Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 30 (2012) 577 – 583

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Communication Technology and System Design 2011

# A socio friendly approach to the analysis of emotive speech

Agnes Jacob[a], P. Mythili[a], a*

*Division of Electronics, School of Engineering, Cochin University ,Kochi, , Kerala 673022,India*

**Abstract**

This paper describes certain findings of intonation and intensity study of emotive speech with the minimal use of signal processing algorithms. This study was based on six basic emotions and the neutral, elicited from 1660 English utterances obtained from the speech recordings of six Indian women. The correctness of the emotional content was verified through perceptual listening tests. Marked similarity was noted among pitch contours of like-worded, positive valence emotions, though no such similarity was observed among the four negative valence emotional expressions. The intensity patterns were also studied. The results of the study were validated using arbitrary television recordings for four emotions. The findings are useful to technical researchers, social psychologists and to the common man interested in the dynamics of vocal expression of emotions.



*Keywords:* Intonation; elicited emotions; perceptual listening test; pitch contour; valence; validation.

## 1. Introduction

Over the recent years, numerous studies have been done on speech, for various purposes. The interdisciplinary nature of speech along with its high degree of variability in both the production as well as perception makes speech analysis a complicated area of research. Even so, some of these speech based studies have succeeded in deducing the age, sex, height and weight of the speaker as well as his/her state of health with respect to certain specific diseases. Regarding emotive speech, several studies have been done in the past in order to assess the emotional state of the speaker from the speech samples usually based on a maze of algorithms and reports of their comparisons are available. According to Chateau *et al* [1] speech emotional quality refers to the emotional content and describes the listeners' global impressions as elicited by their audition. During the past few years, the focus has been on the manner in which emotions are displayed in vocal interactions. It is more so, in the light of the increased social role

* Agnes Jacob; Tel.: 09446522793.
*E-mail address*: nirag_2007@yahoo.co.in.

1877-7058 © 2011 Published by Elsevier Ltd.
doi:10.1016/j.proeng.2012.01.901

of emotional quotient (EQ) in almost all organizations and in the different spheres of life. One significant component of EQ is emotional awareness. In this context, we ought to look out for more user friendly methods of emotional speech analysis that have sufficient reliability to assess emotions without being weighed down by algorithmic complexity. Therefore the motive of this work was to understand the manifestation of emotions using the least amount of signal processing and with least inputs. Scherer [2] has already found out from his studies that even though the energy of an utterance also contains valuable information, it cannot solely differentiate the basic emotions. Hence this study has explored and inferred the universal characteristics of utterances under various emotions, using a combination of their pitch and pitch contours along with their average energy. Banziger and Scherer [3] have earlier reported that the pitch of an utterance is actually indicative of its fundamental frequency and is one of the prime parameters for determining the emotion in speech. Whereas the previous works in this field were mostly based on male speech databases available on the Web, this work was based on the female speech database developed exclusively for a study of the energy and intonation aspects of speech. Literature survey reveals that men and women use their voices differently; one main difference usually manifested is in the pitch of the utterances. The pitch as well as energy profile characteristics for six basic emotions along with the neutral have been identified and these are presented in a manner that is more simple and intelligible to any interested lay person. The results had been validated by comparing the results of this study with those obtained from the analysis of arbitrary TV recordings.

## 2. Methodology

This work was based on neutral and six basic emotions namely, sad, surprise, happy, anger, fear, disgust as these are more universal than the others. These seven emotions were elicited from the speech samples of six females. Female subjects were chosen for this study as their speech is more expressive of emotions. All the six subjects were educated, non-professional, urban, non-native, Indian speakers of English. The study was based on elicited speech in which emotions have been induced**.** As the ultimate results of any research on emotional speech relies on the richness and appropriateness of the database, the speech corpus had been specially designed to accommodate the common features in the speech used by women such as exclamations, hedges and disclaimers. The English database had been designed taking into account various gender, social and linguistic aspects as suggested by Giri.N [4] and Jones [5]. Since spontaneous emotions are very difficult to record and acted emotions have exaggerated expressions, it was decided to investigate on a database of elicited emotions. The database for this experimental study consists of short, but often used utterances. Subjects were in the age group of 32 years to 42 years, aware of the purpose of the recordings. Since emotions had to be elicited, several trials were required in order to get three good samples of each utterance. The recordings were done in different sets on to a computer hard disk, using the Realtek audio system. These wave files were then segmented, labeled and stored. Perceptual listening tests had been conducted to verify the correctness of emotional content in the recordings under each of the seven classes of emotion. The recorded database was evaluated by nine listeners without any hearing difficulties [6]. All the listeners were in the age group 15-50 and had no hearing pathologies. Listeners listened over headphones and indicated the perceived emotion from a list of six emotions apart from the neutral. The listeners rated the sound files for each emotion on a 5 point scale ranging from bad to excellent through fair, good and very good. While complex factors like intelligibility and naturalness are the major criteria in any speech evaluation scheme, here the stress was on the speech emotional quality which is basically a perceptual factor. Table.1 consists of a summary of the listening tests results. Only those sound files rated as good or above it were considered to be recognized properly and included for further pitch contour analysis. One of the challenges in the research in this field of analysis of emotional speech is the difficulty in creation of an authentic database.

