Objective analysis of emotional speech for English and Slovenian Interface emotional speech databases

Vladimir Hozjan, Zdravko Kacic

*Faculty of Electrical Engineering and Computer Science Institute of Electronics University of Maribor
Smetanova17, SI - 2000 Maribor, Slovenia
(vladimir.hozjan@uni-mb.si, kacic@uni-mb.si)

Abstract

In this paper we propose a new approach for analysis of emotional speech prosody features. The aim of the analysis is definition of emotional features that characterise emotions. Analysis was performed on emotional speech databases that were recorded in the framework of the project "Multimodal Analysis/Synthesis System for Human Interaction to Virtual and Augmented Environments" (Interface). The new approach determines emotional features in three steps. In the first step the low-level features were determined, second step includes definition of high-level features and in the last step the emotional features were determined. Emotional features for native English and native Slovenian speakers were analysed. A comparison of emotional features between English male and female speaker, English male and Slovenian male as well as English female and Slovenian female speaker was performed. New approach for analysis of emotional features enables consistent and objective analysis. It enables comparison of emotional features between different speakers and at the same time searches for new correlates of emotions in speech.

1. Introduction

Speech communication between humans includes emotions that are important factor for successful communication. In the future the emotions will play important role also in multimodal human-computer interaction (Sharma et. al. 1998).

The analysis of speech signal that is affected by emotion can be done on different levels. It can be done on acoustic, prosodic or lexical level (Cahn, 1990).

In the work done in the recent past on emotional speech analysis the following findings were reported (Pereira, 1996; Noad et. al. 1997). Emotion anger affects increase of mean value of pitch contour, pitch ranges, variety of pitch and mean value of energy. There is some evidence about increasing values of high frequencies. Lowering of pitch contour and increased degree of articulation was also noticed.

For sadness typically decrease of mean value of pitch contour, pitch ranges, variety of pitch, degree of articulation and decreased value of high frequencies can be seen. Speech spoken in sadness has lower pitch contour. Variety of pitch contour is low.

Scherer and Pittam (1992) pointed on the fact, that increase of mean value of pitch contour, pitch ranges, variety of pitch, energy value and value of high frequencies result in arousal of emotions. They threat happiness and anger as emotions with high arousal, sadness and neutral speech style as emotions with low arousal. In this paper we followed this classification of emotions.

In the work reported in this paper we have focused on analysis of speech signal, which was spoken by human that was affected by various emotions. The aim of speech synthesis systems is to produce speech, which sounds naturally emotional. This is the same goal as that of actors (Johnstone, 1996). To record a collection of natural speech that contains naturally expressed emotion is a complex problem. Some researchers have done an analysis of naturally expressed or spontaneous emotion (Williams and Stevens, 1972).

In this paper we present new approach for objective analysis of emotional speech. The results of the approach are categorised features that are compared between various speakers.

2. Databases

The emotional speech databases used were recorded in the framework of the IST project "Multimodal Analysis/Synthesis System for Human Interaction to Virtual and Augmented Environments" (Interface). Within the project databases for four languages were recorded: Slovenian, English, Spanish, and French (Hozjan et. al., 2002). Slovenian emotional speech database contains one male and one female speaker. English emotional speech database contains two mail speakers and one female speaker.

Speaking style	Abbreviation
Anger	A
Disgust	D
Fear	F
Neutral fast	H
Joy	J
Neutral slow	N
Surprise	S
Sadness	T

Table 1:List of speaking styles used in emotional databases.

Emotional speech databases consist of recordings recorded by professional actors. The selected emotions are standardised in MPEG 4 FAP (facial animation parameters) standard. The FAP protocol describes the facial animation for different emotional expression and for synchronisation with speech signal. The FAP protocol provides parameters for six emotional facial expressions.

The emotional speech databases consist of speech recordings in six emotional states and in two neutral speech styles. These emotions and neutral speaking styles are listed in the table 1. To be able to evaluate the analysis of emotional speech, a reference speech signal has to be defined. The references in our case were two neutral speech styles. The slow/soft neutral speech style served as a reference to emotions with low arousal (sadness and disgust). The fast/loud neutral speech style is more dynamic and served as a reference to emotions with arousal (happiness, fear, and anger).

