language [7]. Accordingly in this paper, the influence of emotional states on pitch frequency and its slope is evaluated.

There are some solutions for emotional speech recognition. Developing an emotional speech corpus is the first approach [8]. However, preparation of a complete emotional corpus is not a simple task.

The second approach is parameter normalization [9-11]. The normalized emotional speech can be recognized with baseline models. Cepstral parameters are used as important features in speech recognition systems, but their variations due to emotion are very

complicated and normalization of cepstral features is also a difficult task [6].

Using a hybrid recognition system, which deploys an integrated two-step speech recognition system, is the third approach. In this system the prosodic parameters are used together with the basic speech recognition system [12, 13].

The last approach uses robust prosodic parameters, such as cepstral coefficients and log energy together [6, 14, 15]. This approach is effective if the emotional states don't have noticeable influence on selected prosodic parameters. In this case, it is likely that the speech recognizer has the ability to follow regular changes in prosodic parameters and also prevents probable deterioration of performance due to severe variations in cepstral parameters.

In this research, the last approach has been selected in which a large corpus is not needed.

2. SPEECH CORPUS AND TOOLS

FARSDAT is used as the baseline speech corpus in this work [16]. This is a continuous speech Farsi corpus including 6000 utterances from 300 speakers with various accents present in Iran. The above utterances have been read by different speakers from a set of 390 sentences. 1800 utterances have been selected from the mentioned utterances for training the recognizer models. These utterances are all with formal Farsi (Tehrani) accent. Meanwhile 890 sentences are used as the test corpus D_0.

The speeches of five male speakers were used for emotional corpus. Sentences of this corpus were selected from FARSDAT. For example, 99 sentences for neutral state, 69 sentences for happy state, 34 sentences for angry state, and 50 sentences for interrogative form. The speakers were trained to utter selected sentences for each emotional state. We call this corpus D_1.

Hidden Markov Models were created using HTK [17] and used as the speech recognizer. Pitch extraction was performed using the technique devised by Medan *et al.* [18]. The implementation available in the speech toolbox [19] as PDA (Pitch Detection Algorithm) and enhancement of pitch values [20] is utilized here.

The ordinary model has six mixture components and was trained using FARSDAT corpus. Speech recognizer is a continuous speaker independent system without language models.

3. EMOTIONS IN SPEECH

3.1. Influence of Emotion on Pitch Frequency and its Slope

In this section, the influence of emotion on pitch frequency and its slope is evaluated for voiced phonemes in Farsi language. To this purpose, precise time alignment is needed for sentences in each emotional state of speakers. To increase alignment accuracy, the baseline models, trained using FARSDAT training corpus, were adapted using neutral speech of each speaker. In alignment of emotional sentences, the adapted model was re-adapted using emotional sentences of each speaker. Experimental results show that in angry and happy sentences in Farsi language, all of the portions of the sentence are influenced by emotion. However in interrogative sentences; only the last section is influenced by emotion. To have a precise alignment for interrogative sentences, each sentence was listened and only the influenced portion by interrogative form was selected.

Fig. 1. shows the phoneme groups in Farsi language. The mean of pitch frequency and its slope are depicted for vowels, plosives, fricatives, liquids, affricatives and glides of voiced Farsi phonemes in Figs. 2 and 3. It is noted that pitch frequency slope was calculated for 10msec frames. The overall results for each voiced phoneme group are reported in Table 1.

As shown in Fig. 1, for most of the voiced phonemes in the angry state, the mean of pitch frequency is increased. Also for almost all of the voiced phonemes in the happy state, the mean of pitch frequency increases. The exceptions are /γ/, /δZ/, /ϖ/ , /l/ and /oω/ . And finally for the voiced phonemes in the interrogative state, the mean of pitch frequency is also increased, except for /γ/, /δZ/, /z/ and /ϖ/.

