Automatic Emotion Recognition in Speech: Possibilities and Significance

Milana Bojanić and Vlado Delić

Abstract—**Automatic speech recognition and spoken language understanding are crucial steps towards a natural human-machine interaction. The main task of the speech communication process is the recognition of the word sequence, but the recognition of prosody, emotion and stress tags may be of particular importance as well. This paper discusses the possibilities of recognition emotion from speech signal in order to improve ASR, and also provides the analysis of acoustic features that can be used for the detection of speaker's emotion and stress. The paper also provides a short overview of emotion and stress classification techniques. The importance and place of emotional speech recognition is shown in the domain of human-computer interactive systems and transaction communication model. The directions for future work are given at the end of this work.**

Index Terms—**Emotional speech recognition, stress, ASR.**

I. Introduction

The advent of new technologies and the introduction of interactive systems have increased the demand for more natural man-machine communication − in the way people communicate among themselves. Using voices people express their emotions. In order to achieve more natural verbal communication between humans and machines, it is necessary to recognize the mood of the speaker during automatic speech recognition (ASR) as well as to generate emotionally "colored" speech during its synthesis (text-to-speech systems, TTS). During the ASR, the recognition of emotion can be useful for proper handling of a man-machine dialogue. For example, detection of speaker's impatience, irritability or frustration, will help to appropriately redirect the dialogue (ticket reservation systems, call centers) [1].

Emotional speech recognition (ESR) strives towards automatic identification of emotional or psychological state of the individual based upon analysis of individual's speech. The mood and emotional condition of the speaker belong to paralinguistic aspects of verbal interaction. Analysis of these elements of verbal man-machine communication is necessary for successful implementation of ASR, regarding spontaneous dialogue between humans and machines [2].

Even though the emotional state can be manifested at the semantic level, the emotional content of the speech is contained in prosodic features to a considerable extent. While the classic ASR is based on correct recognition of word strings, for natural language processing and dialogue systems it is necessary to understand the context of the speech. In that case, the prosody and the emotional content can play an important role.

First studies in this field were conducted in the mid 80's of the last century using statistical properties of certain acoustic features. Subsequent progress of computer architectures allowed the use of more complex algorithms for ESR implementation. Today's studies are focused on combining classifiers in order to improve the efficiency of emotion classification within real applications [3].

There are three types of speech in emotional speech data collections [4]: natural, simulated (acted) and elicited speech. Natural speech is spontaneous speech where all emotions are real. Simulated speech is speech most often expressed by professional actors in order to invoke certain emotional states. In elicited speech the emotions are induced, for example by showing an adequate audiovisual material to the examinees. Simulated speech is most reliable for ESR because the professional actors express emotions vividly, with large amplitude and great power. Additional signals recorded during the data collection are (most often) laryngograph, heart beat rate, blood pressure, and the facial expressions of the speaker [3]. The majority of emotional speech data collections contains 5-6 emotional states alongside with neutral speech. The most frequent emotions are: anger, fear, joy, sadness, disgust, surprise, boredom and similar. Data collections can contain radio and TV excerpts and also recorded conversations with psychologists and phoneticians.

The acoustic features which reflect emotions in speech signal will be presented in the next section. The third section will then briefly explain the techniques of emotion classification, while the fourth section will cover characteristics of speech under stress. The fifth section shows the role and significance of emotion detection with regard to the transaction communication model, including the appropriate example of human–machine dialogue. The sixth section represents the conclusion of this paper and proposes directions for further research.

This paper represents a part of a research which is supported through the technological project of the Ministry of Science of Republic Serbia, "Human–machine Speech Communication" TR11001.