Corpus of Slovenian emotional database consists of 35 isolated words and 155 sentences. 100 sentences are context independent. There are 20 short sentences, 60 middle long sentences and 20 long sentences. Last 55 sentences are context dependent and are of various lengths.

Corpus of English emotional database consists of 35 isolated words and 155 sentences. 100 sentences are context independent. There are 20 short sentences, 60 middle long sentences and 20 long sentences. Last 50 sentences are context dependent and are of various lengths.

Short sentences are composed from five to eight words, middle long sentences are composed from nine to thirteen words and long sentences are composed from fourteen to eighteen words.

The recording was made in studio environment with studio microphone AKG 3000B and laryngograph Portable Laryngograph form Laryngograph Ltd. A laryngograph was used to record the movements of vocal folds. The sampling frequency was 48 kHz with quantisation of 16 bits. Recordings were made in two sessions. The second session was repeated after 14 days. The aim of this was to evaluate the consistency of emotional speech parameters.

3. Low-level features

First step of our approach determines low level features. Pitch contour (F0) of speech signal, duration of phonemes and pauses, energy of speech signal, derivative of pitch contour (ΔF0) of speech signal, and derivative of energy contour of speech signal are defined as low-level features. Those features represent prosody in speech signal.

Method with normalized cross-correlation function (Qian and Kimaresan, 1996) determines F0 from laryngograph signal. This method calculates one value per 10ms. RMS (Root Mean Square) method calculates energy from speech signal. RMS calculates energy in the square window that is 10ms long at the frame rate of 5ms.

Phoneme segmentation was made with the HTK toolkit (Young et. al., 1997). Here, the speech signal was first pre-emphasised with the factor of 0.95. As a feature vector 12 mel-cepstral coefficients and the energy, as well as the first and the second derivative of the mel-cepstral coefficients were used. The cepstral coefficients were calculated with Hamming window (5 ms long) at the frame rate of 2,5 ms. Normalisation of cepstral mean value was used. We used 3-state left-right HMM. The emission probabilities were modelled with continuous Gaussian mixture densities. The models for all phonemes, for pauses between words and start and end points of sentences were used in acoustic modelling.

Training database consisted of all the sentences for one speaker (speech recorded in both sessions and for all emotional states). On initialisation global mean value and variances were estimated. HMM parameters were reestimated with 6 iterations of Baum-Welch algorithm. In the next stage, the number of Gaussian mixture densities was increased from one to 8 in step of 2. After each increase of number of Gaussian mixture densities, two iterations of HMM parameters were performed. Segmentation was done on all sentences of one speaker for all emotional states.

4. High-level features

Statistical presentations of low-level features are defined as high-level features. Calculation of mean value, standard deviation, minimum searching, and maximum searching are statistical operations that calculates high-level features from low-level features.

As high-level features derived from F0 we defined mean value of F0, standard deviation of F0, minimal value of F0, maximal value of F0, and F0 range. High-level features derived from ΔF0 were mean value of ΔF0, standard deviation of ΔF0, minimal value of ΔF0, maximal value of ΔF0, and ΔF0 range.

The second set of defined encompassed features that are calculated from energy contour and derivative of energy contour: mean value of energy contour, standard deviation of energy contour, energy of words, energy of syllables, mean value of derivative energy contour, and standard deviation of derivative energy contour.

The third group consists of features that were extracted from phoneme duration: duration of words, duration of syllables, duration of vowels, duration of fricatives, duration of affricates, duration of sonorants, duration of plosives, and duration of pauses.

High-level features present information about prosody of speaker over entire sentence. They contain information about intonation, tempo and loudness.

