Introduction

Expressing emotions is an essential part of interpersonal communication. Emotions are the carrier of information regarding feelings of an individual and one’s expected feedback. Although computers are now a part of human life, the relation between a human and a machine is not natural. Customer services such as call centers are increasingly using automated recognition support. Knowledge of the emotional state of the customer allows the operator or the machine to better adapt and generally improves cooperation.

Emotion recognition methods utilize various input types i.e. facial expressions [1], speech, gesture and body language [2], physical signals such as electrocardiogram (ECG), electromyography (EMG), electrodermal activity, skin temperature, galvanic resistance, blood volume pulse (BVP), and respiration [3]. Speech is one of the most accessible form the above mentioned signals and because of this most emotion recognition related research focuses on human voice, which has become a relevant trend in modern studies. However, satisfactory efficiency has not been achieved yet.

Most studies are based mainly on samples of acted speech, which is characterized by single specific emotion type. However, this style of consistent prototypical emotion expression does not exist outside of a laboratory environment [4]. Spontaneous speech is more complex, can consists of several emotions or their mixture and proper recognition is more challenging even for human deciders. Moreover, emotion expression depends on context, personality, culture and even mood of the speaker. Emotion classification algorithms created for real-world application should be able to recognize emotional content of an utterance or dialog beyond these and other limitations. Therefore, while designing such a system one should take into consideration the complexity of emotions in spontaneous speech.

This paper presents an analysis of classification techniques that can be utilized to interpret non-prototypical emotion utterances. These methods include modelling of: emotional sequence (changes over time), conflict of emotions (a few emotional state can be expressed the same way like), mixtures of emotions (several emotional states occur at the same time) and similarity of emotional states. Basing on this research the authors recommend to classify natural emotions based on emotional profiles rather than using categorical hard labels.

This paper is organized as follows. The second section presents emotions taxonomy basing on Plutchik’s theory, definition of emotional profiles and difference between prototypical and non-prototypical emotional states. Section 3 presents the database that was created for the purpose of this research. In section 4 extracted features are briefly described. Section 5 presents classification algorithm. In section 6 the process of emotional profiles creation is described and in section 7 results are shown. The experiments are commented in section 8.
Emotion taxonomy

This paper refers to the theory of Robert Plutchik, according to his theory (Fig. 1) [5] there are 8 primary bipolar emotions: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation. These emotions are biologically primitive and have evolved in order to increase the reproductive fitness of an animal. Moreover, primary emotions can be expressed at different intensities and can mix with one another to form different emotional states. This translates to perception of natural emotions, which is a complex and subjective process. Recognizing several different emotional states in a given situation is very common.

Fig. 1. Plutchik’s wheel of emotions
Source: https://pl.wikipedia.org/

Natural perception of emotion is complex, subjective and very often can be assessed differently. An example of woman who learns that her father remains in jail is presented in [6]. A panel of experts evaluated the sample as: anger, disappointment in, sadness and despair. It illustrates the complexity of the human psyche and points to lack of clarity in showing and perceiving natural emotions. There is a constant search for solutions which can avoid determined labeling (i.e. creating so-called emotional profiles), but point out the presence or absence of specific basic conditions and measure the intensity of each [7]. This approach may be useful in identifying ambiguously specified emotions which appear in spontaneous dialogues.

Prototype (basic) emotions are adaptive states designed for the fight for survival of an individual. Plutchik distinguished eight basic states: anger, anticipation, joy, trust, fear, surprise, sadness and disgust. In addition to prototype states one can distinguish a number of secondary emotions, which are combinations or mixtures of the basic emotions.

In 2009, at ACII (Affective Computing and Intelligent Interaction) conference a special session dedicated was held, dedicated to recognizing emotions from ambiguous expressions. This conference started an avalanche of questions about spontaneous speech studies, and researchers began to abandon analysis of acted out recordings.

Main characteristics constituting the ambiguity of the mental state of the speaker and the audience perception:
- mixture of emotions - two or more emotional states occur at the same time; according to Plutchik’s theory, one can experience a theory basic emotions; this involves combining various primary states in one secondary state (emotions lying next to each other on the Plutchik’s wheel);
• masking emotions - hiding emotional state one is currently experiencing with a different state (e.g. masking sadness with joy);
• conflict of emotion expression - a specific state can be achieved as a result completely opposite emotions (i.e. crying out of happiness vs crying from sadness);
• similarity of emotional states - boundaries between individual states are ambiguously defined, overlapping each other (emotions lying next to each other on the Plutchik’s wheel);
• sequence of emotions - consecutively display of various states during speech.

