SOUND FEATURES USED IN EMOTION CLASSIFICATION

Vesna Kirandziska, Nevena Ackovska
Institute of Intelligent Systems, Faculty of Computer Science and Engineering
—SsCyril and Methodius" University in Skopje, Macedonia

ABSTRACT

Human emotion perception is an open problem that can be used in human-to-computer communication. In this paper classification of human's emotions will be done using the human voice and its features as input data. More precisely, different sound features used for human voice emotion classification will be explored and ranked. An application which uses human sound features to classify two categories of human emotions: positive and negative emotions will be presented. The results will be given and explained. A comparison between the classifier built and the human's natural classifier will also be presented.

Keywords — sound feature extraction, emotion detection, classification, machine learning, artificial intelligence, psychology of human voice

I. INTRODUCTION

Humans are emotional creatures. Great part of human-to-human communication is emotionally conducted. Emotions in humans are usually detected through facial expression and other visual movements of humans. Likewise, information for human emotional state can be extracted through MRI and other electrical data from human brain activities. It is expected that this is the most precise method for explaining human emotions. However, this technique has a down side in the actual data extracting, which is mainly invasive.

Alternatively, humans communicate not only face to face, but they communicate voice to voice as well. This means that besides human movement and facial expression, humans use their voice to show their inner emotions. Indeed, speech has features that are indicators of the human emotions [2]. The ease of recording the human voice and the ease of sound extraction from distant communication makes human voice the most practical source of data for emotion classification.

This is the main reason why emotion classification from speech has become one of the major topics of interest in two different research fields: human-to-computer communication and speech processing.

There are varieties of applications that benefit from emotion classification. Some applications are: video and computer games (getting input data from human emotions), call centers (recognizing valuable callers), human intercommunication (improving human-to-human interaction) and human-to-robot interaction (robots that understand human's emotions).

Many examples of the work done in this research field encourage us to do something more in order to improve real time sound emotion classification with better accuracy and speed. Exploring different sound features is important for the continuation of this research work.

In the next chapter of this paper the problem of emotion classification will be explained. Next, meaningful sound emotion features for classifying emotions according to psychological studies will be presented. Afterwards, the course of our research conducted for classifying positive and negative emotions in humans will be described. Later, the data and the results will be explained. The possible future directions and opportunities for further research will also be stated.

II. EMOTION CLASSIFICATION FROM HUMAN VOICE

Emotion sound classification is a process of classifying human speech in two or more emotional states. The conventional method of estimation of emotion in speech has 3 steps. First is the step of human speech collection. Next, sound features are extracted from human speech. Finally, a classifier is created from these sound features, using a learning algorithm [3] (Fig 1.).

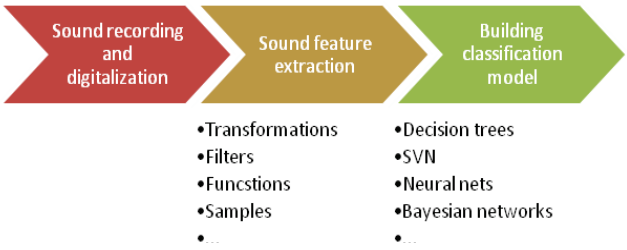

Figure 1: The process of emotion sound classification.

As shown in Fig.1 the first phase is the process of sound recording and sound digitalization. Sound data can be collected with three different approaches [4]:

a) Actors are recorded saying the same words with different emotions

b) Actors are provoked to being in specific emotion and then recorded (—Wizard-Of-Oz" method)

c) Real life humans are recorded and then the emotional state is recognized by other humans.

As stated in [1] the databases taken with the third approach are the most difficult for emotion classification. Example databases used in similar researches are: Linguistic Data Consortium from the University of Pennsylvania [10], Berlin database of emotional speech [11] and the Danish Emotional Speech corpus [12].

