Review Article **Open Access**

Mohammad Rabiei and Alessandro Gasparetto

A Methodology for Recognition of Emotions Based on Speech Analysis, for Applications to Human-Robot Interaction. An Exploratory Study

Abstract: A system for recognition of emotions based on speech analysis can have interesting applications in human-robot interaction. In this paper, we carry out an exploratory study on the possibility to use a proposed methodology to recognize basic emotions (sadness, surprise, happiness, anger, fear and disgust) based on phonetic and acoustic properties of emotive speech with the minimal use of signal processing algorithms. We set up an experimental test, consisting of choosing three types of speakers, namely: (i) five adult European speakers, (ii) five Asian (Middle East) adult speakers and (iii) five adult American speakers. The speakers had to repeat 6 sentences in English (with durations typically between 1 s and 3 s) in order to emphasize rising-falling intonation and pitch movement. Intensity, peak and range of pitch and speech rate have been evaluated. The proposed methodology consists of generating and analyzing a graph of formant, pitch and intensity, using the open-source PRAAT program. From the experimental results, it was possible to recognize the basic emotions in most of the cases.

Keywords: Emotion, Human-Robot Interaction, Speech Analysis

Mohammad Rabiei, Alessandro Gasparetto: DIEGM - Università di Udine, Via delle Scienze 206 - 33100 Udine, Italy

1 Introduction

In recent years, a great deal of research has been done to automatically recognize emotions from human speech [1, 2]. Systems for automatic recognition of emotions can be considered as knowledge-based systems designed to analyze speech (words and sentences), which could be eventually employed to carry out simple emotional interaction between humans and robots.

In the literature, automatic recognition of emotions has been performed from any biological modality such as facial expressions, speech and gesture [3]. Most acoustic features that have been used for emotion recognition can be divided into two categories: prosodic and spectral. Prosodic features have been shown to deliver recognition, including intonation, accent, pitch, mute and rate of speech. Spectral features convey the frequency content of the speech signal, and provide complementary information to prosodic features. The spectral features are usually extracted over short frame duration. In addition we can also express energy features such as low-frequency and high-frequency domain in some kinds of behavior interaction [4]. Although technologies have been developed to improve the effectiveness of speech communication system, affective high-level human interaction with robots is still far from ideal.

Numerous studies have been done on speech emotion recognition, for various purposes. People have interacted with other humans in a social speech, beyond pure cognitive reasoning [5]. The definition of 'human emotion' is; "what is present in life but absent when people are emotionless"; this is the concept of pervasive emotion [6].

The performance of automatic systems for recognition of emotions based on speech analysis is still weak for spontaneous speech in general, and dialog in particular [7]. The task of speech emotion recognition is very challenging for the following reasons. First, it is not clear which speech features are the most suitable, in different cultures, to distinguish between emotions. The acoustic variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates is another factor directly affecting most speech features, such as pitch, and energy contours [8].

The most important works done on emotion recognition through speech analysis are the famous Ekman's and Fox's models. Ekman's model is based on six basic emotions model [9], while Fox's is a multi-level emotional model [10].

Each emotion corresponds to a different portion of the spoken utterance. However, it is very difficult to determine the boundaries between these portions. Another challenging issue is that a certain emotion is expressed in

a way that generally depends on the speaker, his or her culture and environment. Most literature work has focused on monolingual emotion classification [11]; however, beside the emotional states, the voice and speech signal can give information about a speaker's age, sex/gender and regional background.

Some other articles in this field are dealing with the analysis of emotions and different aspects of automatic recognition of emotions in speech, recognition of stress and perspective of automatic speech recognition [12, 13]. In the context of human-robot interaction, analyses of emotional speech expressions are generally aimed at the design of embodied conversational agents [14]. This predominantly relates to application in automated emotional dialog systems [12].

In a smaller group of studies some biological signals, such as: heart rate, brain waves and skin conductivity, are analyzed and the results are combined with those provided by speech analysis for recognition of emotions [15, 16].

Lee and Narayanan exploited and combined acoustic and lexicon information, in order to classify emotions into seven states; moreover, they implemented their results on automated dialog systems [17]. There are a number of studies by Enos that analyzed features in speech, such as: speech rate, pauses, voice quality, hesitations, speech errors, and investigated some frequency based parameters [18].

There is evidence, from human-robot interaction experiments, that language models for dialog can be improved by using additional sources of information and by improving the modeling of acoustic and prosodic features. In some recent work, the analysis of vowel and consonant articulation is carried out and formant frequencies are analyzed in connection to spoken language processing techniques [19].

By analyzing pitch and intensity patterns, Agnes Jacob used two key steps to recognize emotions in speech signals for an Indian woman: the first step was to find effective speech emotion features, the second one was to establish a proper model for emotion recognition from speech [20]. Some authors argue that accurately recognizing speech ultimately requires "mind modeling" [21, 22].

The paper is organized as follows: after the description of the state of the art in this field, we describe the methodology we employed, as well as the procedure for the experimental tests. Then, we deal with the problem of feature extraction from the experimental results, and we propose a set of rules for recognizing the emotions. Finally, the results of the application of the methodology are presented and discussed.

With respect to other works in the scientific literature, the methodology we propose in this paper makes minimal use of signal processing algorithms for feature extraction and emotion classification. Thus, the proposed algorithm is very suitable for implementation in real-time systems, since the computational load is very low indeed.

2 Methodology

Emotion recognition can have interesting applications in human-robot interaction, thus paving the way for a scenario where human-robot interaction will normally take place in the real world. When a speaker expresses an emotion while adhering to an inconspicuous intonation pattern, human listeners can nevertheless perceive the emotional information through the pitch and intensity of speech. On the other hand, our aim is to capture the diverse acoustic cues that are in the speech signal and to analyze their mutual relationship to the speaker's emotion. We propose a technique to recognize several basic emotions, namely sadness (SAD), anger (ANG), surprise (SUR), fear (FEA), happiness (HAP) and disgust (DIS), based on the analysis of phonetic and acoustic properties.