Emotional profiles are helpful in determining most probable mental state of the speaker, examining the evolution of this state during speech, and interpretation of expressions of complex or several emotions.

**Emotional Database**

The task of pattern recognition requires gathering of relevant input data (sets for training and tests). For this study the input consists of emotional speech samples. Considering the theory of Robert Plutchik, a set of basic emotions was created, but in a wider scope in comparison to other studies. The first step was to collect speech samples in seven core emotional states: anger, anticipation, joy, fear, surprise, sadness and disgust. Confidence was deliberately excluded from the set, due to insufficient amount of samples. All recordings were evaluated and labeled by a group of experts.

Speech samples can be divided into three categories, taking into account their source: spontaneous, invoked and acted or simulated emotions. First type of samples can be obtained by recording speakers in natural situations, or using TV programs such as talk shows, reality shows or various types of live coverage. This type of material might not always be of satisfactory quality (background noise, artefacts, overlapping voices, etc.) and may obscure the exact nature of recorded emotions. Moreover collections of spontaneous speech must be evaluated by human decision makers to determine the gathered emotional states.

Another method of sample acquisition is provoking an emotional reaction using drugs or staged situations. Appropriate states are induced using imaging methods (videos, images), stories or computer games. This type of recording is preferred by psychologists, although the method cannot provide desirable effects as reaction to the same stimuli may differ. Similarly to spontaneous speech recordings, triggered emotional samples should be subjected to a process of identification by independent listeners.

Third source of emotional speech are acted out samples. Speakers can be both actors as well as unqualified volunteers. This type of material is usually composed of high quality recordings, with clear undistorted emotion expression. Moreover the easiness in acquiring recordings opens a possibility of obtaining several utterances, representing different emotional states, from a single user. However the acoustic characteristics of such an utterance may be exaggerated, while more subtle features, completely ignored.

Selection of a representative recordings is one of the key elements affecting the research credibility. It is assumed that a sample is representative, when all the values which could affect the test results are present. Because the process of emotion expression is subjective, depending primarily on the gender, age and place of residence of the speaker. While selecting speech sources the above mentioned information was one of the guidelines, in order to retain the right proportions of these variables. However this assumption is largely limited due to lack of personal data of the speakers in recordings obtained from radio auditions.

The emotional state of the speaker can be identified basing on short utterances such as Yes or No [8]. Thus short sentences, or even single words are suitable for emotional analysis. Occasionally additional sounds such as screaming, squealing, laughing or crying carry the information about the speaker's emotional state. Therefore, in addition to full spoken words, such sounds which occur in everyday communication, were featured in the created corpora.

The database consists of more than 80 speech samples for each emotional state (presented in Table 1.). Additionally, differentiation of samples (gender and age) can positively affect classification results. By exploiting different types of utterance, influence of personal features on classification is limited. What is
more, considering the difference between male and female emotion expression, we used additional module of gender recognition.

All recordings were divided into seven groups (basic emotions). This process was conducted by the authors and psychology students from the University of Lodz. The division was performed with the use of video material which allowed access not only to voice and semantics but also the visual display of emotions, such as gestures or facial expressions. In the following step of the process, volunteers labeled the samples basing solely on audio input. The volunteers group consisted of 15 normal hearing people, both male and female, aged 21 to 58 years. The task was to assess the recordings and classify them into the groups of basic emotions. All listeners were presented a set random samples that consisted of at least half of each prequalified basic emotions recordings. The evaluators listened to audio samples one by one, each assessment was recorded in the database. Every sample could have been played any number of times before the final decision, but after the classification, it was not possible to return to the recording. Average recognition amounted to 82.66 % in the range of 63 to 93 %.