M. Bojanić and V. Delić are with the Faculty of Technical Sciences, Trg Dositeja Obradovića 5, Novi Sad, Serbia.

II. Acoustic Features of Emotional Speech

Emotions in speech are expressed through variation of speech characteristics on three levels: (1) prosodic or suprasegmental level through specific frequency, intensity and duration changes, (2) segmental level (changes in articulation quality) and (3) intrasegmental level (general voice quality, whose acoustic correlates are glottal pulse shape and distribution of its spectral energy, amplitude variations (shimmer), frequency variations (jitter)) [4]. Special attention will be paid to short-term acoustic features used in ESR. Short-term features are derived on frame basis:

$$f_s(n;m) = s(n)w(m-n) \tag{1}$$

where *s(n)* is the speech signal and *w(m-n)* is the window of length $N_{w.}$

Furthermore, several such features will be described, of which, according to many authors, the most important are pitch (fundamental frequency), energy of speech signal, vocal tract features and harmonics in the spectrum.

Pitch, fundamental frequency of phonation F_0, is the vibration rate of vocal cords during phonation. The emotional state of speaker affects the tension of vocal cords and the subglottal air pressure, which ultimately affects the pitch. This is why many authors consider it as the most important prosodic feature for ESR [5]. Numbers of algorithms have been developed for pitch estimation. Here will be presented the widely spread autocorrelation method [6]. First, the signal is low filtered at 900Hz, and then it is segmented to short time frames of speech which are later clipped. The clipping is a nonlinear procedure which should prevent the influence of the first formant on the pitch. The clipped signal looks like this:

$$\hat{f}_s(n;m) = \begin{cases} f_s(n;m) - C_{thr} & if \ \left|f_s(n;m)\right| \geq C_{thr} \\ 0 & if \ \left|f_s(n;m)\right| < C_{thr} \end{cases} \tag{2}$$

C_{thr} is set at 30% of the maximum value of $f_s(n;m)$. After calculation of autocorrelation value

$$r_s(\eta;m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} \hat{f}_s(n;m)\, \hat{f}_s(n-\eta;m) \tag{3}$$

the fundamental frequency (*pitch*) of the frame which ends at moment *m* can be estimated by

$$\hat{F}_0(m) = \frac{F_s}{N_w} \ \arg\max_{\eta}\left\{\left|r(\eta;m)\right|\right\}_{\eta = N_w(F_l/F_s)}^{\eta = N_w(F_h/F_s)} \tag{4}$$

where $F_s = 8000Hz$ is the sampling frequency, while F_l and F_h are the lowest and the highest pitch frequencies which humans can perceive. Their typical values are 50 Hz and 500 Hz, respectively. It is necessary to note that this is a relatively wide bandwidth which includes several octaves. This fact additionally complicates the task of automatic detection of fundamental frequency i.e. the pitch and allows mistaken detection of twice as high, or twice as low F_0. Except by the pitch, glottal waveform is characterized by air velocity through glottis, whose measurement is based on the maximum value of autocorrelation of clipped signal frames. One of the insufficiently studied topics is the shape of glottal waveform which is evidently associated with emotional coloring of the speech.

TEO (*Teager Energy Operator*) – on the occasions of emotional states of anger or speech under stress, fast and nonlinear air flow causes vortices located near vocal cords providing additional excitation signals beside the pitch. These additional excitation signals are present in the spectrum as harmonics and cross-harmonics. TEO [7] for signal frame is calculated as follows:

$$\Psi\left[f_s(n;m)\right] = (f_s(n;m))^2 - f_s(n+1;m)f_s(n-1;m)\,. \tag{5}$$

When applied on AM-FM sinewave it gives a squared product of instantaneous amplitude and instantaneous frequency of the signal:

$$\Psi\left[f_s(n;m)\right] = \alpha^2(n;m)\sin(\omega_i^2(n;m))\ . \tag{6}$$

In the case when the signal has a single harmonic, TEO operator gives a constant number, otherwise it is a function of discrete time *n.* Since the speech signal contains more than one harmonic in the spectrum, it is more convenient to split the bandwidth into smaller bands and then observe each one independently. The polynomial coefficients which describe the TEO autocorrelation envelope area could be used for classifying the emotional speech [8]. This method achieves 89% of accuracy when classifying neutral speech versus speech under stress. The pitch frequency affects the number of harmonics in the spectrum, so that more harmonics are present in the spectrum when the pitch frequency is low. This effect, as well as observations that the harmonics from additional excitation signals are more intense than those caused by the pitch, can be a subject of further research.