An experimental methodology was set up, consisting of three different databases built from speakers of different areas of the world. The first database includes five European adults in the age group from 25 to 35 years (2 women and 3 men; mean age 29) from different countries of the European Union (Spain, Italy, Belgium, Romania, and France), the second group contains five Asian (Middle East) adult speakers in the age group from 19 to 45 years (2 women and 3 men; mean age 31) and the third database contains recordings from American English speakers in the age group from 21 to 38 years (3 women and 2 men; mean age 28).

Six simple sentences of everyday life were chosen, namely: "What are you doing here?"- "Are you ready?"- "Come in"-"Thank you"-"You are welcome"-"Where are you?" The participants to the experiment had to repeat these six sentences for three times with a neutral (NEU) intonation, with 1 second of interval between each sentence, in order to distinguish rising-falling intonations and pitch movements.

Then, every participant had to repeat again three times the same six sentences, but with each one of the emotions listed above. All the sentences were recorded, thus obtaining 315 files, which were input to a dedicated program for speech analysis (Figure 1), which could pro-

vide the intensity, pitch (peak, range, values) alignment and speech rate of all the sentences.

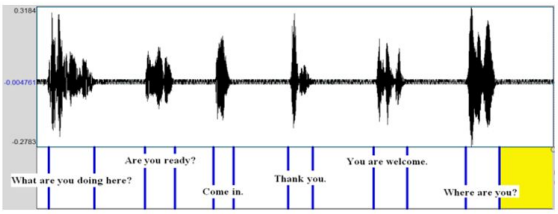

Fig. 1. Example of speech analysis with the PRAAT program.

The program used for speech analysis is the standard PRAAT program. PRAAT is an open-source software that is extensively used by researchers in the field of speech analysis [23]. The PRAAT software organizes the sound file into "frames" for analysis, computing four pitch values within one frame length. The segmented wave files were analyzed one at a time and the pitch contours were saved in separate files.

The technique for emotion recognition proposed in this paper is based on two steps; namely: 1) feature extraction and 2) rule definition.

3 Feature extraction

The most important step in recognition of emotions is the extraction of a meaningful and informative set of features.

We propose to neglect linguistic features and to focus on acoustic and prosodic features, more specifically on pitch, duration and intensity.

Voice characteristics at the prosodic level, including intonation and intensity patterns, carry important features for emotional states. Hence, prosody clues such as pitch and speech intensity can be used to model different emotions and the fundamental frequency pitch contours, pitch values, pitch range, as well as the average intensity can enable one to build a classification of various emotion types. For example, high values of pitch are correlated with happiness and anger, whereas sadness and boredom are associated with low pitch values [25].

Three types of features were considered: pitch (range, value and peak), intensity (energy) and rate speech; hence, the graphs of formant, pitch and intensity were analyzed.

Pitch features are often made perceptually more adequate by logarithmic/semitone function, or normalization with respect to some (speaker specific) baseline. Pitch is a fundamental acoustic feature of speech and needs to be determined during the process of speech analysis [26]. The modulation of pitch plays a prominent role in everyday communication. Pitch extraction can influence the performance of emotion recognition.

The ***Pitch value*** of a sound is the length of the time interval when the sound signal is higher than the average. The ***pitch peak*** of a sound is the maximum intensity (peak) of the signal. The ***pitch range*** is defined as the ratio between the highest and lowest values of intensity of the sound signal.

Intensity features usually represent the loudness (energy) of a sound as perceived by the human ears, based on the amplitude in different intervals [27].

Energy is the square of the amplitude multiplied by the duration of the sound.

Voice quality is a complicated issue in itself, since there are many different measures of voice quality, mostly clinical in origin and mostly evaluated for constant vowels only, though once again standardization in this area is lacking [27].

The ***spectrum*** is characterized by formants (spectral maxima) modeling spoken content. Higher formants amplitude also represents speaker position and characteristics [28].

Non-linguistic vocalizations are non-verbal phenomena, such as breathing, mute and laughter [29].

The ***speech rate*** specifies the speaking rate in words per minute, a rate that varies somewhat by language, but is nonetheless widely supported by speech synthesizers [29].

One of the important questions in the field is how many and which features to choose for automatic recognition of emotions. The answer to this question have main role to improve performance and reliability but also to obtain more efficient models, also in terms of processing speed and memory requirements. Ideally, feature selection methods should not only reveal single or most relevant attributes. Features such as pitch, intensity, duration, formant, voice quality and speech rate (SR) are most common and standard features in emotion recognition.

In this research we propose a methodology for emotion detection based on analysis of the plots provided by the speech analysis software. The features our technique takes into consideration are: pitch, intensity, formant and speech rate.

3.1 Pitch analysis

Pitch is a fundamental acoustic feature of speech and needs to be determined during the process of speech syn-

thesis. The modulation of pitch plays a prominent role in everyday communication fulfilling very different functions, like contributing to the segmentation of speech into syntactic and informational units, specifying the modality of the sentence, regulating the speaker–listener interaction, expressing the attitudinal and emotional state of the speaker, and many others. Automatic pitch stylization is an important resource for researchers working both on prosody and speech technologies [30]. Pitch range was considered as a necessary feature for emotion recognition. Pitch contour extraction was done using the PRAAT software. Figure 2 shows some pitch plots for the sentences spoken by the participants to the experiment (the clearest results were chosen, among the three repetitions made by each participant).

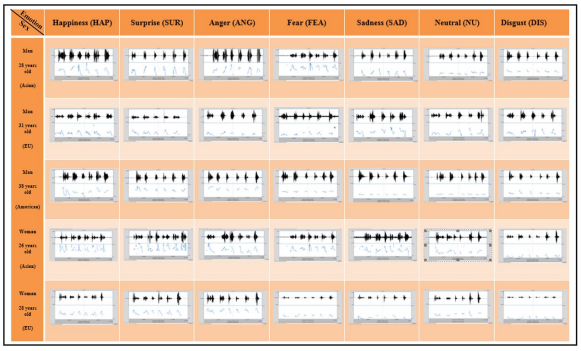

Fig. 2. Some pitch results from the 15 interviewed persons (Europeans, Americans, and Asians).