Table 1. Percentage of emotions recognized correctly

| Happy | Surprise | Neutral | Anger | Disgust | Sad | Fear |
|---|---|---|---|---|---|---|
| 90 | 92 | 90 | 96 | 94 | 93 | 95 |

The high percentage of successful recognition of emotions by the listeners indicates the validity of the speech corpus design as well as the methods used for eliciting emotions and was followed by Jacob and Mythili[6]. The pitch was analyzed using the autocorrelation method explained by Rabiner and Schafer [7]. Initially the mean values of pitch were found programmatically using the functions in the Matlab Signal Processing Toolbox. For pitch contours, Praat's autocorrelation based pitch tracking algorithm was used to extract pitch values incorporating a gender dependent pitch range. The algorithm performs acoustic periodicity detection on the basis of an accurate autocorrelation method, as described in Boersma [8]. It is more accurate than other methods and gives $F_0$ with an accuracy of $10^{-6}$. The autocorrelation $r_x(\tau)$ of a time signal x(t) as a function of the time lag $\tau$ shows the similarity of the signal with displaced versions of itself, where $\tau$ is the displacement time or lag time and is defined as,

$$r_x(\tau) = \int x(t)x(t+\tau)\, dt \quad (1)$$

Since the concept of a long-time autocorrelation measurement is not really meaningful for a non stationary signal such as speech, it is reasonable to define a short- time autocorrelation function, which operates on short segments of the signal, as explained by Boersma [9]. Candidates for the fundamental frequency of a continuous signal $x(t)$ at a time $t_{mid}$ can be found from the local maxima of the autocorrelation of a short segment of the sound centred on $t_{mid}$. As per the algorithm for the speech-like signal $x(t)$, a piece with duration T (the window length), centred around $t_{mid}$ is taken from which, the mean $\mu_x$ is subtracted and the result is multiplied by a window function w(t), so that we get the windowed signal as,

$$a(t) = [x(t_{mid} - \tfrac{1}{2}T+t) - \mu_x]\, w(t) \quad (2)$$

The window function $w(t)$ is symmetric around $t = \frac{1}{2}$ T and zero everywhere outside the time interval [0, $T$]. The sine-squared or Hanning window was chosen which is given by

$$w(t) = \tfrac{1}{2} - \tfrac{1}{2}\cos(2\pi t/T) \quad (3)$$

Thus Praat's pitch detection method estimates a signal's short-term autocorrelation function on the basis of a windowed signal, by dividing the autocorrelation function of the windowed signal by the autocorrelation function of the window

$$r_x(\tau) \approx r_{xw}(\tau) / r_w(\tau) \quad (4)$$

Where x(t) is the signal, xw(t) represents the windowed signal and $r_x(\tau)$ is the short term autocorrelation function of the signal. Finally, the places and heights of the maxima of the continuous version of $r_x(\tau)$ are found out. Praat always computes 4 pitch values within one window length, i.e., the degree of oversampling are 4. Praat software used in this analysis employs a pitch floor of 75Hertz and a ceiling pitch value of 500Hertz. The segmented wave files were analyzed one at a time and the pitch contours were saved in separate files. Typical pitch contours possessing the salient, common features were selected to represent the numerous speech files under the different emotion classes and these are

presented in the next section. These pitch contours were then examined for intra group similarities within any particular emotion class and for inter group similarities between the various emotion classes. The contours give a comprehensive picture of the entire pitch profile over the duration of an utterance rather than just a few instantaneous or representative values of pitch as given by Table. 2. The results of this experimental study to assess emotions through pitch contour profiles were verified using selected excerpts from TV interviews, talk shows, and news reports. This validation was done by using the audio recordings on appropriate topics of suitable semantic content. These sound files too were further evaluated for their emotional content using the perceptual listening tests. The pitch contours of only those files rated good, very good or excellent were examined, out of all the different recordings. In all the cases considered, the obtained pitch contours were in conformity with the pitch contours obtained for similar emotions in our experimental study. In this experimental study, the pitch contour profiles were further verified by discrete measurements of the mean pitch values also.