Vocal tract features – The shape of the vocal tract is changed under the influence of the emotional state of the speaker. The features which describe the shape of the vocal tract during the emotional speech production are: formants, cross-section areas of the tubes modeling the vocal tract, and coefficients derived from frequency transformations of the speech signal. Formants, as emphasized parts of the spectrum (spectral peaks), reflect locations of the vocal tract resonances. Their position in the spectrum (formant center frequency) and their bandwidth depend upon the shape and dimensions of the vocal tract, and those change depending on emotional state of the speaker. Experimental analysis [3] has shown that the emotional state largely influences the first and the second formant. MFCC (*Mel-frequency Cepstral Coefficients*) represent the signal spectrum in frequency bands which correspond to human auditory frequency response (Mel frequency scale). These coefficients are used in some emotion classifications, but better results are achieved with LFPC (*Log-frequency Power Coefficients*) which include the pitch information.

Energy of speech signal – the short-term speech energy is directly related to level of emotions in speech, therefore it can be efficiently applied in algorithms for emotion recognition. The short-term energy of the speech frame ending at *m* is:

$$E_s(m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} \left|f_s(n;m)\right|^2 \tag{7}$$

Contours of short-term acoustic features – The contours and their trends (rising and falling slopes, plateaux) also provide information useful for emotion recognition. These contours are formed when the value of some feature, calculated on the frame level, is assigned to all samples belonging to that frame. For example, the energy contour is calculated as follows:

$$e(n) = E_s(m), \qquad n = m - N_w + 1, \dots, m \tag{8}$$

Frequently used statistics for extracted features and their contours are: mean value, variance and pitch contour trends, mean and range of the intensity contour, rate of speech and transmission duration between utterances. For example, in the state of anger it is characteristic that speech possesses high energy (especially male voices) and the high pitch level (F_0). Table I shows that compared to men, and under the similar circumstances women express anger with higher speech rate. Emotional state of sadness corresponds to lower pitch frequency F_0, lower intensity in relation to neutral speech as well as falling slope in pitch contour. Men express sadness with higher speech rate as opposed to women.

TABLE I
INFLUENCE OF SEVERAL EMOTIONS ON SELECTED PROSODIC FEATURES IN RELATION TO NEUTRAL SPEECH (ADAPTED FROM [3])

	Pitch				Intensity		Duration	
	Mean	Range	Variance	Contour	Mean	Range	Speech rate	Transmission Duration
Anger	>>	>	>>		>>M >W	>	<M >W	<
Disgust	<	>M <W			<		<<M <W	
Fear	>>	>		↘	≥			<
Joy	>	>	>	↗	>	>		<
Sadness	<	<	<	↘	<	<	>M <W	>

>: greater, <: less than neutral, ↗: increasing and ↘: decreasing contour trend, m: men and w: women.

III. EMOTION CLASSIFICATION TECHNIQUES

The classification techniques can be divided into two groups:

- Techniques which use prosody contours (sequence of short-time prosody features)
- Techniques which use statistics of prosody contours (the mean, the variance, etc.)

The classifier output should evaluate, or give an assumption of emotion content of some statement, word, phrase.

The classification techniques which use prosody contours are: (1) the technique of artificial neural networks (*ANNs*), (2) the multi-channel hidden Markov model (*multi-channel HMM*), (3) the mixture of HMMs. In the technique based on ANNs, the short-time features are extracted for the frames grouped into phoneme groups. ANN is being trained on the *k*th emotional state of the *m*th phoneme group. The output node gives the likelihood of the part of some frame given specific emotional state and phoneme group. The multi-chanell HMM actually represents a group of several Markov chains which independently model the speech for certain emotional state. Transitions are possible within the single chain as well as between the chains which correspond to certain emotions. This technique achieved a correct speech recognition rate of 94.4%, and 57.6% for stress classification (this low result is attributed to insufficient training data collection) [3]. The third technique, so called the mixture of HMMs consists of two phases. The first phase uses iterative clustering algorithm in order to obtain M clusters in the feature space of training collection. Then, the second phase involves training of *C* HMMs where every HMM corresponds to single emotional state *c* = 1, 2, *C.* The correct classification rate of 4 emotional states using mixture of HMMs was 62% using energy contours in different frequency bands [3].