As depicted in Figure 3, 4 and 5, the pitch contours under positive valence emotions (such as surprise and happiness) are similar: the value of the pitch at the end of the sentence is lower than the value at the beginning, but surprise has a bigger pitch value. we can see that the highest pitch value is for surprise and the lowest corresponds to disgust. Also, we can see that the pitch peak under positive valence emotions is sharper among Asian speakers, while European and American speakers more or less have similar pitch contours under positive valence emotions. Happiness and anger have the highest average pitch peak for European speakers (see Figure 6) while sadness has the lowest pitch peak. In our experiment, we can also see that surprise and anger for Asian and American speakers have the highest average pitch peak (see Figure 4, 5, 7 and 8).

Among the negative valence emotions, anger has the highest pitch peak (see Figure 6, 7 and 8). Sadness decreases sharply for Asian and American speakers, but sadness slop decreases slowly.

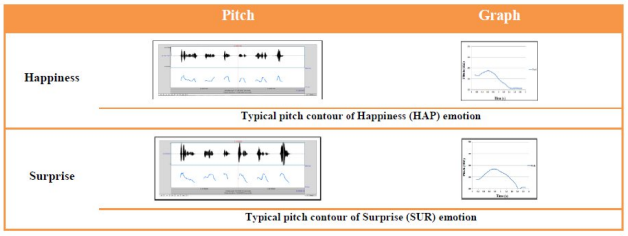

Fig. 3. European pitch contours for Happiness and Surprise.

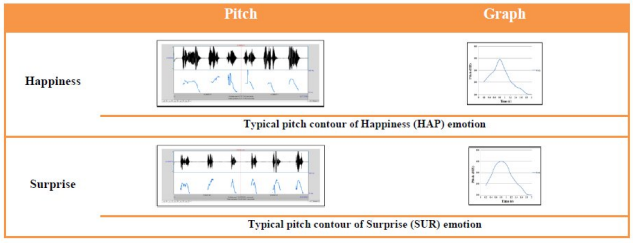

Fig. 4. Asian pitch contours for Happiness and Surprise.

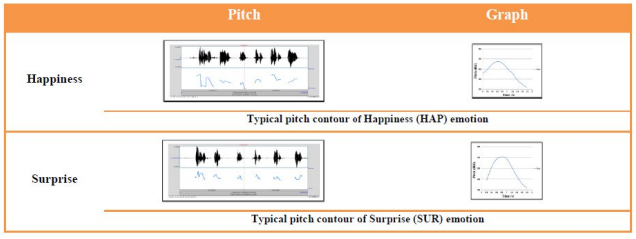

Fig. 5. American pitch contours for Happiness and Surprise.

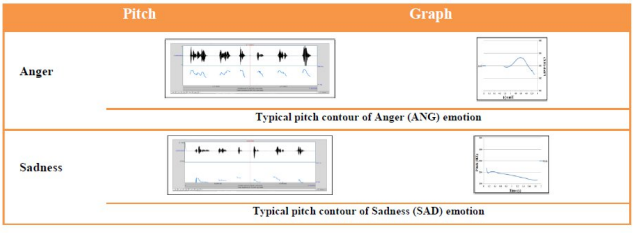

Fig. 6. European pitch contours for Anger and Sadness.

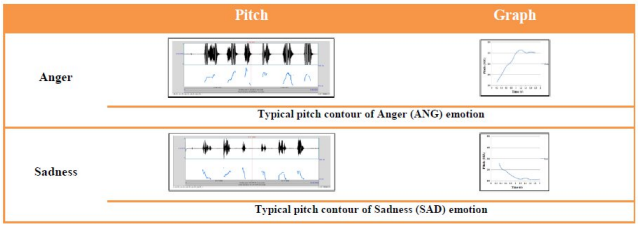

Fig. 7. Asian pitch contours for Anger and Sadness.

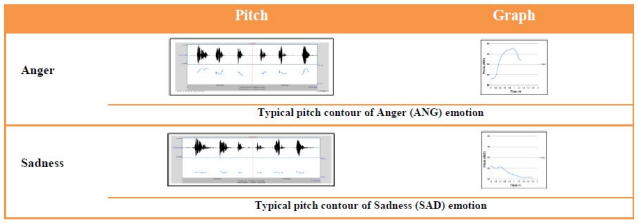

Fig. 8. American pitch contours for, Anger and Sadness.

If we compare sadness and neutral for all groups of speaker (Figure 6, 7, 8 and 12) the neutral emotion does not have a distinct peak and is similar to sadness; however, sadness has lower ending pitch signals.

Asian speakers were more sensitive to sad emotion, while the pitch graphs of Americans and Europeans were similar. Anger is associated with the highest energy for Asian and American speakers but for Asian speakers the anger slope decreases slowly, while sadness is associated with the lowest energy for Asian and European speakers.

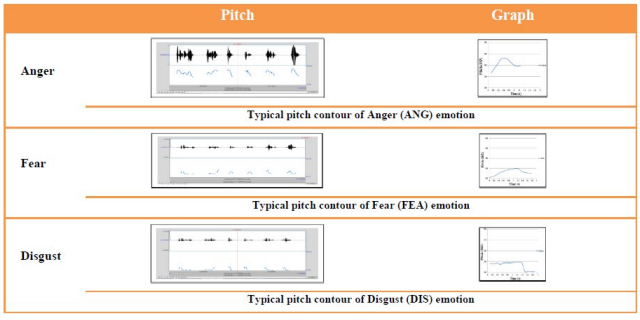

Fig. 9. European pitch contours for Anger, Fear and Disgust.

