BASIC ANALYSIS ON PROSODIC FEATURES IN EMOTIONAL SPEECH

X.Arputha Rathina[1], K.M.Mehata[1], M.Ponnavaikko[2]

[1] Department of Computer Science and Engineering, B.S.Abdur Rahman University, Vandalur, Chennai, India.
[2] SRM University, Kattankulathur, Chennai, India.

ABSTRACT

Speech is a rich source of information which gives not only about what a speaker says, but also about what the speaker's attitude is toward the listener and toward the topic under discussion—as well as the speaker's own current state of mind. Recently increasing attention has been directed to the study of the emotional content of speech signals, and hence, many systems have been proposed to identify the emotional content of a spoken utterance.

The focus of this research work is to enhance man machine interface by focusing on user's speech emotion. This paper gives the results of the basic analysis on prosodic features and also compares the prosodic features of, various types and degrees of emotional expressions in Tamil speech based on the auditory impressions between the two genders of speakers as well as listeners. The speech samples consist of "neutral" speech as well as speech with three types of emotions ("anger", "joy", and "sadness") of three degrees ("light", "medium", and "strong"). A listening test is also being conducted using 300 speech samples uttered by students at the ages of 19 -22 the ages of 19-22 years old. The features of prosodic parameters based on the emotional speech classified according to the auditory impressions of the subjects are analyzed. Analysis results suggest that prosodic features that identify their emotions and degrees are not only speakers' gender dependent, but also listeners' gender dependent.

1. INTRODUCTION:

One can take advantage of the fact that changes in the autonomic nervous system indirectly alter speech, and use this information to produce systems capable of recognizing affect based on extracted features of speech. For example, speech produced in a state of fear, anger or joy becomes faster, louder, precisely enunciated with a higher and wider pitch range. Other emotions such as tiredness, boredom or sadness, lead to slower, lower-pitched and slurred speech. Emotional speech processing recognizes the user's emotional state by analyzing speech patterns.

Previous researches have engaged in the study of emotions, in how to recognize them automatically from speech and they have tried to incorporate this technology into real world applications. However, it has been difficult to find the features that describe the emotional content in speech. It has not been reached a reliable set of features for discriminating emotional states in spontaneous speech [1]. We can find information associated with emotions from a combination of prosodic and spectral information. Much of the work done to date has been focused on features related to prosodic aspects.

As information communication technology (ICT) advances, there are increasing needs for better human-machine communication tools. Expressive speech is more desirable than non-expressive speech as a means of man-machine dialog. However, the capability of synthesizing expressive

DOI : 10.5121/ijcsea.2012.2410

speech including emotional speech is currently not high enough to match the needs. We have to explore features from natural speech to achieve a method for a variety of expressive-speech synthesis. Among expressive speech, we have so far placed a focus on emotional speech.

New applications such as speech-to-speech translation, dialogue or multimodal systems demand for attitude and emotion modeling. Humans would choose different ways to pronounce the same sentence depending on their intention, emotional state, etc. Here we present a first attempt to identify such events in speech. For that purpose we will try to classify emotions by means of prosodic features, which provide a rich source of information in speech processing.

Prosody has long been studied as an important knowledge source for speech understanding and also considered as the most significant factor of emotional expressions in speech [1, 3]. In recent years there has been a large amount of computational work aimed at prosodic modeling for automatic speech recognition and understanding. The basic idea of the paper is to focus on the study of the basic correlation between Prosodic features like Pitch contour, energy contour and utterance timing and emotional characteristics. A minute approach instead of generally investigating various types of emotional expressions is taken into consideration [4]. The degree of emotion are categorized into “Neutral”, “Low” and “Strong”,[5] and accordingly the prosodic features of each category have been analyzed.