## 3. Results and Discussions

The typical pitch contours characterizing each of the basic emotions are as presented from Figure.1 to Figure.7. In this method of pitch contour extraction using the Praat software, the actual sound extends beyond the limits of the contour as can be noted from the figures given below wherein the marking on the X-axis (time) is beyond the end of the pitch contour. This is so since the pitch can be calculated only for the voiced region and cannot be found for the unvoiced part of the sound. In almost all of the basic emotions considered here, the last quarter duration of the utterance was observed to be unvoiced. There are variations in the pitch contours with the choice of speaker, choice of words in the utterance as well as with the emotion contained in the utterance. Even so, within any particular emotion class we have successfully identified certain characteristic features and these formed the basis for classification of different emotions and the correct identification of the class representatives. As depicted in Figure.1 and Figure.2 the ending pitches are lower than the values at the beginning of utterances under happiness and surprise.

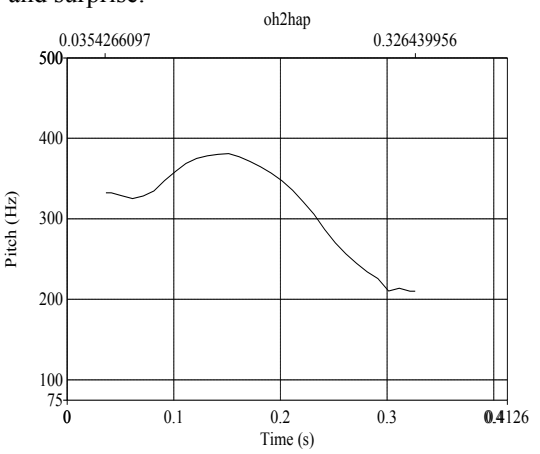


Fig.1.Typical pitch contour of happy emotion

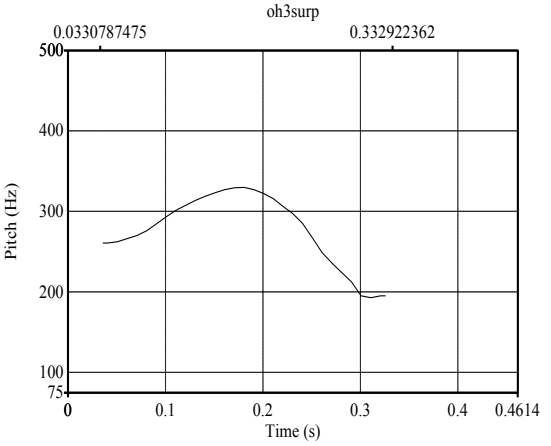


Fig.2. Typical pitch contour of surprise emotion

The pitch contours of like utterances under positive valence emotions such as surprise and happiness are similar, though the variations are more pronounced in the case of surprise. Also surprise will have higher average pitch values than happy emotion. The least variations in pitch are seen in Figure.3, Figure.4 and Figure.7 corresponding to neutral, sad and disgusted utterances respectively. Pitch contours

under the neutral emotion do not have a distinct peak and are similar to the sad emotion pitch contours except that the pitch values are often higher than for sad. Sad utterances are marked by nearly horizontal traces indicating a constancy of pitch over a major part of the utterance.

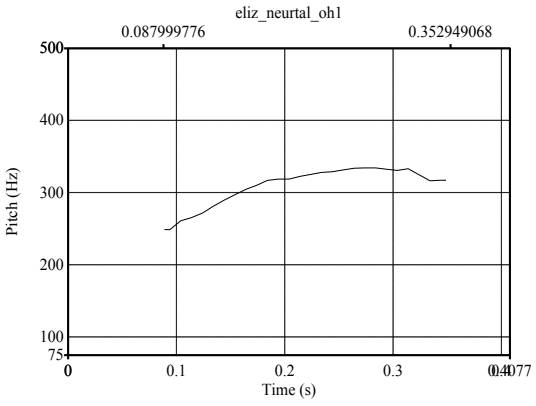


Fig.3. Typical pitch contour of a neutral utterance

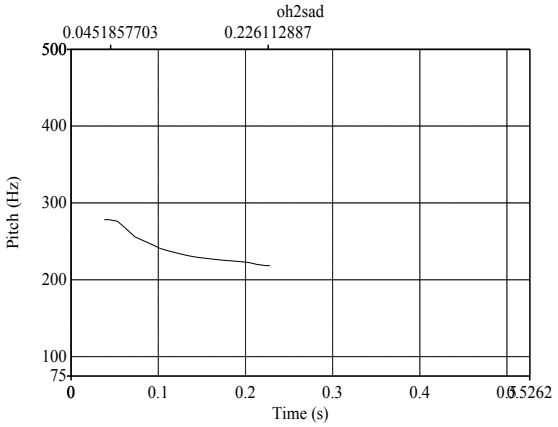


Fig.4. Typical pitch contour of sad emotion

The mean pitch values in Table.2 summarize the discrete pitch measurements for the emotions and validate the observations based on pitch contours. In terms of intensity it is seen that anger has the highest energy while neutral and sad are at the lower end.