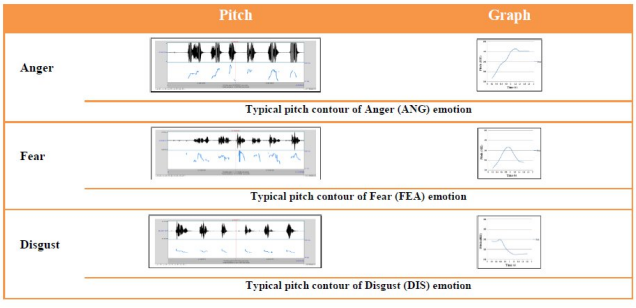

Fig. 10. Asian pitch contours for anger, fear and disgust.

Even though no universal similarities are observed among negative valence emotions, similarities are noted between certain utterances under anger and fear. In Figure 9, 10 and 11 anger is characterized by a rising peak followed by either a decrease or a leveling out of the pitch values and the utterance duration is observed to be small. In almost all utterances under anger and fear, the pitch increases to a peak and then decreases slightly left-skewed. European and American speakers more or less have similar pitch contours under fear emotion.

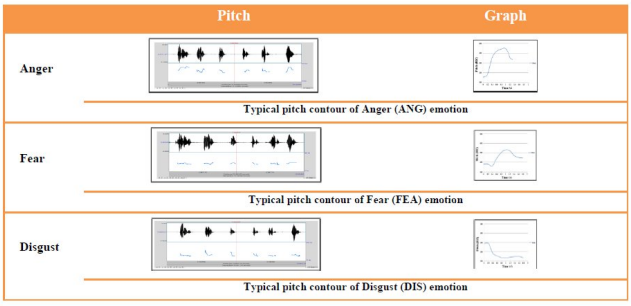

Fig. 11. American pitch contours for Anger, Fear and Disgust.

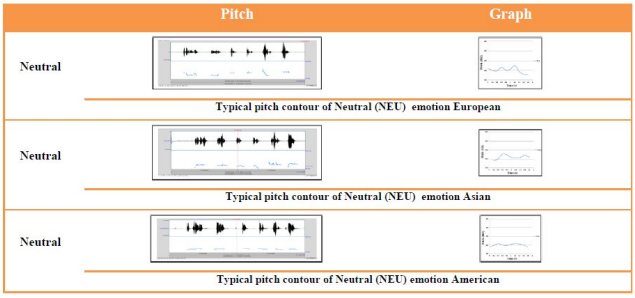

Fig. 12. European, Asian and American pitch contours for Neutral.

In Figures 10 and 11 it can be seen that the highest mean pitch values are for American speakers, while Asians have sharper pitch peaks. Pitch values and speech rate are connected together. We can see that usually the speech rate of American speakers is higher than Asian and European speakers. As depicted in Figure 12, the beginning and ending of pitches in neutral emotion for Americans after rising and falling have similar frequencies. This is probably due to the fact that the mother language of American speakers was English, while Europeans and Asians (whose mother language was not English) show a bit of stress in neutral speech.

3.2 Formant analysis

Formants are the meaningful frequency components of human speech and contents of the vowel sounds. We can change the position of the formants by moving around the tongue and the lip muscles so as to show the emotion in speech. In the PRAAT software the maximum value of the formant should be set to about 5000 Hz for a male speaker, 5500 Hz for a female speaker and even higher for children [31]. The plot of the formant displays the amplitude of the frequency components of the signal over time. For most analyses of human speech, we may want to extract 5 formants per frame. As depicted in Figure 13, 14 and 15, the formant contour in anger and happiness for European speakers has the highest power, while we have the lowest spectral power in fear. Formant contour in Figure 14, 15 explain that anger, fear and happiness have the highest power for Asians and Americans, while we have a lot of wave and formant dots in the fear plot. Asians and Europeans have the lowest spectral power in sadness, while Americans have the lowest spectral power in neutral emotion.

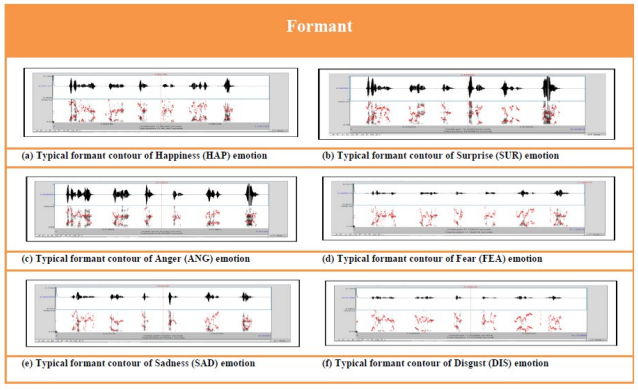

Fig. 13. European speakers typical formant contour of basic emotions.

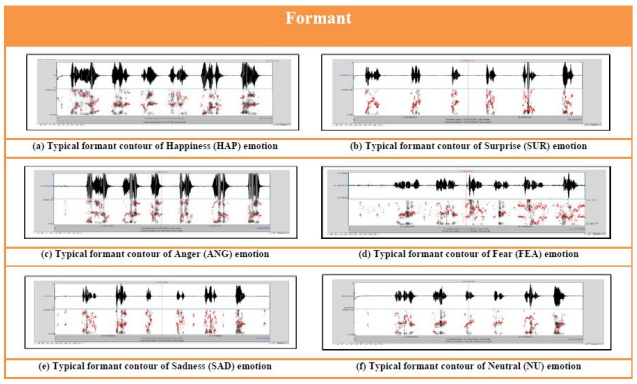

Fig. 14. Asian speakers typical formant contour of basic emotions.

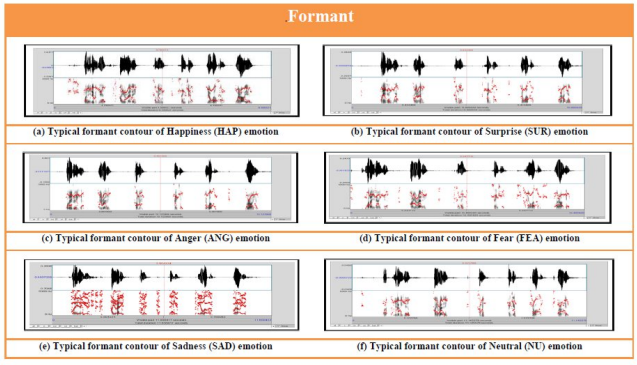

Fig. 15. American speakers typical formant contour of basic emotions.