We will use a new database gathered with the first approach from only one actor. Our goal is to analyze the best features extracted from the sound, which is invariant to the object that is recorded. Also, we are going to assume that sound is invariant to the word semantic, so the words spoken will be ignored. One reason for this is due to the different languages, cultural differences and diversity. Still, emotions are present in the human voice. As stated latter, sounds in our database are classified in two classes: positive and negative emotion.

In this paper special attention will be given to the second classification phase. Mathematically, sound features can be multidimensional or one-dimensional functions of the input sound. In order to extract features, sound is first segmented. The size of these segments is determent by a parameter called a window size. Two neighbouring segments can overlap so that the same value information is taken in account twice. Some information can be extracted from the sound represented in time domain, but transformations like the Fast Fourier transformation change the vector space to a frequency space. The vector space could be time space, frequency space, correlation space etc. Many different filters like the frequency passes, filtering the extremes, the Bark-filter, the ERB filter [8] are used to translate the vectors in the same vector space.

The final phase is the classification itself. Lots of famous algorithms for classification can be used. Classification methods like classification trees, Neuron networks, Support vector Machines, Hidden Markov models, Bayesian networks etc. are most commonly found in the literature on emotional speech classification [1][4]. All these classifiers need training data from which they learn the parameters. In fact, the database has testing and training data. After learning the parameters, testing data is used to show the precision of the classification method. As an example, Lee [13] with 75% accuracy distinguished between two emotions: negative and positive in a call centre environment. Similarly, Paeschke [14] gained 77% accuracy of classifying two emotions: agitation and calm.

Our research will use Neural Networks. This algorithm was used in many similar researches and has proven to be significantly accurate. For example, in [7] classifying hot anger and natural emotions was done with accuracy of 90.91%, while happiness and sadness were truly classified with 80.46%.

III. SOUND FEATURES USED FOR EMOTION DETECTION

Sound carries information about the emotional state of one human. This information is described by many sound features. Firstly, these features could be global or local functions of the recorded sound. Global functions are calculated by some statistical measure (mean, standard deviation, minimum, maximum, percentile, slope, etc.) given local values. Local functions are calculated from the original data or from some transformation. Data given by global functions could be used for clarifiers like Neural Networks. Contrary, data given by local functions could be classified with Hidden Markov Models

Today researchers in this area are not yet agreed on which sound features are the most important in the emotion classification problem. Lots of features have been used in different studies. According to some of them, these can be divided into different categories (Fig. 2).

Figure 2: Sound features categories.

The number of sound characteristics for one recorded sound can vary. It has to be large enough so all valuable information is included, but small enough so that there is no redundancy and correlation between the features.

Frequency features are mostly used for emotion classification. They include pitch-based measures which are related to voiced speech generation mechanism and vocal tract formation. Pitch is the perpetual correlate of fundamental frequency of voice. This feature is used in almost every survey on emotion classification. Jitter [6] and Mel Frequency Cepstral Coefficients (MFCC) [1] are also features that are usually found in many researches for emotion sound classification. Features in the second group are related to speech production processes. Statistical measures like mean, minimum, maximum of the energy can be used, as well as characteristic feature like shimmer [6]. The third group contains features which are related to behavioural speech production processes. Examples are voiced duration, utterance duration, segment duration, pause duration, their ratio, mean, minimum, maximum, standard deviation etc. Segmental features include hyper-clear speech, pauses inside words, syllable lengthening, off-talk, inspiration, expiration, mouth noise, laughter, crying, and unintelligible voice [1].

Emotion classification is open for new sound features which would improve the classification models. Combination of more emotional characteristics is also one possible option for improving the feature space. For the time being, no feature set has proven significantly better than the others.

IV. Emotion state categories

To start a study for emotion sound classification we need to be sure that there is relevant information about the human emotional state in their speech. According to some analysis in psychology, information about which sound features are most relevant in specific emotion perception has been extracted. For example, in [2] emotions were split in 10 different types of emotions: Pleasant, Activity, Potency, Anger, Boredom, Disgust, Fear, Happiness, Sadness and Surprise. For each, a set of sound characteristics has been given. The results are drawn from an experiment and they reflect the human's perception of other human's emotions.