In this previous study so far, the type and degree of each emotion has been determined by the speakers themselves. In conversational communication, however, a speaker’s emotion inside his/her mind is not necessarily reflected in his/her utterances, nor is exactly conveyed to the listener as the speaker intended. The purpose of our study is therefore to clarify quantitatively (1) how much the speaker’s internal emotion (speaker’s intention) is correctly conveyed to the listener and further, (2) what type of expression is able to convey the speaker’s intention to the listener correctly. As the style of emotional expressions is gender-dependent, the gender features are also taken into consideration.

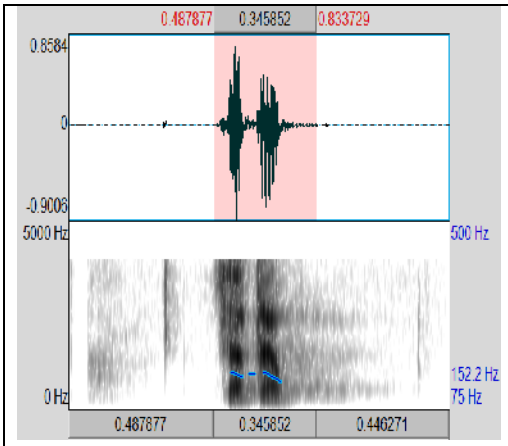
We first conducted a listening test to examine how much the speaker‘s intended emotions agreed with the listeners auditory impressions, using 130 word speech sample uttered by the college students at the age of 19-22 year old. The test results show that the subjects need not necessarily perceive emotional speech as the speakers intended to express.

From these results, we therefore analyzed the features of prosodic parameters based on the emotional speech classified according to the auditory impressions of the subjects. Prior to analysis, we calculated an identification rate of each type and degree of emotion, which was rate of the number of identifying as the specific type and degree of emotion to the total number of listeners. We selected 5 speech samples whose identification rates ranked the top 4 for each type and degree of emotion.

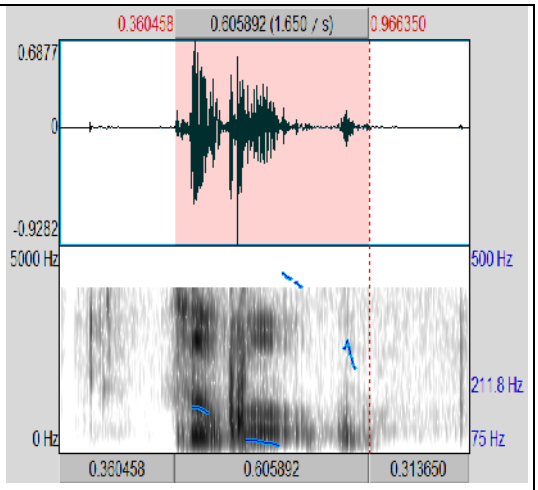
2. IMPLEMENTATION

2.1 Speech samples

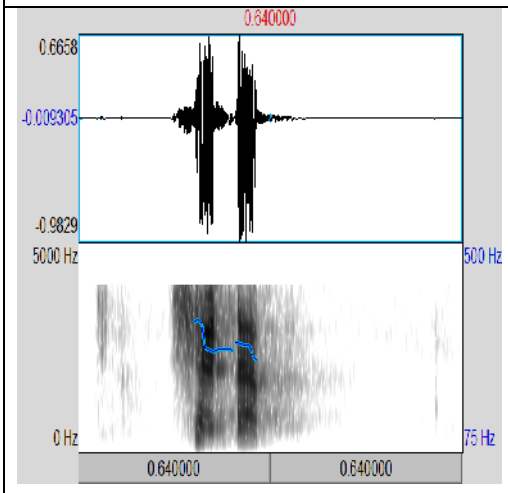
The speakers are college students in the age groups of 19 to 22. As speech samples, we use 2-Tamil words: “nalla iruku”, and ”ennada panura”. The types of emotions are “anger”, “happy”, and “sad”. Each word is uttered with following 2 degrees of emotions: “high”, “low” and sample speech signals are given in figure 1.