Table 1. Structure of the natural corpus - primary emotions

<table>
<thead>
<tr>
<th>Emotion name</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>111</td>
</tr>
<tr>
<td>Joy</td>
<td>165</td>
</tr>
<tr>
<td>Fear</td>
<td>128</td>
</tr>
<tr>
<td>Surprise</td>
<td>87</td>
</tr>
<tr>
<td>Anticipation</td>
<td>88</td>
</tr>
<tr>
<td>Sadness</td>
<td>115</td>
</tr>
<tr>
<td>Disgust</td>
<td>90</td>
</tr>
<tr>
<td>Neutral</td>
<td>235</td>
</tr>
</tbody>
</table>

Under Polish law, recording of live TV programs or downloading such recordings is permitted without the consent of the author. These recordings may be used only to the extent permitted private use. Any application which is commercial in nature or leading to the provision from the downloaded work is prohibited. Thus, the samples collected can be used in research only, but making them publicly available is not possible at the moment. The authors have taken steps to obtain permission from TV stations. Granting public access to such a database would constitute a basis for development of research on spontaneous emotions. Moreover, we are open for cooperation and we can provide models of emotion in the form of features vectors for research.

Features extraction

Representation of the signal in time or frequency domain is a complex image. Therefore, features are sought to determine signal properties. In this part extracted features will be presented.

Fundamental frequency is the frequency of vocal folds. It is responsible for the scale of the human voice and accent. It plays an important role in the intonation, which has a significant impact on the nature of the speech. F0 changes during articulation. The rate of those changes depends on the speakers intended intonation [9]. There are many methods to determine the fundamental frequency. In this paper F0 have been extracted using autocorrelation method. The analysis window was set to 20 ms with 50 % overlap. It is difficult to objectively assess the behavior of F0 based on the chart. Therefore statistical parameters related to F0 have been extracted.
Formant frequencies are the frequencies, at which local maxima of the speech signal spectrum envelope occur. They are the properties of vocal tract. Basing on them, it is possible to determine who the speaker is and about what and how he is speaking [10]. In practical applications from 3 to 5 formants are used. In this paper 3 formant frequencies were estimated.

Speech signal energy, which refers to the volume or intensity of speech, also provides information that can be used to distinguish emotions i.e. joy and anger have increased energy levels in comparison to other emotional states.

Perceptual approach is based on frequency conversion, corresponding to subjective reception of the human hearing system. For this purpose, the perceptual scales such as Mel or Bark are used. In this paper Mel Frequency Cepstral Coefficients MFCC [11], Human Factor Cepstral Coefficients HFCC [12], Bark Frequency Cepstral Coefficients BFCC [13], Perceptual Linear Prediction PLP [14], RASTA-PLP [15] and Revised Perceptual Linear Prediction RPLP [16] coefficients were taken into consideration. The entire scheme for perceptual feature extraction is shown in Fig. 2.

![Perceptual feature extraction process for different methods](image)

Source: own work

**Classification algorithm**

The structure of emotion recognition algorithm is presented in Figure 3. The algorithm consists of two levels. Feature vector describing an unknown object and all objects from training set was divided into sub-vectors represented by groups of features. The fragmentation creates $m$ separate features vectors, for example F0 or MFCC subvectors. The first level consists of $M$ classifiers where $m$ is the number of group of features. Each classifier produces an output (particular emotional label) using k-NN algorithm. One the second level all $m$ decisions are subjected to voting. Using a weighted or equal vote the dominant class is selected.

Considering the volatility of emotional states during the speech, segmentation of speech is the next step of the algorithm. It was assumed that the base fragment should be 1s long - at such a short time the state should not change. An example of segmentation of speech samples to one second pieces illustrated in Figure 4.
A similar approach was presented in previous studies [17], where an utterance was divided into three fragments of equal length: beginning, middle and end of the speech. Subsequently, the parts were subjected to separate classification and the result was determined by voting. Due to the nature of samples in the corpora (short speeches, single words or sounds only), recordings such as laughter, scream or screech were not used in that research. In this work, we applied division into 1-second fragments. If the statement is less than 1s, it remains undivided. In a case where the end portion is longer than 500ms, the whole sample is divided into parts of 1 second and a completive fragment. If the final fragment is shorter than 500ms, it is added to the last fragment of 1 second. Each of the fragments is subjected to individual classified using k-NN algorithm. Each fragment resulting from the applied division is subjected to separate classification. The number of classifiers is equal to the groups of parameters used in describing the speech. The last step of the classification is voting basing on the results of classified parameter groups.