The sound characteristics showed in [2] involve some of the features explained earlier in the previous section. From [2] we extracted a model of the main sound features for positive and negative emotions. The result is shown in Table 1. This categorization, with just two non-overlapping categories or classes, promises good results.

The information presented in Table 1 is the starting point of our research. We would like to sort these features according to the best classification results. The result of the research would be a model of positive and negative emotions according to the tempo (quality feature), amplitude (energy feature) and pitch (frequency feature). Mathematically, all these measures are functions of small segments of the recorded sound. In order to get proper data, global statistical values like mean and standard deviation are calculated from these measures.

Table 1: Characteristic important sound emotion features.

Emotion type	Tempo	Pitch	Amplitude
Positive	Fast tempo	Pitch contour down Large pitch variation	Law mean amplitude
Negative	Slow tempo	Pitch contour up High pitch Small pitch variation	High mean amplitude

Other than building the model and showing the classification results from different features, one could also compare the classification model - generated with the one described in Table 2 - obtained by a psychological research.

V. Building the classifier

The phases of building the emotion sound classifier are shown on Fig. 4. In the following sections these phases will be described in detail. Despite the accuracy of the emotion classifier, a great importance will be given to rating scores for the sound features in emotion classification problems. Also the classifier will be compared with the results from the psychological studies presented in Table 1.

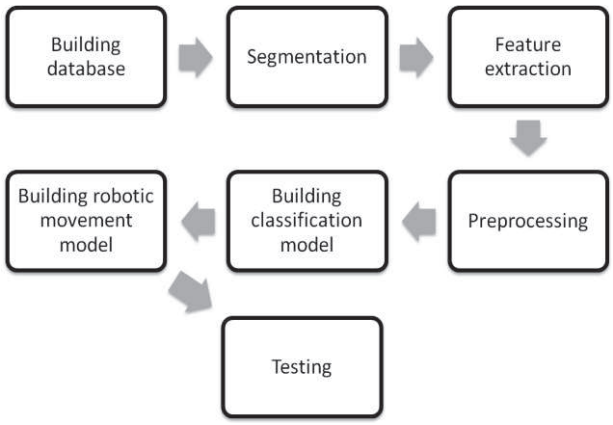

Figure 4: The flow of the classifier building process.

A. Sound database

Our database was created from the voice of just one actor. He was recorded saying „feel good" 20 times, so that in each of these recordings he acted some emotion. These emotions were previously well chosen and divided in two emotional states: positive and negative.

The recorded stereo audio files were saved in WAV format. The sample rate of the recordings is 44100Hz and for each sample, 16 bits were saved. The time length of each sound recording was about 2-4 seconds. For each file, approximately 300000 samples were taken.

In the segmentation phase (Fig. 4) the sound signals were segmented so that each segment has information for about 25ms, while the time included in the overlapping segments is 10ms. Sound signals were segmented in order to get the sound features as global statistics form segment's features.

Taking into account the results of the psychological studies presented in Table 1, the sound features: pitch's mean, pitch's standard deviation, tempo, amplitude's mean and amplitude's maximum were extracted from the recordings in the database. The local functions: pitch and amplitude, were calculated [8] for each segment. Mean value, standard deviation and maximum value ware calculated as global functions from the array of local function's values. The phase of feature extraction is the most difficult for implementation, because it is still an open topic in the field of signal processing.

B. Feature extraction

Here, the algorithms for extracting pitch, amplitude and tempo for sound segments are described. Many algorithms for calculating pitch are presented in [15], although an algorithm with satisfactory accuracy is not found. The pitch is the lowest frequency component, or partial which relates well to most of the other partials. A weighted autocorrelation function [16] for noisy speech was used in the algorithm implemented in our application. The algorithm uses the concept of a correlation function. A correlation function of two data series has values in the range [-1; 1], where 1 means that the series are linearly correlated, while -1 means they are negatively linearly correlated. Data series that are completely uncorrelated have a value around zero. In order to find the pitch we are looking for greater correlations in sound data series that are on the same distance apart. This is called a log number and this is the parameter of the autocorrelation function. The domain of the autocorrelation function matches the human voice's pitch period domain (0.0025s – 0.02s). The pitch was taken as the log number that has a maximal value for the weighted autocorrelation function. This value is the pitch period and accordingly the pitch is its reciprocal value.