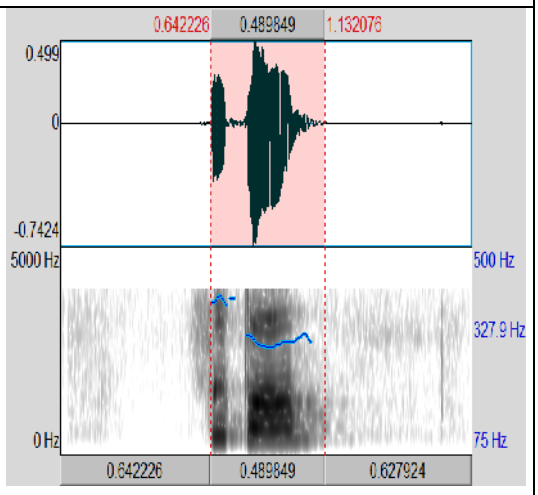
Male Speech sample for nalla irruku
Type & Degree of Emotion : Anger High



Male Speech sample for nalla irruku
Type & Degree of Emotion: Happy Low



Female Speech sample for nalla irruku
Type & Degree of Emotion : Anger High



Female Speech sample for nalla irruku
Type & Degree of Emotion: Happy Low

Figure 1: Speech Samples.

2.2. Prosodic Feature Parameters

Prosodic-feature parameters used in this work are Pitch contour, Utterance Timing and Energy contour. We did not use the speech power because the distances between the subjects and the microphone varied largely by their body movements during recording and we could not collect reliable power data.

These features have been extracted by means of MATLAB programs using Signal processing toolbox.

- **Pitch Contour**

The *pitch signal*, also known as the glottal waveform, has information about emotion, because it depends on the tension of the vocal folds and the subglottal air pressure. The pitch signal is produced from the vibration of the vocal folds. Two features related to the pitch signal are widely used, namely the pitch frequency and the glottal air velocity at the vocal fold opening time instant. For pitch, female pitch voice ranges from 150-300, for male pitch voice ranges from 50-200 and for child pitch voice ranges from 200-400. The time elapsed between two successive vocal fold openings is called *pitch period T*, while the vibration rate of the vocal folds is the *fundamental frequency of the phonation F*0 or *pitch frequency*. The glottal volume velocity denotes the air velocity through glottis during the vocal fold vibration. High velocity indicates music like speech like joy or surprise. Low velocity is in harsher styles such as anger or disgust. The pitch estimation algorithm used in this work is based on the autocorrelation and it is the most frequent one. A wide spread method for extracting pitch is based on the *autocorrelation of center-clipped* frames. The signal is low filtered at 900 Hz and then it is segmented to short-time frames of speech *fs*(*n*;*m*). The clipping, which is a nonlinear procedure that prevents the 1st formant interfering with the pitch, is applied to each frame *fs*(*n*;*m*) yielding

$$\hat{f}_s(n;m) = \begin{cases} f_s(n;m) - C_{thr} & \text{if } |f_s(n;m)| > C_{thr} \\ 0 & \text{if } |f_s(n;m)| < C_{thr} \end{cases},$$

where *Cthr* is set at the 30% of the maximum value of *fs*(*n*;*m*). After calculating the short-term autocorrelation

$$r_s(\eta;m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} \hat{f}_s(n;m)\hat{f}_s(n-\eta;m),$$

where η is the lag, the pitch frequency of the frame ending at *m* can be estimated by

$$\hat{F}_0(m) = \frac{F_s}{N_w} \operatorname{argmax}_\eta \{|r(\eta;m)|\}_{\eta=N_w\,(F_l/F_s)}^{\eta=N_w\,(F_h/F_s)},$$

where *Fs* is the sampling frequency, and *Fl*, *Fh* are the lowest and highest perceived pitch frequencies by humans, respectively. The maximum value of the autocorrelation (max{|*r*(*η*;*m*)|}*η*=*Nw* (*Fh*/*Fs*) *η*=*Nw* (*Fl*/*Fs*)) is used as a measurement of the glottal velocity during the vocal fold opening. *The Screenshot of the Pitch contour estimation for a single sample is given in figure 2.*