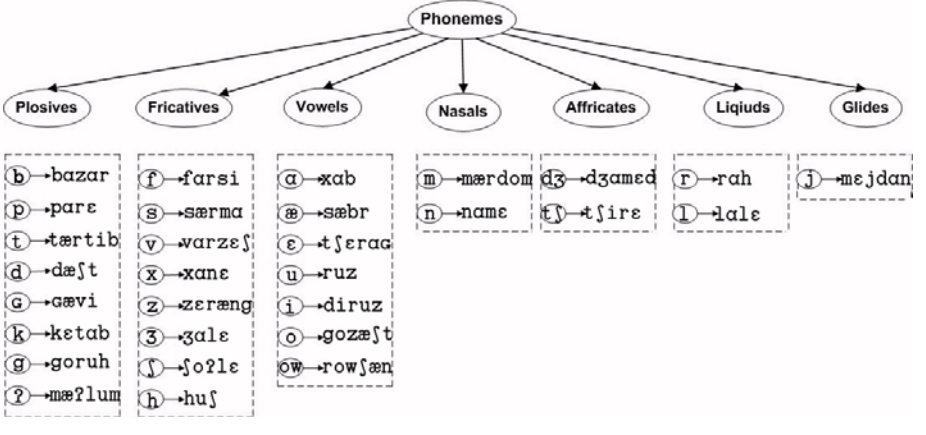

Fig. 1. The phoneme groups in Farsi language.

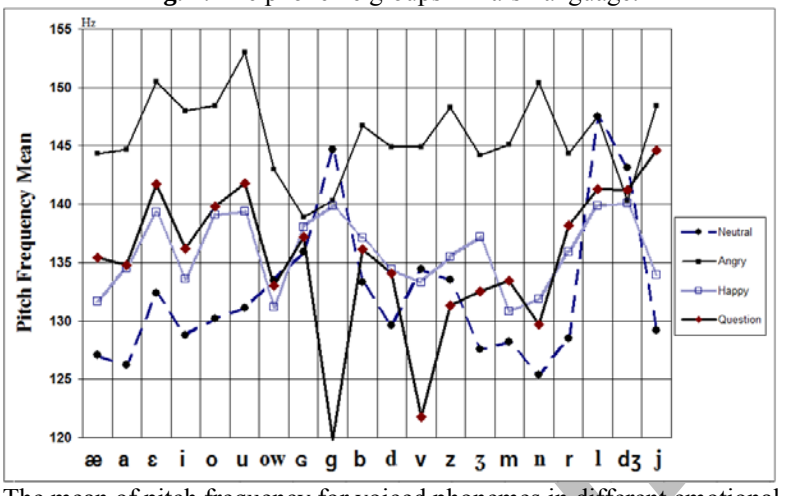

Fig. 2. The mean of pitch frequency for voiced phonemes in different emotional states.

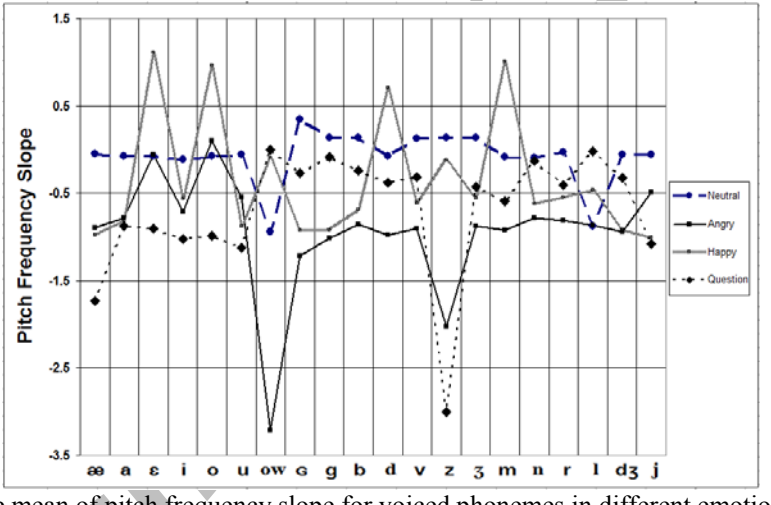

Fig. 3. The mean of pitch frequency slope for voiced phonemes in different emotional states.