The classification techniques which use statistics of prosody contours can be divided into two subgroups: (1) those which use the estimation of the probability density function (pdf) of the features, and those (2) which do not use the estimation of pdf of acoustic features. The Bayes classifier belongs to the first subgroup, using different methods for $P(y|\Omega_c)$ estimation (conditional distribution of the feature vector given the emotional state Ω_c). These class pdfs are modeled as Gaussians, mixtures of Gaussians, or via Parzen windows. The other subgroup contains classifiers: (a) the k-nearest neighbors, (b) the support vector machines and (c) the artificial neural networks *(ANN).*

IV. SPEECH UNDER STRESS

Many researchers narrow down the notion of emotions and their variety to the speech under stress. In that case the task is not to recognize the particular emotion of the speaker, but to binary classify if the speech is under stress or not. The speech under stress can be a result of the outside factors (pressure at work, danger) and/or the emotional states (fear, anger, anxiety, excitement…), and it manifests itself through speech changes in relation to neutral speech (situations without stress and pressures), through the modus of speech (stuttering, tongue-slip..), the selection and the usage of certain words, sentence duration and articulation of phonemes. These speech changes are consequence of physiological changes when a person is under stress. They consist of increased respiration rate, increased subglottal pressure, increased F_0, dry mouth, changes in muscles of larynx and vibrations of vocal cords [9]. There are numerous situations and jobs which change speaker's physical and mental condition in the way it affects the implementation of ASR. These include: police officers, fire-fighters, emergency services, air traffic controllers, pilots in noisy environment, deep sea divers, astronauts, nuclear plant operators and others.

Here will be presented the results obtained by CRSS group of researchers [9] while studying changes of speech characteristics for different speaking styles and stress conditions (fast and slow speech, quiet and loud, anger, Lombard effect, speech during moderate and high computer workload tasks, neutral speech). Within the speech characteristics analysis for the speech under stress, the following were considered: fundamental frequency (pitch), intensity, duration, formant locations, spectral slope. The fundamental frequency is a good stress indicator in the wide range of stress conditions, especially angry speech, Lombard effect and loud speech. Observed characteristics for F_0 are contour, mean, variance and distribution. The mean and the variance are significantly different for neutral speaking style in relation to other styles.

In order to improve existing ASR algorithms, it is necessary to detect the period of speech under stress. Prior to stress detection, it is necessary to estimate the acoustic features from the input signal, and then detect periods of speech under stress and periods of neutral speech. There are several ways to perform the stress detection: detection-theory-based method, methods based on distance measure and methods of statistical modeling [9]. For given input feature vector *x*, two conditional probability density functions are estimated, $p(x|H_0)$ and $p(x|H_1)$, that the input signal belongs to a neutral speech and that the input signal belongs to the speech under stress, respectively. Comparing the ratio of these two densities with chosen threshold (depending on particular criteria and application), according to detection theory the decision of whether the input speech is stressed or not is made. The distance measure method reflects the distance of the input feature vector *x*, in relation to feature distribution for neutral speech, and in relation to feature distribution for speech under stress.

V. Emotion Recognition and the Transaction Communication Model

The knowledge of the variations in acoustic features during emotional speech expression, or speech under stress, may improve human-machine communication which is deteriorated or even disabled under those conditions. The computer may have problems in speech recognition due to acoustic features affected by presence of emotions and stress. Although, from the aspect of ASR, those changes do not carry a linguistic message, they do affect the accuracy of recognition. On the other hand, employing the knowledge about variations in acoustic and prosodic features into the TTS module, it would achieve better quality and more natural synthetic speech.

For the purpose of better understanding, the machine model in human-machine speech interaction [2] is envisioned as a combination of information from two sources: (1) by processing the speaker picture, and (2) by processing his/her speech, as shown in Fig. 1.

Audiovisual nature of speech perception in the human-machine dialogue is reflected in two phenomena. First, watching the speaker, especially speaker's face and lip movement which are synchronized with articulated speech, enhances and facilitates the perception of speech, especially in the noisy environment or in situations of lower speech intelligibility.