3.3 Intensity

Sound or acoustic intensity is defined as the sound power and is measured in dB. The typical context in this field is the listener's location for the measurement of sound intensity. Sound intensity is a specifically defined quantity and is very sensitive to the location of the microphone. In our experiments using the PRAAT software, we put the microphone at a distance of 30 cm form each participant. In terms of intensity, as it can be seen in Figure 16, 17 and 18, when we have strong power on the source of sound signals, the energy and the intensity of the sound increase. Anger and surprise for European speakers have the highest energy and intensity, while neutral and sadness have the lowest intensity.

For Asian speakers, as it can be seen in Figure 17, anger and happiness have the highest energy and intensity, while fear has the lowest intensity.

For American speakers, as it can be seen in Figure 18, anger, surprise and happiness have the highest energy and intensity, while fear has the lowest intensity.

It is straightforward to infer that the difference between results is due to the difference between the cultures to which the speakers belong. Categorizing the emotions into "high intensity" or "low intensity" can be of great help to increase and design algorithms for emotion recognition.

3.4 Speech rate

For emotion recognition in sound signals, speech rate is an important factor as well. Human listeners are able to understand both fast and slow speech. Speech recognizers have been implemented for Human-Robot interaction.

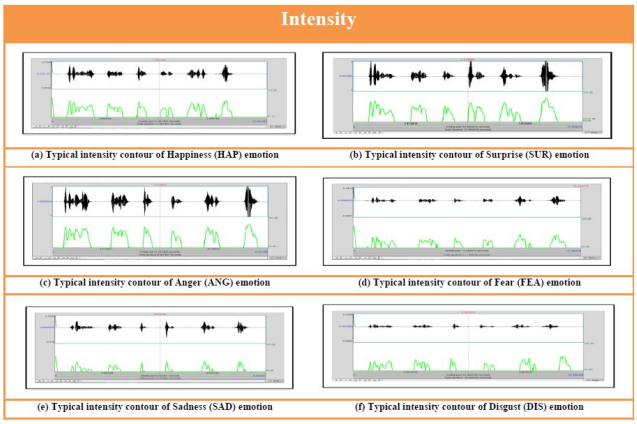

Fig. 16. European speakers typical intensity contour of basic emotions.

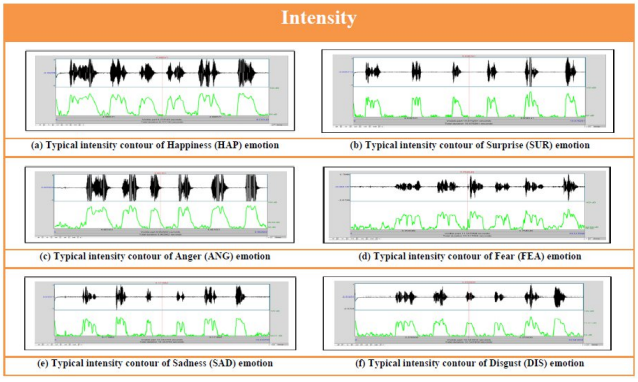

Fig. 17. Asian speakers typical intensity contour of basic emotions.

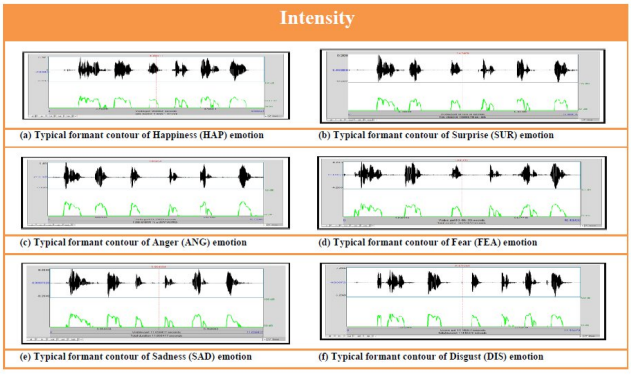

Fig. 18. American speakers typical intensity contour of basic emotions.

Speech rate is typically defined as the number of words spoken divided by the time of speech. The PRAAT software expresses speech rate in seconds, which means the time taken to pronounce the analyzed sentence. A notable result (see Table 1) is that anger and fear have the lowest speech rate for European speakers, meaning that the sentences pronounced with anger or fear are pronounced faster, while happiness has the highest speech rate. Sadness and disgust have the lowest speech rate for Asian speakers, while anger and happiness have the highest speech rate: this results is probably due to the fact that Asian people have bigger emotional reaction to happiness and anger. For American speakers anger and disgust have the lowest speech rate, while happiness and fear have the highest speech rate. In general, happiness and surprise have the highest speech rate, while anger and sadness have the lowest speech rate. Moreover, Americans have the highest speech rate.

4 Rule extraction

Researchers mostly focus on defining a universal set of features that convey emotional clues and try to develop classifiers that efficiently model these features. The system for recognition of emotion needed to present methods for discovering emotions, modeling and evaluating the results. For practical purposes, the important outcome of this section is the discovery of which features are the most informative and meaningful for recognition of emotions. We defined rules for emotion recognition based on human sound signals and evaluated and tested these rules. Some examples of these rules are shown in the following:

Observation1. Anger and disgust are associated with low speech rate, but anger is associated with the highest energy, while disgust (and sadness) are associated with the lowest energy.

Observation2. Anger was found to have the highest pitch values while disgust has the lowest.

Observation3. Positive valence emotion (happiness and surprise) have right-skewed pitch contours, while negative valence ones (anger and fear) have slightly left-skewed pitch contours. Neutral and sadness have the lowest ending pitch contours.

Observation4. Disgust is characterized by a smooth decline in the average pitch from a higher level, like neutral. However, disgust decreases sharply, neutral does not.

Observation5. Happiness and anger have the highest pitch peak but anger has higher average pitch peak and higher energy (actually, anger has the highest intensity).