Table 1. The mean of pitch frequency and its slope for various farsi phoneme groups in different emotional

Phonemes	Neutral		Angry		Happiness		Interrogative	
	Mean	Slope	Mean	Slope	Mean	Slope	Mean	Slope
Vowels	132.1	-0.21	146.6	-1.30	136.4	-0.18	135.6	-0.96
Plosives	134.1	0.10	143.2	-1.00	137.1	-0.83	138.8	-0.26
Fricatives	135.4	0.14	146.8	-1.27	135.7	-0.35	137.9	-1.90
Nasals	130.0	-0.08	149.1	-0.85	133.3	-0.51	133.3	-0.37
Liquids	130.7	-0.45	144.2	-0.84	137.1	-0.50	135.9	-0.22

So, emotion increases pitch frequency in most of the voiced phonemes. This is the rule achieved from the above results. The phonemes that don't convey this rule are /γ/, /δZ/, /ϖ/, /oω/, /ζ/, /l/ and /γ/. The relative numbers of these phonemes in FARSDAT are 1.37%, 0.89%, 1.57%, 0.59%, 1.85%, 1.98% and 1.37%, respectively [6]. Therefore, it is observable that for the majority of voiced phonemes this rule is applicable.

Statistical dispersion of pitch frequency slope is very high. Therefore, the evaluation is performed for each phoneme groups separately.

Vowels:

In neutral sentences, the mean of pitch frequency slope is negative for all the vowels.

The pitch slope is also negative in the angry state, except for /o/. As another rule, the absolute values of pitch slope are also increased. These rules are applicable for the happy state; except for /o/ αvδ /E/. Vowels in interrogative sentences also convey these rules.

Nasals:

The statistical mean of the pitch frequency slope for all emotional states is negative, except for /μ/ in the happy state. The absolute values of pitch frequency slope are also increased.

Voiced Affricatives, Glides and Voiced Fricatives:

The statistical mean of the pitch frequency slope is always negative and for emotional phonemes is even more negative.

Voiced Plosives:

The statistical mean of the pitch frequency slope is positive, except for /δ/, and are increase in the degree of emotion makes it even more negative.

Liquids:

The statistical mean of the pitch frequency for phonemes in this group is negative for the neutral state. The slope for /ρ/ becomes more negative and the slope of /λ/ becomes more positive in emotional sentences.

The values reported in Table 1 show that the mean of pitch frequency increases in emotional sentences and the amount of this increase depends on the types of the phoneme group and the emotion. Angry and interrogative sentences have the most variation.

The pitch frequency slope is also negative for neutral and emotional sentences. Emotionalizing the sentences makes it more negative except for vowels in the happy state. The amount of this variation depends on the type of emotion, but the angry state makes the most variation.

The overall results show that there is a regular change in the pitch frequency and its slope due to emotion. Therefore, this regular variation can help emotional speech recognition systems to improve their accuracy. In the next step, we use pitch frequency and its slope as additional features in emotional speech recognition systems.

3.2. Emotional Speech Recognition

3.2.1. Neutral Speech

To apply the pitch frequency features, the value of pitch frequency, its slope and PSSZN (Per-Story-Syllable-Z-Normalization) [21] are used as the pitch features in this research . These features are added at the end of the speech recognizer feature vector. To find PSSZN, we use Eq. (1):

$$p_n = \frac{(f_n - \mu)}{\sigma} \quad (1)$$

In this equation, f_n is the pitch value for frame n, μ and σ are the mean and standard deviations of the pitch in the utterance, respectively. We also use Eq. (2) to find the pitch frequency slope for each frame:

$$d_t = \frac{\sum_{n=1}^{\theta} n(f_{t+n} - f_{t-n})}{2\sum_{n=1}^{\theta} n^2} \quad (2)$$

The original feature vector was made of 12 cepstral coefficients and the logarithm of energy, their delta and delta-delta parameters. Therefore, this feature vector has 39 components. After extracting one of the pitch features, this feature is added to the end of the main feature vector. The following results show that adding a pitch frequency feature to the end of feature vector can improve speech recognizer performance in compare to the baseline model.