Table 2. Pitch and intensity values

| Emotions | Happy | Surprise | Neutral | Anger | Disgust | Sad | Fear | Validation by TV recordings | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Happy | Neutral | Anger | Sad |
| Mean Pitch (Hertz) | 294 | 303 | 228 | 208 | 200 | 222 | 234 | 298 | 260 | 303 | 266 |
| Intensity(dB) | 74.4 | 74.5 | 71.5 | 78.4 | 72.2 | 71.3 | 74.2 | 82.6 | 79.5 | 82.8 | 80.5 |

From Table.2 it is seen that the highest mean pitch values are for utterances under surprise and the lowest correspond to disgust. Among the negative valence emotions, fear has the highest mean pitch value. In Figure.5, a representative of angry utterance of the second speaker in the database; the voiced portion is confined to almost the first half of the utterance. Even though no universal similarities are observed among negative valence emotions, similarities are noted between certain utterances under anger and fear. Anger is characterized by rise to a peak followed by either a decrease or leveling out of the pitch values and the utterance duration is observed to be small. In almost all utterances under fear, the pitch increases to a peak and then decreases slightly only.

From the validation results it is seen that here too neutral and sad have nearby pitch values and the pitch of happy utterances is higher than these two. But anger was found to have higher pitch values than in our study, probably because the voice samples from TV recordings belonged to acted emotions as

obtained from serial telecasts or projected emotions as in interviews. In both case, the emotional expressions are exaggerated. Hence the utterance under anger was in many cases a furious utterance.

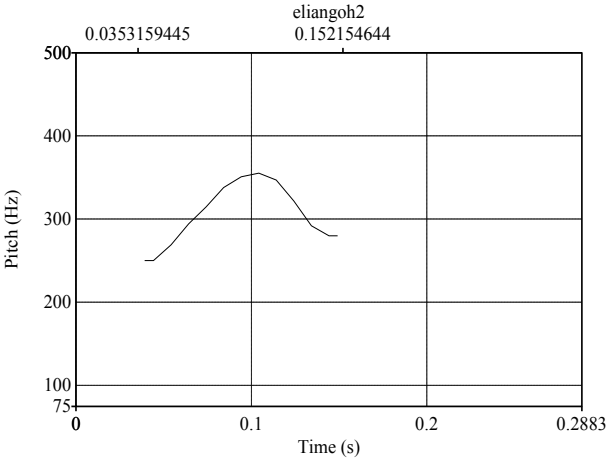


Fig.5.Typical pitch contour of an angry utterance

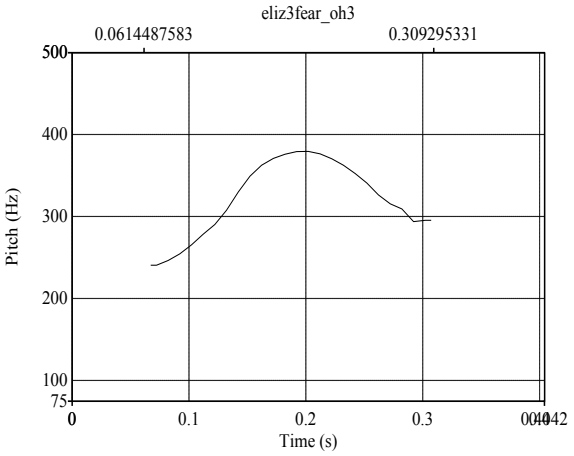


Fig.6.Typical pitch contour of an utterance under fear

The disgust contour of the second speaker in Figure.7, shows that pitch values are lowest for this emotional expression. A few other pitch contours for disgust have sharp decrease in pitch and are therefore wavy towards the end of the utterance. From Table. 3 it is seen that the pitch range is the least for sad, disgust and neutral, while it s the largest for happy and fear. Negative valence emotions such as sadness and disgust are mostly expressed by lower pitch values as compared to the positive valence emotional expressions. Further a large pitch variation occurs for utterances under happiness, surprise or fear. There is concurrence between the observed pitch profiles represented graphically and that obtained from discrete pitch measurements conducted separately.