Table 1. Speech rate of basic emotions (average of the 15 experimental tests) for European, Asian and American.

Emotion Quality	Speech rate for European speakers	Speech rate for Asian speakers	Speech rate for American speakers
Happiness (HAP)	2.0921 s	1.9457 s	1.9931 s
Surprise (SUR)	1.6439 s	1.7001 s	1.6128 s
Anger (ANG)	1.3176 s	1.8832 s	1.2204 s
Fear (FEA)	1.4863 s	1.7121 s	1.6585 s
Disgust (DIS)	1.5521 s	1.4401 s	1.5736 s
Sadness (SAD)	1.7071 s	1.3764 s	1.4750 s
Neutral (NU)	1.6889 s	1.5343 s	1.5158 s
SUM RATE	11.489 S	11.616 S	10.482 S

A block diagram of the proposed for automatic multi-level emotion recognition system is given in Figure 19.

The input of the system is the file obtained from the PRAAT software. The algorithm described in the foregoing analyzes the plots of pitch, intensity and formant, thus recognizing if the emotion belongs to the category of "high intensity" or "low intensity". If the speech falls into the "high intensity" category, it is further analyzed in order to distinguish between fear, anger, surprise and happiness. In the same way, if the speech falls into the "low intensity" emotion, it is further analyzed in order to distinguish between neutral and sad. We can use speech rate and the graphs of pitch signals in low intensity categories to distinguish between neutral and sadness emotion. If we compare sadness and neutral, sadness emotion has lower ending pitch signals. High intensity category comprises anger, fear, surprise and happiness. In order to distinguish between these emotions, we must first draw our attention to the shape of signals. Namely, if the shape of signal is "left skewed" and the ending of the signal is higher than the beginning, the emotion must be fear or anger emotion. In order to distinguish between fear and anger, the algorithm must compare the intensity: anger emotion has the highest intensity, thus it is easily distinguishable. If the shape of the signal is "right skewed", it must further analyzed in order to distinguish between surprise and happiness. To this aim, the algorithm must check the pitch value and intensity: happiness has the highest pitch range and pitch peak while surprise has the highest pitch value. If the speech does not belong to any of the aforementioned emotions, it is classified as disgust.

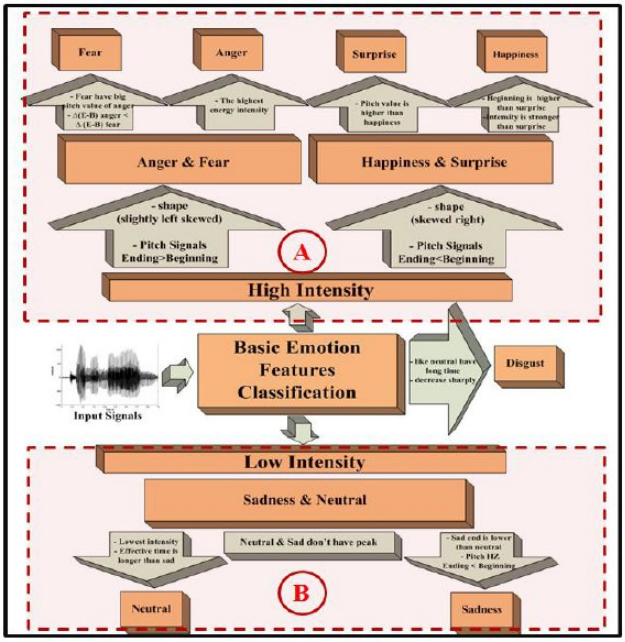

Fig. 19. Model of the system for emotion recognition.

5 Results and Discussions

In almost any category of emotion we have successfully identified certain characteristic features and these formed the basis for classification of different emotions. The typical pitch and intensity contours characterizing each of the basic emotions are as presented in Figure 19. Happiness and surprise have the highest pitch range and pitch peak. In pitch peak and intensity analysis, happiness and anger are distinguished faster than other emotions for European speakers, while in order to distinguish fear and disgust, the algorithm must check all the acoustic features. In pitch value analysis, surprise, happiness and fear can be distinguished quicker than other emotions. For Asian speak-

ers in pitch peak, happiness and anger are distinguished faster than other emotions. Also, anger has the highest and sadness has the lowest range of intensity for Asian speakers. The procedure for emotion recognition from speech can be implemented using a Likert type scale (see Table 2), which categorizes the basic emotions based on discrete values of pitch peak, pitch range, pitch value, intensity and speech rate. Perceptual listening tests had been conducted to verify the correctness of emotional content in the recordings under each of the seven categories of emotion. Human experts (3 persons) listened to the sample sentences 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 five point scale ranging from excellent to bad through very good, good and fair.

From the validation results (see Table 3) it appears that fear, sadness and disgust are not so easy to distinguish from speech signals; however, this is also true for humans. This kind of emotions could be more easily detected if the speech-based algorithm proposed in this work is combined with an emotion recognition algorithm based on face analysis. In future work, we intend to study such a hybrid algorithm.

Figure 20 shows the characteristics of the six basic emotions in a bar diagram. This representation is complementary to those described above.

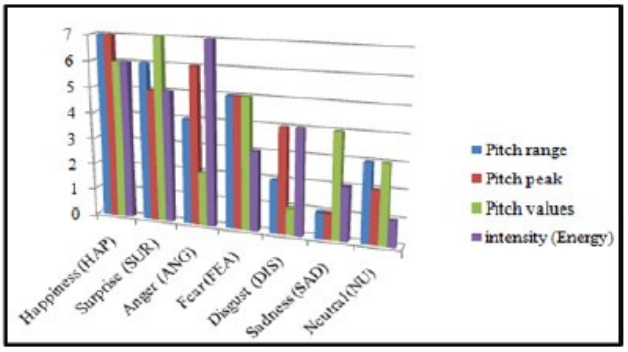

Fig. 20. Bar chart of the six basic emotions.