In the first step, models M_0 to M_3 were made and trained with the features shown in Table 2, using FARSDAT training corpus.

Table 2. M_0 to M_3 recognition models and their corresponding features

Models	Features
M_0	C+LE+Δ(C+LE)+Δ^2(C+LE)
M_1	C+LE+Δ(C+LE)+Δ^2(C+LE)+F_0
M_2	C+LE+Δ(C+LE)+Δ^2(C+LE)+ΔF_0
M_3	C+LE+Δ(C+LE)+Δ^2(C+LE)+PSSZN

In this table, C represents cepstral coefficients C_0 to C_{12}, LE is the logarithm of energy, Δ and Δ^2 are delta and delta-delta parameters. To enable comparison of the recognizer performance for FARSDAT with emotional corpus, Table 3 depicts recognition results for FARSDAT using M_0 to M_3 models.

Table 3. M_0 to M_3 recognition models and their corresponding features

Models	Accuracy
M_0	78.02
M_1	75.48
M_2	80.93
M_3	79.66

These results show that the pitch frequency slope and PSSZN improve neutral speech recognition accuracy, whereas the addition of pitch frequency to the feature vector does not. It is noted that pitch frequency is a gender dependent parameter. The FARSDAT speakers are from both genders and the recognizer models are trained with both of them. Therefore, it is predictable that pitch frequency cannot improve speech recognizer performance. These results also show that slope of pitch frequency and PSSZN are nearly gender independent features, because in calculation of these two features, the effect of pitch mean is eliminated.

3.2.1. Emotional Speech

In this section, the effect of emotion in the performance of speech recognition systems is evaluated. As mentioned before, our emotional corpus was made of five speakers in three states: angry, happy and interrogative.

The recognition rates for different emotional states using M_1 to M_3 models are reported in Table 4, respectively.

Table 4. Recognition rates in different emotional states using M_0 to M_3 models

Models	NEUTRAL	INTERROGATIVE	EMOTIONAL STATES	
			Angry	Happiness
M_0	52.10	35.57	6.04	20.57
M_1	55.26	42.19	18.03	23.70
M_2	61.43	45.15	13.15	28.40
M_3	56.63	43.59	15.08	24.31

We can conclude the following results from this table:

- Adding pitch frequency, its slope and PSSZN to the end of feature vector can improve speech recognizer performance.
- For neutral speech, considering the pitch frequency slope causes the most improvement in the recognizer performance (relatively 17.9%). As depicted earlier, pitch frequency cannot make notable improvement in speech recognition accuracy (relatively 6% compared to M_0 model). We also conclude that PSSZN improves recognition rates about 8.6%. Therefore, the pitch frequency slope and PSSZN cause more improvement in speech recognizer performance compared to the pitch frequency.
- In the angry state, pitch frequency is the most effective parameter among the other pitch features. The angry state has vast influence on speech parameters. It is obvious that the pitch frequency slope is influenced more than the pitch frequency. Therefore, it is predictable that pitch frequency slope cannot improve the speech recognizer performance as much as the pitch frequency. In the angry state, the recognition rates are not noticeable but these results depict that in the best case, using M_1 model, the accuracy can be improved by about 198% relatively.
- For the angry state, M_1 and M_3 models have more improvement in the speech recognizer performance.
- In the happy state, pitch frequency slope is the most effective added pitch parameter. As it was mentioned in the last section, for the happy state, pitch frequency slope variation is more regular than the angry state. Therefore, in the neutral case, the pitch frequency slope (M_2 model) has the most improvement on speech recognition accuracy (38% relative to M_0).
- M_2, M_3 and M_1 models have relative improvement about 26.9%, 22.5% and 18.6% compared to M_0, respectively.