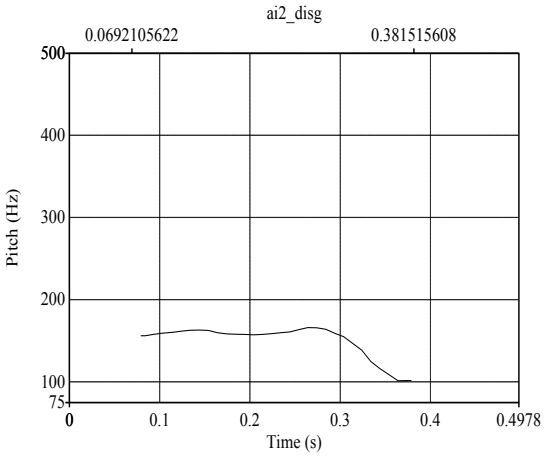


Fig.7.Typical pitch contour of utterance under disgust

Table 3. Pitch range

| Emotions | Pitch range (Hertz) |
|---|---|
| Happy | 172 |
| Surprise | 137 |
| Neutral | 86 |
| Anger | 105 |
| Disgust | 66 |
| Sad | 60 |
| Fear | 139 |

## 4. Conclusion

Some of the significant features of the emotive speech files have been summarized in terms of their pitch contours, pitch values, pitch range as well as the average intensity in order to enable classification into various emotion types. Validation of pitch and intensity measurements was done using TV recordings. Characteristic features of the utterances under each emotion were identified by their representative pitch contours. Positive valence emotions of happiness and surprise have right skewed pitch contours; while negative valence ones like anger and fear have slightly left skewed pitch contours. Neutral and sad have similar pitch contour profiles except that in most cases, neutral has comparatively higher values in a similar pitch profile. Disgust is characterized by a smooth decline in the average pitch from a higher level. Such graphical representations provide a more simple, yet intelligible characterization of emotions rather than the stream of numerical values of F0s (fundamental frequencies) at various time instants. This can be used in further analysis of speech by researchers, students studying speech technology or even by any person without the adequate technical background required to understand the deeper levels of signal processing. Further it can also be used in emotional speech synthesis where the researcher requires guidelines on the pitch patterns to be followed in the synthesized speech so as to convey any particular emotion. These results are more significantt in the Indian context as the data bases were locally developed along the lines mentioned in relevant literature by foreign researchers in speech analysis. This experimental work also attempts to inspire people especially women who are vocally more expressive to be conscious of the vocal expression of emotions. Such a basic awareness can help one to modify his or her vocal behaviour to suit a particular situation or need. Thus the results of this study provide better understanding on the production and manifestation of various emotions and are useful for the purpose of analysis and synthesis of emotional speech as well as for manipulating vocal emotional expressions.

## References


[1] Noel Chateau, Valerie Maffiala, Thibaut Ehrette, "Modeling the emotional quality of speech in a telecommunication context", *Proceedings of the 2002 International Conference on Auditory Display*, Kyoto, Japan, July 2-5, 2002.

[2] K.R. Scherer. Vocal communication of emotion: a review of research paradigms. *Speech Communication*, vol. 40, issues1-2, April 2003, pp 227-56, doi: 10.1016/S0167-6393(02)00084-5.

[3] Banziger and K.R Scherer. , "The role of intonation in emotional expression", *Speech Communication*, July 2005, pp. 252- 67, doi:10.1016/j.specom. 2005.02.01. Vol 46, issues3-4,

[4] Vijai N. Giri, " Gender role in communication style*"*, IIT Kharagpur, Concept publishing House, New Delhi. 110005.

[5] Daniel Jones. , "Cambridge english pronouncing dictionary", Fourteenth Edition, Cambridge University Press, 2004..

*[6]* Agnes Jacob P. Mythili., "Developing a child friendly text-to-speech system" , *Hindawi Journal. Advances in Human Computer Interaction.*Volume, 2008. doi: 10.1155/2008 /597971.

[7] Rabiner L R, Schafer RW., " Digital processing of speech signals", *3rd ed. Englewood Cliffs, New Jersey :Prentice Hall*, 1978.

[8] Boersma P, "Weenink D. Praat:doing phonetics by Computer (Version4.6.09)", 2005 [Computer Program], http://www. Praat.Org/.

[9] Boersma P. , " Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", *Proceedings 17,1993.* Institute of Phonetic Sciences,University of Amsterdam, pp. 97-110.