Figures 21 and 22 show the results arranged in three-dimensional graphs using a discrete approach for the classification of emotions. For instance, Figure 21 shows the location of the six basic emotions in the three-dimensional graph whose axes are: the pitch peak, the pitch range and the pitch value.

Figure 22 shows the location of the six basic emotions in the three-dimensional graph whose axes are: the total pitch score, the intensity and the speech rate. Other three-dimensional graphs can be built by selecting a different set of three features.

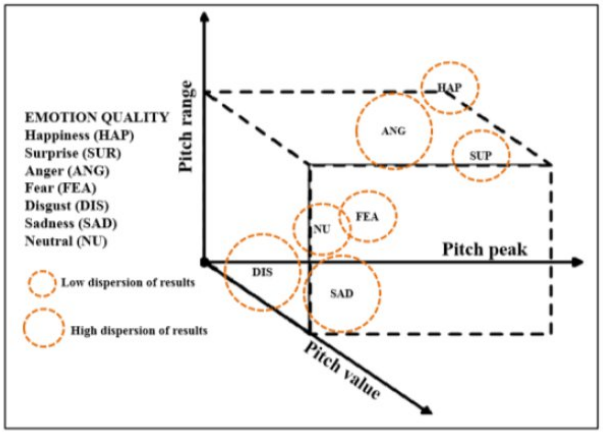

Fig. 21. The location of the emotions in the three-dimensional graph with axes: pitch (range, peak and value).

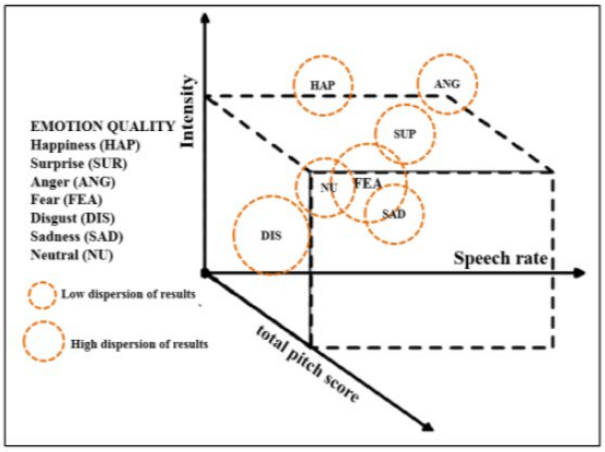

Fig. 22. The location of the basic emotions in the three-dimensional graph with axes: total pitch score, intensity and speech rate.

From the analysis of both Figures, it results that when pitch range, pitch value and pitch peak are considered, happiness, surprise and anger can be distinguished faster than other emotions, and with a high degree of accuracy. In order to distinguish other emotions, we can use other features such as intensity and speech rate, as well as the total pitch score (Figure 22). In this way, we can distinguish all the six basic emotions more easily, because the boundaries between emotions are more distinct.

Table 2. Likert type scale for emotion recognition.

EMOTION QUALITY	Difference between highest and lowest of signals		Top of the signal shape		For long time is higher than the average		sound power	Model Result	Total quality distinguish					Total Result
	Pitch range		Pitch peak		Pitch values		intensity (Energy)	SUM	Likert type scale					SUM
Happiness	European	7	European	7	European	6	6	945	EX	VG	G	F	B	EX
	Asian	6	Asian	6	Asian	4	6		EX	VG	G	F	B	
	American	6	American	5	American	5	5		EX	VG	G	F	B	
Surprise	European	6	European	7	European	5	7	880	EX	VG	G	F	B	EX
	Asian	5	Asian	5	Asian	6	3		EX	VG	G	F	B	
	American	5	American	6	American	6	6		EX	VG	G	F	B	
Anger	European	4	European	6	European	2	7	815	EX	VG	G	F	B	VG
	Asian	7	Asian	7	Asian	2	7		EX	VG	G	F	B	
	American	7	American	7	American	3	7		EX	VG	G	F	B	
Fear	European	5	European	3	European	5	3	640	EX	VG	G	F	B	G
	Asian	4	Asian	4	Asian	7	5		EX	VG	G	F	B	
	American	3	American	4	American	5	2		EX	VG	G	F	B	
Disgust	European	2	European	4	European	1	4	425	EX	VG	G	F	B	F
	Asian	3	Asian	3	Asian	1	1		EX	VG	G	F	B	
	American	4	American	3	American	1	1		EX	VG	G	F	B	
Sadness	European	1	European	1	European	4	2	590	EX	VG	G	F	B	G
	Asian	2	Asian	2	Asian	3	4		EX	VG	G	F	B	
	American	2	American	1	American	2	4		EX	VG	G	F	B	
Neutral	European	3	European	2	European	3	1	680	EX	VG	G	F	B	G
	Asian	1	Asian	1	Asian	5	2		EX	VG	G	F	B	
	American	1	American	2	American	7	3		EX	VG	G	F	B	

Table 3. Percentage of emotions recognized correctly.

Emotion quality	Results of model	Human expert
Happiness (HAP)	94.5%	Excellent
Surprise (SUR)	88%	Excellent
Anger (ANG)	81.5%	Very good
Fear (FEA)	64%	Good
Disgust (DIS)	42.5%	Fair
Sadness (SAD)	59%	Good
Neutral (NU)	68%	Good

6 Conclusion

In this paper, we proposed a methodology for recognition of emotions, based on different speech features, which can be employed for human-robot interaction. The features that are taken into account are prosodic features, such as pitch, intensity and formant.

The proposed technique is based on a first analysis of pitch graph contours (namely: pitch peak, pitch value, pitch range), followed by a second analysis of the intensity and the speech rate in the dialogue, which is considered complementary to the first analysis in order to recognize all types of emotions.

The open source PRAAT soft was used to in the process of carrying out both analyses simultaneously. In the presented model, emotions are first categorized in two main classes; namely high and low intensity emotions, then a more precise distinction is performed within each category.

One of the challenges in the research in this field of analysis of emotional speech is the difficulty in creating a worldwide database of emotional speech.