The overall results depict that adding one parameter as 40th feature to the end of feature vector improves the recognizer performance.

In the next step, M0 to M2 models were adapted with the neutral speech of each speaker. HTK tool was used to do this adaptation. Table 5 shows emotional speech recognition accuracy using these models.

In Fig. 4, the recognition accuracy for the three emotional states using M_0 to M_2 models and adapted models are compared. In this figure, recognition rates with adapted models are represented with "_A" label. These results depict that if the recognition models are adapted with the neutral speech of each speaker, the speech recognizer performance for emotional speech will be improved significantly. Using this adaptation, recognition rates for the angry state become similar to other rates.

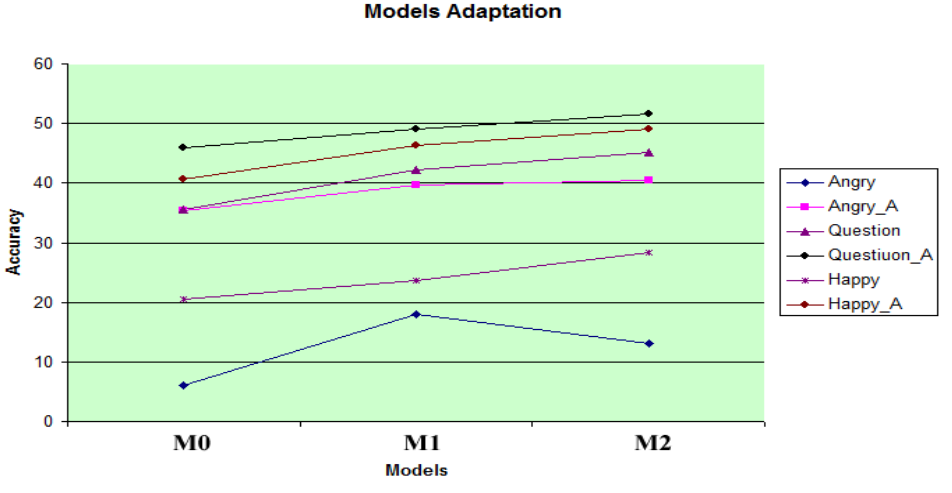

Fig. 4. Speech recognition accuracy of various emotional states using M_0 to M_2 models and adapted models

Table 5. Emotional speech recognition results for five speakers in various emotional states using adapted models M_0 to M_2

Models	Angry	Happiness	Interrogative
M_0	35.41	40.67	45.94
M_1	39.65	46.28	49.01
M_2	40.43	48.99	51.59

4. CONCLUSION

In this research, the influence of pitch frequency parameters has been evaluated as an additional feature to the end of an ordinary speech recognizer's input feature vector for emotional speech recognition. The speech recognizer was trained using neutral speech of FARSDAT corpus. The results depict that the performance of emotional speech recognition using ordinary models is degraded. By utilizing additional features such as the pitch frequency, its slope and PSSZN for training new models, the speech recognizer performance is improved.

REFERENCES

[1] Wang C., and Seneff S.; **"Robust Pitch Tracking for Prosodic Modeling in Telephone Speech"**, in Proc. .Int. Conf. .Acoustics, Speech, and Signal Processing, Vol. 3, pp. 1343 - 1346, (2000)

[2] Huang X., Acero A. and Hon H-W.; **Spoken Language Processing, A Guide to Theory, Algorithm, and System Development**, Prentice Hall, (2005)

[3] Shneilerman B.; **"The Limits of Speech Recognition"**, *Communication of The ACM*, Vol. 43, pp. 63 - 65, (2000)

[4] Yuan J., Shen L. and Chen F.; **"The Acoustic Realization of Anger, Fear, Joy and Sadness in Chinese"**, in Proc. 7th Int. Conf. Spoken Language Processing (ICSLP'02), pp. 2025 - 2028, (2002)