4.1. Emotional features of a speaker

Analysis of high-level features selects those high-level features that determine specific speaking style. These features can be characterized as emotional features. Equation (1) normalizes high-level features to reduce intra-speaker variability.

$$xn_{i,j} = \frac{x_{i,j}}{\bar{x}_i}, \qquad (1)$$

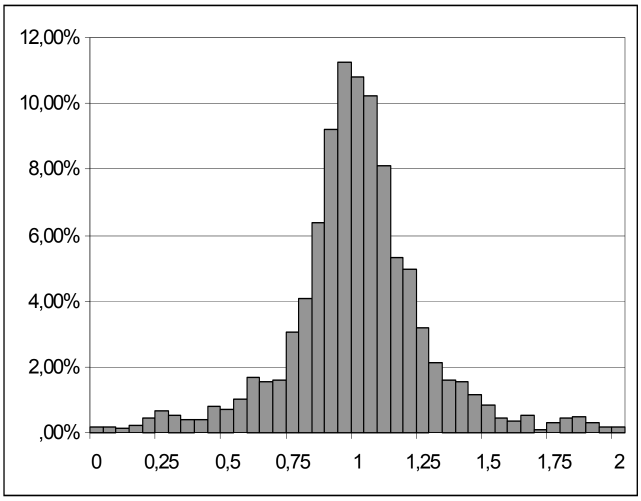

Figure 1: Histogram of normalized high-level features for English male speaker (M1) for all values of high-level features.

where $xn_{i,j}$ is normalized value of i^{th} feature in the j^{th} sentence, $x_{i,j}$ is i^{th} feature in j^{th} sentence, $\bar{x}_i$ is mean value of i^{th} feature. Each high-level feature was normalized across all speakers' high-level features using the equation (1).

Sentences were divided into five groups. First group (G1) contained sentences with isolated words, second group (G2) contained short sentences, third group (G3) middle long sentences, fourth group (G4) long sentences, and fifth group (G5) was composed from context dependent sentences. Mean value for all features and for each group (G1 – G5) was calculated.

Equivalent to the Banse and Sherer's (1996) classification of emotional feature values, we also used five classes of emotional feature values. Histogram (see figure 1) of all high-level feature values is in the range between 0 and 2. Mean value of all normalized features is 1 and variance is 0.094. The range of values was divided into five classes. In table 2 the values of all classes shown. Values 1.2 and 0.8 present deviation from mean value that is equal to 2 × variance. Values 1.05 and 0.95 are deviation from mean value for ½ variance. Two classes are above value 1 and two classes are below value 1 (see table IV). Class 1 presents values that are extremely below mean, class 2 presents values that are below mean, class 3 presents values that are around mean, class 4 presents values that are above mean, and class 5 present values that are extremely above mean.

1	$xn_{i,j} < 0.8$
2	$0.8 \le xn_{i,j} < 0.95$
3	$0.95 \le xn_{i,j} < 1.05$
4	$1.05 \le xn_{i,j} < 1.2$
5	$1.2 \le xn_{i,j}$
$xn_{i,j}$ - normalized high-level feature of ith feature in the jth sentence	

Table 2: Classes and range of high-level feature values.

Normalized high-level features were compared accordingly to different type of sentence length. Mean values of normalized high-level features for each group and for each session ($\overline{X}_{i,e,r,s}$) were calculated. $\overline{X}_{i,e,r,s}$ is mean value of i^{th} normalized high-level feature of eth speaking style, of r^{th} group, and of s^{th} session. If $\overline{X}_{io,eo,ro,E1}$ and $\overline{X}_{io,eo,ro,E2}$ are in the different classes, is feature x_{io} is denoted as a non-emotional feature. If feature $\overline{X}_{io,eo}$ is for all groups (G1 to G5) of the same emotion in different classes, then feature x_{io} is denoted as a non-emotional feature. If feature $\overline{X}_{io}$ has same values in all eight speaking styles, then feature x_{io} is also denoted as non-emotional feature. If feature x_{io} is not non-emotional feature, then it is denoted as emotional feature. Described rules of emotional features selection affirm emotional feature independence of sentence length and session of recording but dependence on speaking style.

5. Results of analysis

We compared emotional features between English male and female speaker, between English male and Slovenian male speaker and between English female and Slovenian female speaker.