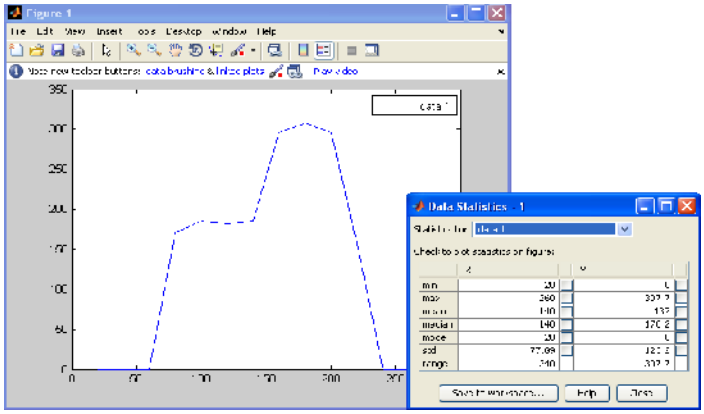


Figure 2: Pitch Extraction.

- **Utterance Timing**

There are two ways of extracting the utterance timing features. The first one is based on the length of syllables and the second one is based on the duration of pauses into the utterance voice periods. [10] To calculate the former type, we have to use the technique where syllables are detected automatically without needing a transcription. After detecting syllables, the total sounding time for every recording has to be calculated. The speech rate for every recording is obtained dividing the total amount of detected syllables by the total sounding time of the recording. From this procedure we can calculate: speech rate, syllable duration mean and syllable duration standard deviation.

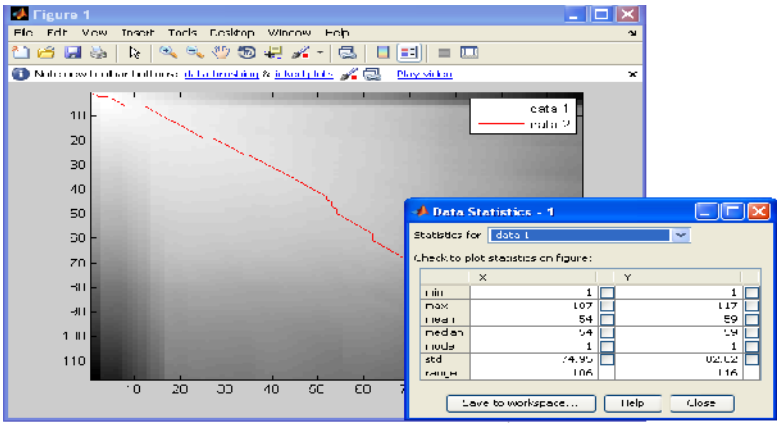


Figure 3 : Utterance Timing for a sample speech.

The latter type of Utterance Timing features is extracted from the calculation of the silences and voice period's durations in the utterance. This work uses this method and the features extracted are: pause to speech ratio, pause duration mean, pause duration standard deviation, voice duration mean and voice duration standard deviation. *Figure 3 gives the details of the Utterance timing features extracted for a sample speech.*

- **Energy Contour**

Energy is the acoustic correlated of loudness; their relation is not linear, because it strongly depends on the sensitivity of the human auditory system to different frequencies. Coupling of the loudness perception with the acoustic measurement is as complex as the coupling of the tone pitch perception and the computable F0. The sensation of loudness is both dependent on the frequency of the sound and on the duration, and the other way round, pitch perception depends on the loudness [10]. The short time speech energy can be exploited for emotion detection, because it is related to the arousal level of emotion. The short time energy of the speech frame ending at m is

$$E_s(m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} |f_s(n;m)|^2.$$

Figure 4 gives the extraction of energy contour features from a given sample and Figure 5 shows the combination of all the above three prosodic feature extraction of a given speech sample.