[5] Yuan J., Shih C. and Kochanski G.P.; **"Comparison of Declarative and Interrogative Intonation in Chinese"**, in Proc. Int Conf Speech Prosody, Aix-En-Provence, pp. 711 - 714, (2002)

[6] Gharavian D.; **Prosody in Farsi Language and Its Use in Recognition of Intonation and Speech**, Ph.D Thesis, Electrical Engineering Department, Amirkabir University of Technology, Tehran (In Farsi), (2004)

[7] Almasganj F.; **Structural Analysis of Farsi Language Using Prosodic Information of the Speech Signal**, Ph.D Thesis, Tarbiat Modarres University, Tehran (In Farsi), (1998)

[8] Athanaselist T., Bakamidis S., Dologlou I., Cowie R., Douslas E. and Cox C.; **"ASR for Emotional Speech: Clarifying the Issues and Enhancing Performance"**, *Journal of Neural Networks*, Vol. 18, pp. 437 - 444, (2005)

[9] Liu J., Zheng T.F. and Wu W.; **"Pitch Mean Based Frequency Warping"**, in Proc. Int. Symp. .Chinese Spoken Language Processing (ISCSLP'06), Vol. I, pp. 87 - 94, (2006)

[10] Cui X. and Alwan A.; **"MLLR-Like Speaker Adaptation Based on Linearization of VTLN with MFCC Features"**, in Proc. 9th Europ. Conf. Speech Communication and Technology (EUROSPEECH'05), pp. 273 - 276, (2005)

[11] Schuller B., Müller R., Eyben F., Gast J., Hornler B., Wollner M., Rigoll G., Hothker A. and Konosu H.; **"Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application"**, *Image and Vision Computing*, Vol. 27, pp. 1760 - 1774, (2009)

[12] Magimai M., Stephenson T.A. and Bourlard H.; **"Using Pitch Frequency Information in Speech RecognitionC"**, in Proc. 8th Europ. Conf. .Speech Communication and Technology (EUROSPEECH'03), pp. 2525-2528, (2003)

[13] Hirose K. and Iwano K.; **"Detection of Prosodic Word Boundaries by Statistical Modeling of MORA Transitions of Fundamental Frequency Contours and its Use for Continuous Speech Recognition"**, in Proc Int. Conf. .Acoustics, Speech, and Signal Processing, Vol. 3, pp. 133 - 136, (2000)

[14] Hung H.C.H. and Seide F.; **"Pitch Tracking and Tone Features for Mandarin Speech Recognition"**, in Proc. Int. Conf. .Acoustics, Speech, and Signal Processing, Vol. 3, pp. 1523 - 1526, (2000)

[15] Cernak M., Benzeghiba M. and Wellekens C.; **"Diagnostics of Speech Recognition: On Evaluating Feature Set Performance"**, in Proc. 12th Int. Conf. Speech and Computer (SPECOM'07), Vol. 1, pp. 188 - 193, (2007)

[16] Bijankhan M., Sheikhzadegan J., Roohani M.R., Samareh Y., Lucas C. and Tebiani M.; **"The Speech Database of Farsi Spoken Language"**, in Proc. 5th Australian Int.Conf. .Speech Science and Technology (SST'94), pp. 826-831, (1994)

[17] Young S.J., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V. and Woodland P.; **The HTK Book**, Cambridge University, (2001)

[18] Medan Y., Yail E. and Chazan D.; **"Superresolution Pitch Determination of Speech Signals"**, *IEEE Trans. on Signal Processing*, Vol. 39, pp. 40 - 48, (1991)

[19] Edinburgh Speech Tools Library, Available: http://estvox.org/docs/speech-tools-1.2.0/x2152.htm.

[20] Vepreka P. and Scordilis M.S., **"Analysis, Enhancement and Evaluation of Five Pitch Determination Techniques"**, *Journal of Speech Communication*, Vol. 37, pp. 249 - 270, (2002)

[21] Surendran D.; **Analysis and Automatic Recognition of Tones in Mandarin Chinese**, Ph.D Thesis, University of Chicago, (2007)