**Emotional profiles**

According to [18] emotional profile is a vector describing the presence or absence of a basic emotion labels in particular utterance. Thus one can avoid hard-labeled assignment and instead provide multiple
labeling. This kind of classification indicates the contents of several emotional states in a given speech sample. It is possible to use fuzzy-labelling using membership function for each emotional state.

Classification algorithm consists of a series of classifiers basing on the distribution of features groups. Extending this module by another – segmentation speech module – creates a hierarchical structure of k-NN classifiers. Final voting is based on the outputs from all of the classifiers. Replacement of voting by the analysis of the responses allows creation of emotional profiles of speech samples. Counting particular outputs one can built a histogram representing the content of emotional state in sample. This approach does not replace the classification process, however can be used as a support module in ambiguous samples classification.

**Emotional profiles – results**

For selected speech samples, labeled by volunteers as ambiguous, emotional profiles were determined. Every sample was subjected to classification. Then, basing on the outputs of separate classifiers, corresponding histograms were created. Examples of four different models are presented below.

![Fig. 5. Emotional sequence – consecutively display of various states during speech. Both, by experts and by volunteers, the sample was determined initially as anger, changing gradually into sadness. On segmentation level this sample was divided into three parts, which were classified as follow: first and second part as anger, the third part as sadness. Source: own work](image)

![Fig. 6. Mixture of emotion – two or more emotional states occur at the same time. Both, by experts as and by volunteers sample can be classified as a mixture of anger and sadness. Results obtained by machine confirmed human evaluation. Source: own work](image)
Fig. 7. Conflict of emotions – a specific state can be achieved as a result completely opposite emotions. The sample was determined by experts as joy (crying with happiness). In blue – subsequent results obtained by machine. In red – decisions of ten volunteers.

Source: own work

Fig. 8. Similarity of emotions – boundaries between individual states are ambiguously defined, overlapping each other. Both, by experts and by volunteers sample may be classified as anger or as anticipation. Machine confirms human decision.

Source: own work

Conclusions

The main assumption of these research was to develop a system allowing automatic emotional state recognition based on natural, ambiguous speech. Emotion expression is a complex process involving various dependencies such as: speaker individual features, context, mood, personality, and culture. However, emotion classification system for real-world must be able to interpret the emotional content of an utterance or dialog beyond this dependencies. Prototypical emotions often do not exist outside of a laboratory environment. This paper presents an algorithm for interpreting the emotional content of ambiguous utterances.

To achieve this we created a polish spontaneous emotions database consisting of over 700 samples divided into seven sets representing primary emotional states. Quantitative description of the problem is based on, commonly used in this type of research, speech descriptors compared with hybrid perceptual coefficients never before applied in this area. The process of classification was multimodal and based on emotional profiles creation.

An unchallenged advantage of spontaneous speech database are samples, containing ambiguous emotional states. Conducted analysis shows complexity of conception, perception and expression of emotions. In natural conditions a speaker can be filled with various emotions simultaneously, and a listener can perceive signals differently. The research shows that suggested algorithm successfully deals with ambiguous samples, what indicates that some parameters of voice cannot be changed, even on purpose.
The proposed studies are implementable particularly for business purposes. One potential application is the detection of emotional state of speaker in call center services (recorded messages, a conversation between the operator and the customer). This kind of system can provide a feedback to an operator or a supervisor for monitoring purposes. For example one can determine the level of frustration or anxiety of the speaker. On the basis of such information one can sort voice mail messages according to the emotions expressed or allocate appropriate operator to the caller.

Abstract

Emotion recognition system can improve customer service especially in the case of call centers. Knowledge of the emotional state of the speaker would allow the operator to adapt better and generally improve cooperation. Research in emotion recognition focuses primarily on speech analysis. Emotion classification algorithms designed for real-world application must be able to interpret the emotional content of an utterance or dialog beyond various limitation i.e. speaker, context, personality or culture. This paper presents research on emotion recognition system of spontaneous voice stream based on a multimodal classifier. Experiments were carried out basing on natural speech characterized by seven emotional states. The process of multimodal classification was based on Plutchik’s theory of emotion and emotional profiles.

Keywords: emotion recognition, speech signal analysis, man-machine interaction

References