5.1. Comparison between English male and female speaker

Comparison of emotional features between English male and English female speaker showed that average mean value of pitch contour is lower for male speaker. Emotions anger, joy, and surprise have higher mean values as emotion surprise and neutral slow speaking style (see figure 2). Female speaker has high value of mean value of pitch contour for emotions disgust and sadness and low value for emotion fear. Male speaker has low value of mean value of pitch contour for emotions disgust and sadness and high value for emotion fear. Both speakers have similar behavior of high level features.

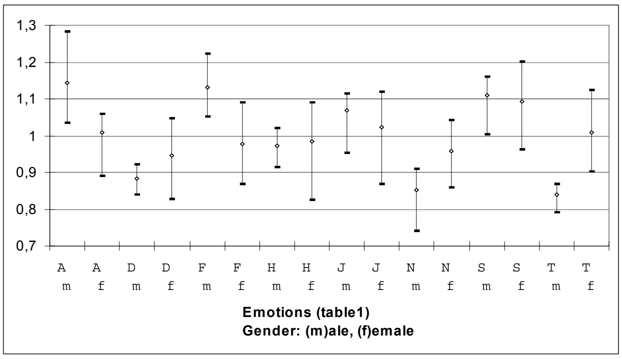

Figure 2: Mean values of pitch contour for English male and English female speaker, normalised to speaker's mean value of pitch contour. Abbreviations for emotions are given in Table 1.

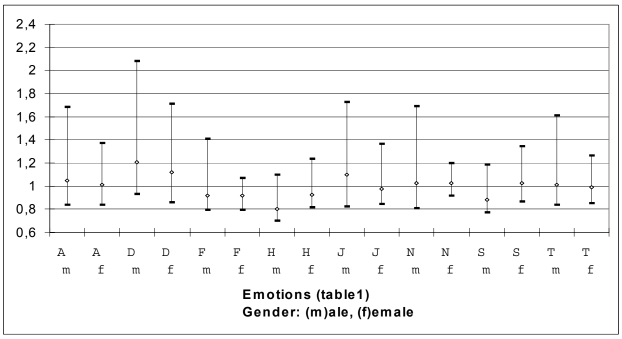

Figure 3: Duration of vowels for English male and English female speaker normalised to speaker's vowel duration. Abbreviations for emotions are given in Table 1.

Emotions anger, fear, joy and surprise cause increase of emotional features that are calculated from pitch contour (standard deviation of pitch contour, minimal value of pitch contour, maximal value of pitch contour, standard deviation of derivative of pitch contour, minimal value of pitch contour, maximal value of pitch contour, derivative of pitch contour range).

The highest values of vowel duration have male speaker for emotions anger, joy, and disgust. Female speaker has highest values for emotions disgust and sadness. The lowest values have both speakers for fast neutral speaking style (see figure 3). The average value of vowel duration is lower for male speaker.

Male and female speakers have high values of emotional features that are calculated from duration (duration of syllables, duration of fricatives, duration of sonorants, duration of plosives) for emotion disgust and low values for fast neutral speaking style.

5.2. Comparison between English male and Slovenian male speaker

English and Slovenian male speakers have high mean values of pitch contour for emotions anger, fear, joy and surprise. Low mean values of pitch contour are for emotions disgust and sadness (see figure 4). The difference between English speaker and Slovenian speaker is for emotions disgust and sadness. English speaker has lower average value of mean values of pitch contour as Slovenian speaker.

Emotional features that are calculated from pitch contour have similar behaviour as mean value of pitch contour. Emotions anger, fear, joy, and surprise have increased values and emotions disgust and sadness have decreased values. In case of high level features that were calculated from pitch, the English speaker had lower values as Slovenian speaker.

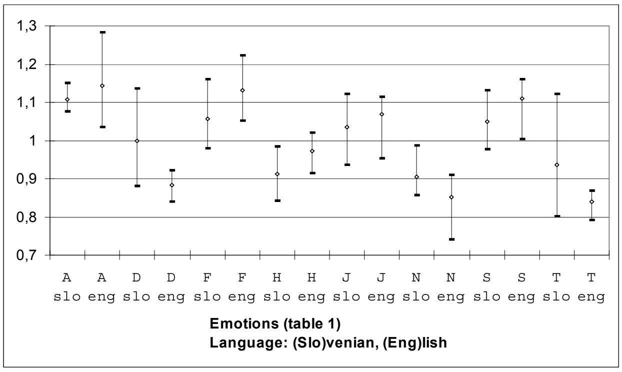

Figure 4: Mean value of pitch contour for English male and Slovenian male speaker, normalised to speaker's mean value of pitch contour. Abbreviations for emotions are given in Table 1.