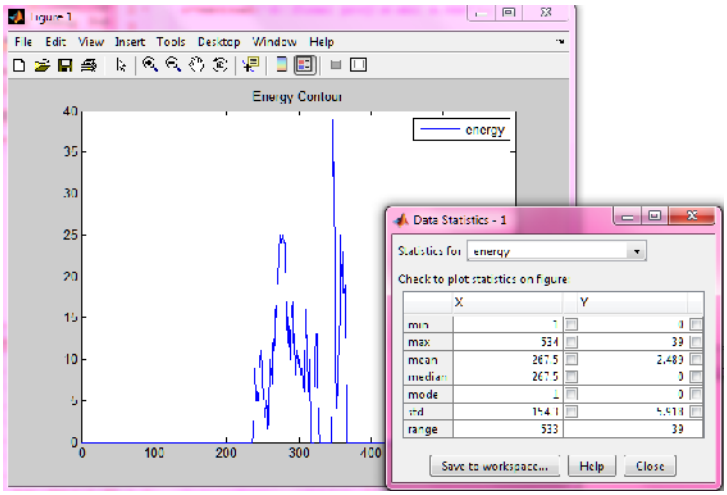


Figure 4: Energy Contour.

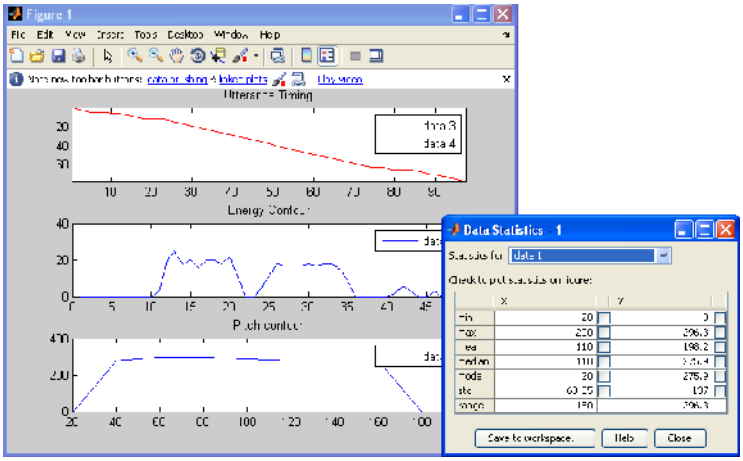


Figure 5: Extraction of Pitch, Energy Contour and Utterance Timing from a sample speech.

3. EXPERIMENTAL RESULTS

3.1. Experimental Conditions

Listening tests have been conducted to investigate whether or not there are Listeners and speakers gender dependent difference in the prosodic features of speech samples that have the same auditory impression on the type and degree.

In the listening tests, speech samples were presented to the subjects in random order. There were 6 dummy samples ahead and 100 test samples. We conducted two sessions for the purpose of cancelling the order effect. I n the second session, the speech samples were presented to the subjects in the reverse oder of those presented in the first session. Fifty subjects used a headphone of the same maker and the same sound pressure. Among them, 25 subjects were male at the ages of 20 and 21 years old and 25 subjects were female college students at the ages of 19 and 20 years old, both with a normal auditory capacity.

3.2. Identification Rate

To quantify the strength of listener's auditory impressions, an "identification rate r" [11] was introduced. We defined "an identification number" as the number of listeners' identification regardless of the type and degree of emotion that speaker intended to express. In the same way, we defined "*an identification rate r*" as a rate of the identification number to the total number of listeners.

Table 1 lists the top 5 speech samples in identification rate for each gender combination of speakers and listeners extracted from all types and degrees of emotional speech.