An experimental test, with the participation of five European, five Asian and five American individuals, was set up, in order to experimentally validate the proposed methodology. The results of this exploratory study show that it could be feasible to build a technique which is effective in recognizing emotions. Thus the results of this study provide a better understanding on the manifestation and production of basic emotions and can be useful for the purpose of analysis and synthesis of emotional speech for technical researchers, social psychologists and human-robot interaction.

In future work, we intend to combine the proposed algorithm, based on speech analysis, with a technique based on face analysis, in order to design a hybrid technique for emotion recognition, to be employed in human-robot interaction.

References

[1] D. Ververidis, C. Kotropoulos, Emotional speech recognition – resources features and methods, Journal of Speech Communication, 48, 2006, 1162–1181

[2] F. Metze, T. Polzehl and M. Wagner, Fusion of acoustic and linguistic speech features for emotion detection, Procedings of. International Semantic Computing Conference on (14-16 Sep 2009, Berleley, CA, USA)

[3] F. Metze, A. Batliner, F. Eyben, T. Polzehl and B. Schuller, Emotion recognition using imperfect speech recognition. Proceedings.Annual Conference. of the International Speech Communication Association (2009, Makuhari, Japan), 1-6

[4] C. Peter, Affect and Emotion in Human–Computer Interaction, From Theory to Applications, 6 (2008), 48-68

[5] B. Schuller, Recognizing realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge - Speech Communication, 53 (2011), 1062–1087

[6] R. Cowie, N. Sussman and A. Ben-Ze'e, Emotions: concepts and definitions, The HUMAINE Handbook, Springer, 2010

[7] N.G. Ward, A Vega, Studies in the use of time into utterance as a predictive feature for language modeling, Technical Report UTEP-CS- 22, University of Texas, Department of Computer Science, 2010

[8] Wu. Dongrui, D. Narayanan and S. Shrikanth, Acoustic feature analysis in speech emotion primitive's estimation, Procedings of. International Inter Speech Conference, (2010, Makuhari, Chiba, Japan), 2010, 785-788

[9] P. Ekman, Are there basic emotions? Psychological Review, 99 (1992), 550–553

[10] N. Fox, If it's not left, it's right: Electroencephalograph asymmetry and the development of emotion, Am. Psychol, 46 (1991), 863–872

[11] P. Ekman, Darwin's Compassionate View of Human Nature, JAMA February 10, 303 (2010), 557–558

[12] T. Bnziger, K. Scherer, The role of intonation in emotional expressions, Speech Common, 46 (2005), 252–267

[13] A. S. Cohen, K. S. Minor and G. Najolia, laboratory-based procedure for measuring emotional expression from natural speech, Journal of Behavior Research Methods, Instruments and Computers, 41 (2009), 204–212

[14] S. Paulmann, S. Jessen and S.A. Kotz, Investigating the multimodal nature of human communication, Journal of Psychophysiol. 23,2 (2009), 63–76

[15] M. Swerts, E. Krahmer, Facial expression and prosodic prominence: Effects of modality and facial area. Journal of Phonetics, 36,2 (2008), 219–238

[16] T. Sobol-Shikler, P. Robinson, Classi?cation of complex information: Inference of co-occurring affective states from their expressions in speech, IEEE Trans, 32, 7 (2010), 1284–1297

[17] Lee. Sungbok, S. Narayanan, Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection, IEEE Transactions on,17,4 (2009), 582-596

[18] F. Enos, Detecting Deception in Speech, PhD thesis, Submitted to Columbia University, 2010, URL http://www.cs.columbia.edu/wfrank/enos_thesis.pdf

[19] T.F. Yap, J. Epps, E. Ambikairajah and E.H.C. Choi, Formant frequencies under cognitive load: effects and classification, EURASIP Journal on Advances in Signal Processing (In press), ID: 219253

[20] A. Jacob, P. Mythili, socio friendly approach to the analysis of emotive speech, Procedings of International Conference on Communication Technology and System Design, 2012, 577–583

[21] N.G. Ward, A. Vega, Towards the use of inferred cognitive states in language modeling, 11th IEEE Workshop on Automatic Speech Recognition and Understanding (2009, Merano, Italy), IEEE, 2010, 323–326

[22] L. I. Perlovsky, Toward physics of the mind: Concepts, emotions, consciousness, and symbols, Journal of Physics of Life Reviews, 3 (2006b), 22–55

[23] P. Boersma, D. Weenink, Praat (Version 4.5.25)

[24] , Latest version available for download from www.praat.org

[25] Z. Zeng, M. Pantic, G.I. Roisman and T.S. Huang, A survey of affect recognition methods: Audio, Visual, and Spontaneous expressions, IEEE Trans, Pattern Anal. Mach. Intell, 31 (2009), 39–58

[26] T. Polzehl, A. Schmitt and F. Metze, Salient features for anger recognition in German and English IVR systems, Spoken Dialogue Systems Technology and Design, Springer, Boston, 2010, 83-105

[27] M. Grimm, K. Kroschel, E. Mower and S. Narayanan, Primitives based evaluation and estimation of emotions in speech, Journal of Speech Communication, 49 (2007a), 787–800

[28] J. Kaiser, On a simple algorithm to calculate the 'energy' of a signal, International Conference on Acoustics, Speech and Signal Processing, 1 (1990), 381–384

[29] B. Kreifelts, T. Ethofer, T. Shiozawa, D Wildgruber and W. Grodd, Cerebral representation of non-verbal emotional perception, fMRIreveals audiovisual integration area between voice- and face-sensitive regions in the superior temporal sulcus, Journal of Neuropsychologia 47, 14 (2009), 3059–3066

[30] D. Bitouk, R. Verma and A. Nenkova, Class-level spectral features for emotion recognition, Journal of Speech Communication, 52 (2010), 613–625

[31] L. Zhang, J. Barnden, Affect sensing using linguistic, semantic and cognitive cues in multi-threaded improvisational dialogue, Journal of Cognitive Computation, 4 (2012), 436–459

Received October 16, 2013; accepted February 28, 2014.