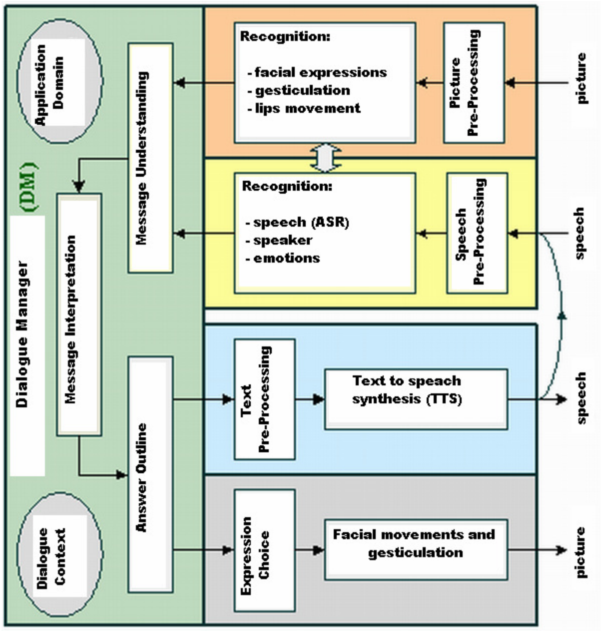

Fig. 1. Functional model of the machine in the verbal dialogue. (taken from [1]).

The other example is the McGurk effect, which denotes changes in auditory speech perception in case when the lip movement does not match spoken words [10]. Besides its verbal expression, emotions have their form of characteristic facial expressions. For example, happiness is followed by a smile on face. Within the framework of multimodal human-machine communication, the machine listens with the assistance of ASR module. The implementation of ESR algorithm (emotion recognition) into ASR module will increase the reliability of speech recognition and ensure natural and spontaneous human-machine dialogue. The module which processes the picture of human face (lips movement, facial expressions and gesticulations) provides the information which are video correlates of spoken phonemes, words and expressed emotions as well. The synergy of these two modules complements the semantics of spoken words, while verbal interaction is complemented with paralinguistic and nonverbal elements of dialogue, as shown in Fig. 1.

Based on extensive prior knowledge about application and language, the machine performs postprocessing with purpose of spoken language understanding (SLU), that is, the integration of speech recognition and natural language understanding (NLU). This would also be valid in the opposite direction, from machine to the human, if the TTS module would be able to synthesize emotionally "colored" speech, and along with the adequate animation of the speaker's face in output picture, it would lead to a specific illusion of having a conversation with other "person", i.e. having a natural communication.

Transaction communication demands from its participants to recognize the influence of one message on the other and broaden the field of mutual experience through active process of language understanding [2]. For better understanding of meaning and intentions of the utterance, it is necessary to consider not only its verbal elements but also its nonverbal elements. In that sense, the automatic emotional speech recognition could have an important role because the emotional state of the speaker points out his/her needs as the conversational partner and his/her reactions towards current dialogue, and so helps proper dialoge handling or accommodates the initialized communication.

Here is the example of human-machine dialogue where the machine recognizing the key words in the user's answers as well as the level of his satisfaction (based on recognition and emotion classification), directs the dialogue and gives answers in order to successfully complete communication. This example is shown in Table II. The machine has understood the need of the person after the first sentence – the ticket reservation to city of Niš, and then directed the question to obtain more data (the date and the time of the trip). While answering, the person has shown some degree of uncertainty and confusion, and then has corrected himself/herself. The machine "understood" this using ESR algorithm, accepted different timing as requested, and then provided the answer. The machine has correctly connected the time specification "day after tomorrow" to a concrete date in order to avoid misunderstanding. The next user's answer, "yes, yes, yes" was detected by using emotion detection and classification as a delight and satisfaction of the person. Comfort and natural dialogue was achieved with additional pronunciation instructions to the machine. The effect of emotionally "colored" speech was achieved by varying prosodic features within the synthesized speech.