Speaker	Rank	Male Listener	Female Listener
		Degree & Emotion (Identification Rate r %)	
Male	1. 2. 3. 4. 5.	Anger High (100.0) Happy High (94.0) Happy Low (92.0) Anger Low (91.0) Sad High (86.0)	Sad High (100.0) Sad Low (98.0) Happy High (97.0) Happy Low (94.0) Anger High (940)
Female	1. 2. 3. 4. 5.	Anger Low (89.0) Happy Low (87.0) Sad Low (83.0) Anger High (80.0) Sad High (79.0)	Anger Low (99.0) Anger Low (90.0) Happy High (82.0) Sad High (80.0) Sad Low (78.0)

Table 1: The rank of Identification rate r for all speech.

In the case of speech uttered by Female speakers, the emotions and degrees perceived by both male and female listeners are having more or else the same identification rates. In the case of speech uttered by male speakers, on the other hand, only the "Anger High" and "Happy High" emotions are placed on the top positions. There is a mismatch in recognizing the other types and degrees of emotions.

3.3 Database samples and Results

The Database contains 300 speech samples, 150 each for the two Tamil words: "nalla iruku", and"ennada panura". Each word is uttered in different types of emotions: "anger", "happy", "sad" and in different degrees: "High" and "Low". Using MATLAB programs we have extracted the prosodic features for all the samples. Thus the Database contains all the samples and their corresponding prosodic features. We computed and compared the average prosodic features from all the 300 samples and the results are given below:

Figure 6 gives the comparison of the average Pitch range all the Male and Female speech samples present in the database. Figure 7 gives the comparative chart of different emotions of both male and female speech samples under different degrees. Figure 8 says that Female pitch is greater than Male but the energy counter and utterance timings of male are slightly greater than female.

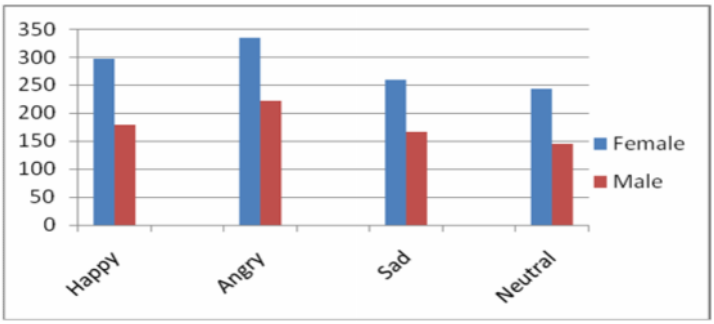


Figure 6: Average Pitch Comparison of Male and Female speakers'.

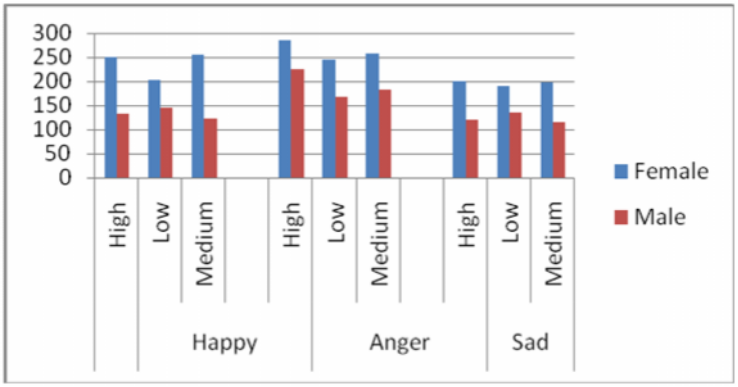


Figure 7: Comparisons' of all emotional in different degrees.