One energy feature (Fig. 2) which we considered in our study is the amplitude. The amplitude in each segment is calculated as a maximum from the absolute values of the samples in each segment. The bytes recorded for each sample were interpreted as double values in the range from -1 to 1.

The tempo is a quality feature. It is calculated by dividing the estimation on the number of samples that come from unvoiced sound with the number of all samples. In order to estimate witch samples come from human voice and witch are noise only, a correlation function between sound and noise samples was used. Segments with correlation greater than a given threshold were classified as unvoiced. If otherwise, the segment was considered to be voiced. After a database with the extracted sound features was built, in the preprocessing phase the data was scaled.

C. Building the classification model

Neural networks are used as a classification method in our research. Neural networks were used in many researches in sound classification and showed good results. We used a simple version of a neural network that has one neuron with more weighted input data. This model of a neuron network is called the McCulloch and Pitts model (MCP). The input is presented by five extracted sound features. The output is binary and expresses positive, or alternatively negative, emotional state. The neuron was trained to give (almost always) the correct output for given sound features with some predefined allowed error.

The weights play crucial role in training the neuron. Each input has a specific weight, so the effect that each input has at the output depends on the weight of the particular input. The weight of an input is a number which when multiplied with the input gives the weighted input. The MCP neuron works in a way that if the sum of the weighted inputs exceeds a pre-set threshold value, the output would match a positive emotional state. The MCP neuron has the ability to adapt to a particular situation by changing its weights. The Delta rule was used for adapting a MCP neuron. This rule changes the weights by a learning step if the neuron doesn't give the correct error. The result of the training of the MCP neuron is a vector of weights that corresponds to each feature.

D. Testing and results

In the training phase of the MCP neuron, we experimented with different threshold values, predefined errors and initial weights. The best results (Table 2) were shown when the initial weights were set to a random value and the threshold value had a fixed value of zero. The trained neuron has accuracy of 85%. The training and testing phase were done using the leave-one-out cross validation algorithm.

Table 2: Weights for the trained MCP neuron model.

Features	Weights
Mean Amplitude	0.376685
Max Amplitude	-0.22787
Mean Pitch	0.63438
Pitch Std	0.165609
Tempo	0.251441

From the weights, it is clear that negative emotions are closely related to low mean amplitude, high max amplitude and low mean pitch. If we compare the results from the psychological studies (Table 1) and the weights in Table 2 we can conclude that there is some correlation between the human's natural emotion classifier and the classifier build here. Higher values of the max amplitude and low tempo values for negative emotions are results gained from the neuron and the psychological studies also. But, for the pitch mean and the amplitude mean the results are contrast. We should discover why this is the case and then to try to use that information in the future work. Nevertheless, this gives another view of the process of building an emotional sound classifier in which the algorithm is not only data, but, human driven.

Using the weights we can rank the features by importance. The values close to 1 by absolute value are more important because they have bigger affect in the classification of the negative and positive emotional state, accordingly. Using this knowledge we could rank the features with the given order starting from the one that is most important: 1. Pitch Mean, 2.

Amplitude Mean, 3. Tempo, 4. Amplitude Max and 5. Pitch Std.

VI. Future work

In the future, other features should be in detail explored and ranked. Each new important feature should be ranked and given a proper weight in the classification model. This should improve the accuracy of the model. Also, comparison of the classification model to the human emotion perception should be made just to be sure that there is a correlation.

Emotions differentiation in more categories would be essential for getting more precise emotion perception of the human speech. In this study only two disjoint emotional states were considered. In real life, humans are in many overlapping emotional states at the same time. This makes the problem of emotion perception and classification more challenging. Proper differentiation is very important, so people with psychological background knowledge should be included in feature research.