TABLE II
THE EXAMPLE OF THE DIALOGUE BETWEEN THE MACHINE (M) AND THE HUMAN (H)

M 1	Hallo, **good day!** [a bit faster] This is service of inter-city bus station Novi Sad. [short pause, slower] **How can I help you?**
H 1	*I would like to make a reservation to city of Niš.*
M 2	[clear and slowly] Please, give me the **date** and the **time** of your travel.
H 2	*Day after tomorrow around 6 o'clock, er, 16 o'clock.* // recognized confusion and the subsequent correction by the user
M 3	[clear] For March 10th there is a departure at 16:15, Nišekspres. [polite and inquiring] **Does it** suit you?
H 3	*Yes, yes, yes….* // recognized user's satisfaction
M 4	[] All **right.** Your reservation has been confirmed, to city of Niš, [slow], **March 10th** at **16:15.** [inquiring] Can I help you with anything else?
H 4	*No*, thanks.
M 5	[joyfully] Have a nice day and **bon voyage**!

Bold text should be emphasized. Text in the brackets gives instructions for speaking style and intonation. Words in *italics* are the keywords for dialogue flow.

With the correct word recognition and their semantics, the recognition of user's emotional state allows correct understanding of the user's message and needs, confirming the success of the dialogue and achieving the goal of the conversation.

VI. Conclusion

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

This paper deals with several topics which are related to automatic emotional speech recognition. Firstly, the extraction of relevant acoustic features within speech has been presented as well as space of their variations oposed to neutral speech. In order to do such task it is necessary to have formed speech databases whose diversity and size are still unsatisfactory, especially for Serbian language.

The review of some widespread classification techniques is given, but the problem which remains is that their accuracy cannot be directly compared because the results were obtained using different databases and experiment protocols. It is necessary to detect the speech under stress, as a way of expressive speech, in order to improve existing ASR algorithms. One of the general approaches is the equalization of stress, that is, normalization of parameters variability due to presence of the stress within the speech signal.

The importance and application of emotion recognition has been emphasized within the human-computer communication. It contributes to higher accuracy of recognition within the ASR module, and within the human-machine dialogue it helps better understanding of the meaning of the message, and also needs and intentions of individuals. For TTS module knowledge of characteristic prosodic variations will give the possibility that machine speech can be as close to the human speech.

One of the future research directions refers to ESR algorithm implementation within the ASR module as well as their integration with the module for processing face pictures. This would be a step forward in aspirations to model machine as an audiovisual "conversational partner" within the multimodal human-machine communication.

References

[1] L. Bosch, Emotions, speech and the ASR framework, Speech Communication 40, pp. 213-225, 2003.

[2] V. Delić, M. Sečujski, "Transaction model of human–machine verbal interaction", DOGS, Kelebija, 2-3.10.2008, pp. 8-15.

[3] D. Ververidis, C. Kotropoulos, I. Pitas, Automatic emotional speech classification, In. Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 593-596, Montreal, 2004.

[4] S. T. Jovičić, Z. Kašić, M. Đorđević M. Vojnović, M. Rajković, J. Savković, Forming corpus of emotion and attitude expression in Serbian language-GEES, XI Telekomunikacioni forum TELFOR 2003, Beograd, 25-27.11.2003.

[5] Y. Li, Y.Zhao, Recognizing emotions in speech using short-term and long-term features, In: Proc. ICSLP 1998, pp. 2255-2258, 1998.

[6] M. M. Sondhi, New methods of pitch extraction, IEEE Trans. Audio and Electroacoustics 16, pp. 262-266, 1968.

[7] H. M. Teager, S. M. Teager, Evidence for nonlinear sound production mechanisms in the vocal tract, NATO Advanced Study Institute, Series D, vol. 15, Boston, MA: Kluwer, 1990.

[8] G. Zhou, J. H. L. Hansen, J. F. Kaiser, Nonlinear feature based classification of speech under stress, IEEE Trans. Speech and Audio Processing 9 (3), 201-216, 2001.

[9] C. Műler (editor), Speaker Classification I, LNAI 4343, pp. 108-137, Springer-Verlag Berlin Heidelberg, 2007.

[10] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264, pp. 746-748, 1976.