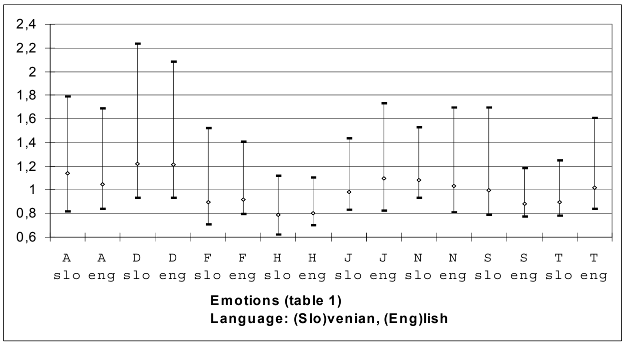

Figure 5: Duration of vowels for English male and Slovenian male speaker, normalised to speaker's vowel duration. Abbreviations for emotions are given in Table 1.

Highest values of vowel duration can be seen for emotion disgust, joy, and slow neutral style of speech. The lowest values of vowel duration are for fast neutral speaking style and for emotions fear and surprise (see figure 5). Slovenian speaker had lower value of vowel duration for emotion sadness, as this was the case for English speaker. Slovenian male speaker has lower average value of vowel duration as English male speaker.

Other emotional features that were calculated from duration have similar values as vowel duration. Emotional features that are calculated from duration have high values for emotions disgust and low values for fast neutral speaking style.

5.3. Comparison between English female and Slovenian female speaker

English female and Slovenian female speakers have high mean values of pitch contour for emotions anger, surprise and joy (see figure 6). Both female speakers have low mean value of pitch contour for emotion sadness and disgust. Difference in mean value of pitch contour was seen in emotion fear, where the Slovenian female speaker has high values of pitch contour and English female speaker has low mean values of pitch contour.

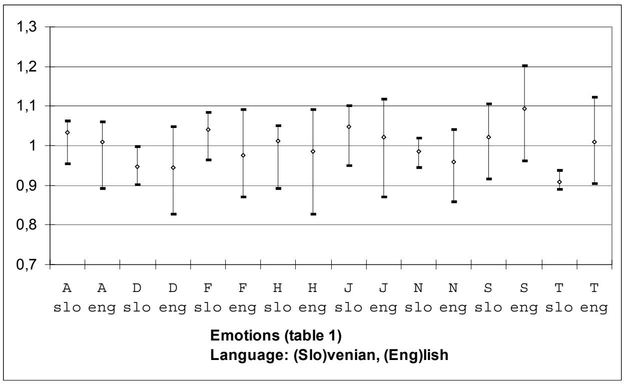

Figure 6: Mean value of pitch contour for Slovenian female and English female speaker, normalised to speaker's mean value of pitch contour. Abbreviations for emotions are given in Table 1.

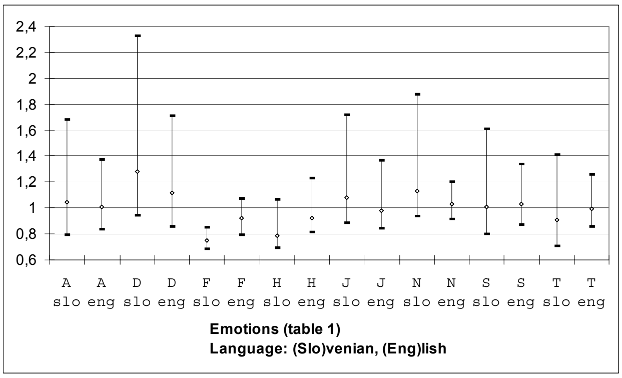

Figure 7: Duration of vowels for Slovenian female and English female speaker normalised to speaker's vowel duration. Abbreviations for emotions are given in Table 1.