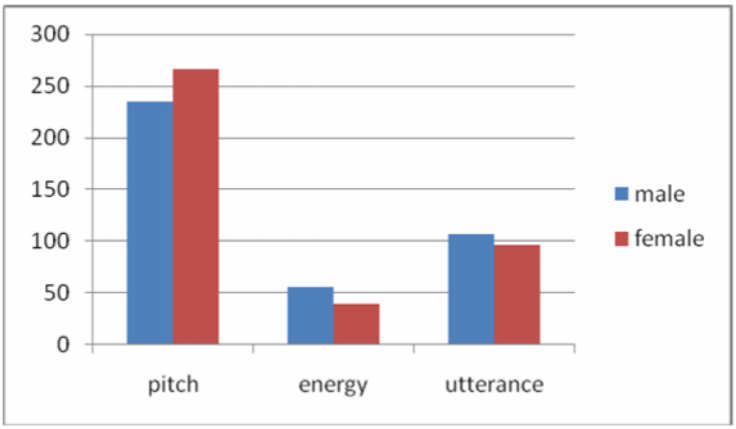


Figure 8: Comparison of Pitch Contour, Energy & Utterance Timing.

CONCLUSION:

The test results have suggested that there are Listeners as well as Speakers gender dependent differences in prosodic features to identify the type and degree of emotions.

Analysis results are summarized as follows: 1. For anger speech uttered by female speakers increases compared to that for neutral speech and significant difference has been observed from

male speech. 2. Prosodic features that characterize their emotions' are speaker gender- dependent. 3. Pitch for all emotion of female speech increases, and also significant difference has been observed from male speech.

Experimental results show the possibility in which our approach is effective for enhancement in Human Computer Interactions. Future works of our research are the following. We have to collect synthetic speech and put emotion labels on them. We have to reconsider how to estimate emotion in speech based on the results of experiments. For example, which features do we focus on? Which combination features and learning algorithms do we use? We have to reconsider evaluation of our approach and do it, too. People express and estimate more than one emotion in human speech. So we should think processing multi emotions in speech to develop better human computer interaction.

REFERENCES:

[1] Masaki Kuremastsu et eal, "An extraction of emotion in human speech using speech synthesize and classifiers foe each emotion", International journal of circuits systems and signal processing", 2008.

[2] Jordi Adell, Antonio Bonafonte et al, "Analysis of prosodic features: to-wards modeling of emotional and pragmatic attributes of speech", Proc. Natural lang proc, 2005.

[3] Hashizawa, Y.Takeda, S.Muhd Dzulkhifee Hamzah and Ohyama.G, "On the Differences in Prosodic Features of Emotional Expressions in Japanese Speech according to the Degree of the Emotion", Proc.2nd Int.Conf.Speech Prosody, Nara, Japan, pp.655-658, 2004.

[4] Takeda.S,Ohyama G, and Tochitani.A, "Diversity of Prosody and its Quan-tative Description" and and example: analysis of "anger" expression in Jap-anese Speech, Proc.ICSP2001, Taejon, Korea, pp.423-428, 2001.

[5] Mudh Dzulkhifee Hamzah, Muraoka T, and Ohashi T, "Analysis of Prosodic features of Emotional Expressions in Noh Farce speech according to the Degree of Emotions", Proc.2nd Int.Conf.Speech Prosody, Nara, Japan, pp651-654, 2004.

[6] Humberto Perez Espinosa, Carlos A. Reyes Garcıa, Luis Villase nor Pineda, "Features Selection for Primitives Estimation On Emotional Speech", ICASSP, 2010.

[7] Shiqing Zhang, Bicheng Lei, Aihua Chen, Caiming Chen and Yuefen Chen,"Spoken Emotion Recognition Using Local Fisher Discriminant Anal-ysis," ICSP, 2010.

[8] Jackson Liscombe, "Detecting Emotions in Speech: Experiments in three domains", Proc. Of the Human lang. tech conf. of the North America chap-ter of ACL, Pages: 251-234, June2006.

[9] P. Dumouchel N. Boufaden, "Leveraging emotion detection using emotions form yes-no answers," Interspeech, 2008.

[10] D. Galanis, V. Darsinos and G. Kokkinakis, "Investigating Emotional Speech Parameters for Speech Synthesis", ICECS, 1996.

[11] Jordi Adell, Antonio Bonafonte et al, "Analysis of prosodic features: towards modeling of emotional and pragmatic attributes of speech", Proc. Natural lang proc, 2005.

[1]