Next, more data should be included in the researches. Bigger databases should be made and more real life sounds should be used to improve the validity of the results. This is in direction of using voice from more than one human - not all actors.

In the future, some usable applications could be made based on the emotion sound classifier. One possible application is in the field of human-to-robot interaction, where robots could really understand humans by identifying their emotions.

VII. Conclusion

Classification of emotion is gaining attention due to the widespread applications into various domains. Classifying emotion with high accuracy still remains an elusive goal due to the lack of complete understanding and agreement of emotion in human minds. Although various speech-based systems have been proposed for cognitive classification, their effect is still not well understood.

The results of this research show the relevance of different sound features. Our research helped us get closer to the best set of sound features. To achieve the final goal of improving emotion sound classification in the future, the best sound features should be used along with the best classification method. This method should resemble the human's perception.

This research gave a slightly different approach for making a classifying model. Our approach is not only data driven, but also biologically driven by the characteristics of the human perception gained in some psychological studies. Great importance is given to those features that are more used in human's perception.

References

[1] Thurid Vogt, Elisabeth Andr´e, and Johannes Wagner. *Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realization*, Affect and Emotion in HCI, Springer-Verlag Berlin Heidelberg 2008, LNCS 4868, pp. 75–91

[2] Scherer, Klaus R., Rainer Banse, Harad G. Wallbott, and Thomas Goldbeck. *Vocal Cues in Emotion Encoding and Decoding*, Motivation and Emotion, 1991, pp. 123-148

[3] Masaki Kurematsu, Jun Hakura and Hamido Fujita. *An Extraction of Emotion in Human Speech Using Speech Synthesize and Classifiers for Each Emotion*, International Journal Of Circuits, Systems And Signal Processing

[4] Sherif Yacoub, Steve Simske, Xiaofan Lin, John Burns. *Recognition of Emotions in Interactive Voice Response Systems*, HPL, 2003

[5] Laura Caponetti, Cosimo a Buscicchio and Giovanna Castellano. *Biologically inspired emotion recognition from speech*, Eurasip journal on advances in signal processing, 2011

[6] Jarosław Cichosz, Krzysztof Ślot. *Emotion recognition in speech signal using emotion extracting binary decision trees*, Institute of Electronics, Technical University of Lodz,, Poland

[7] Keshi Dai, Harriet J. Fell, and Joel MacAuslan. *Recognizing Emotion in speech using neural networks*

[8] Ingo Mierswa and Katharina Morik. *Automatic Feature Extraction for Classifying Audio Data*, Kluwer Academic Publishers, 2004

[9] N. Ackovska, S. Bozinovski. *Biped Robots: From Inverted Pendulum to Programming 12dof Dancing Postures*, Proc. Seventh International Conference for Informatics and Information Technologies, 2010, pp.3-7

[10] Linguistic Data Consortium, *Emotional Prosody Speech*, www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S2, University of Pennsylvania.

[11] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B. *A database of German emotional speech*. In: Proceedings of interspeech 2005, Lisbon, Portugal (2005)

[12] Engberg, I.S., Hansen, A.V.: *Documentation of the Danish motional Speech Database (DES)*. Technical report. Aalborg University, Aalborg, Denmark (1996)

[13] Kwon O., Chan K., Hao J., Lee T. *Emotion Recognition by Speech Signals*, Proc. of Eurospeech 2003, Genewa, p. 125-128, September 2003.

[14] Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B. *A Database of German Emotional Speech*; Proc. of Interspeech 2005, pp. 1517 – 1520, Lizbon 2005.

[15] David Gerhard (2003); *Pitch Extraction and Fundamental Frequency:History and Current Techniques*; Technical Report, Department of Computer Science University of Regina Regina, Saskatchewan, CANADA

[16] Tetsuya Shimamura and Hajime Kobayashi (2001); *Weighted Autocorrelation for pitch extraction of noisy speech*; IEEE Transactions on speech and audio processing, Vol. 9, No. 7