Slovenian female has higher average value of mean value of pitch contour as English female speaker.

High values of emotional features that are calculated from pitch are encountered for emotions joy and surprise. Both speakers have low values of high level features that are calculated from pitch for emotions sadness and disgust.

English and Slovenian female speaker have high values of vowel duration for emotion disgust and low values for emotion fear (see figure 7) and fast neutral speaking style. Slovenian female speaker has lower average value of vowel duration as English female speaker.

Emotional features that are calculated from duration have high values for emotions disgust and sadness and low values for emotion surprise and fast neutral style.

6. Conclusion

Accomplished analysis of high level features has shown that the emotions anger, fear, joy and surprise have increased values for emotional features that are calculated from pitch and pitch derivative contours. Decreased values of these features were seen for emotions disgust and sadness. The selected features that were calculated from duration showed increased values for emotions disgust and sadness. Decreased values of these selected features were found for fast neutral speaking style. These results are in accordance with the findings reported in other works (Noad et. al., 1997; Pittam and Scherer, 1992; Pereira and Watson, 1998).

English female speaker had smallest deviation from average values of emotional features. This could indicate that English female speaker did dot express emotions as forceful as other speakers.

The prosody features between male and female speakers and between English and Slovenian speakers were found to be similar. Differences were found in average values of emotional features. For instance difference between male and female were found to be in the average value of mean value of pitch contour. On the other hand, deviation from average value of mean value of pitch contour for male and female speaker is similar among the emotions. Similar was true for emotional features that were calculated from duration. The analysis showed that deviation from average value of emotional features for the speakers used in the analysis is independent from language and gender of speakers.

New approach for analysis of emotional features enables consistent and objective analysis because it is data driven. It enables comparison of emotional features between different speakers and at the same time searches for new correlates of emotions in speech.

7. References

Cahn, J.E., (1990). Generating Expression in Synthesised Speech. *Master thesis, Massachusetts Institute of Technology Montero.*

Hozjan V., Kacic Z., Moreno A., Bonafonte A, Nogueiras A. (2002). *Proceedings of the Language Recourses and Evaluation 2002.*

Johnstone I. T. (1996). Emotional speech elicited using computer games. *Proceedings of the 4th International Conference on Spoken Language Processing*, 3, 1985-1988.

Koike, K., Suzuki, H., Saito, H., (1998). Prosodic parameters in Emotional Speech. *In Proceeding of the 6th International Conference on Spoken Language Processing.*

Monter, J.M., Gutierrez-arriola, J., Colas, J., Macias, J., Enriquez, E., Pardo, J.M., (1999). Development of an Emotional Speech Synthesizer in Spanish, *In the Proceeding of the Eurospeech '99*.

Noad, J.E., Whiteside, S.P., Green, P.D., (1997). A macroscopic analysis of an emotional speech corpus, *In the Proceeding of the Eurospeech '97*.

Pereira, C. (1996). Angry, happy, sad or plain neutral? The identification of vowel affect by hearing-aid users, *In Proceeding of the Sixth Australian International Conference on Speech Science and Technology*, Adelaide.

Pereira, C., Watson, C., (1998). Some Acoustic Characteristics of Emotion. *In Proceeding of the 6th International Conference on Spoken Language Processing,.*

Pittam, J., Scherer, K. (1992). The encoding of affect: a review and direction for future research. *In Proceeding of the Fourth Australian International Conference on Speech Science and Technology, Brisbane.*

Qian X. and Kimaresan R., (1996). A variable Frame Pitch Estimator and Test Results. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Atlanta, GA, pp. 228-231.

Sharma, R., Pavlovič, V.I., Huang T.S., (1998). Toward Multimodal Human-Computer Interface, *Proceedings of the IEEE*, Vol 86, No. 5, May.

Williams, C. E. and Stevens, K.N., (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, 52, 1238 -1250.

Young, S., Ollason, D., Valtchev, V., Woodland p., (1997). The HTK Book (for HTK 2.1). *Entropic Cambridge Research Laboratory, March.*