Automatic classification method of the speaker’s emotional state by voice

K T Koshekov¹, V Yu Kobenko², R K Anayatova¹, A A Savostin³ and A K Koshekov³

¹JSC «Civil Aviation Academy», Almaty, Republic of Kazakhstan
²Omsk State Technical University, 11, Mira ave., Omsk, 644050, Russia
³Manash Kozybayev North Kazakhstan University, Petropavlovsk, Republic of Kazakhstan

Abstract. This article addresses the problem of developing an effective method for automatically classifying the aviation personnel emotions (announcer) by voice. To this end, it is possible to create a dictatorial independent algorithm capable of performing a multi-grade classification of the seven emotional states of a person (joy, fear, anger, sadness, disgust, surprise and neutrality) on the basis of a set of 48 informative features. These features are formed from the digital recording of the speech signal by calculating Mel Frequency Cepstral coefficient and the main tone frequency for individual recording frames. The increase of informativeness and the reduction of the dimension for the Mel Frequency Cepstral coefficient is achieved by processing said coefficients with the aid of a deep, convergent neural network. The model of the classifier is realized by means of logistic regression, which was trained on the basis of emotionally colored English speech samples by these informative features. As a result of the training on the test sample, the correct recognition response accuracy is equal to 0.96. The inventive solution can be used for improving human-machine interfaces, as well as in the field of aviation, medicine, marketing etc.

1. Introduction

Over the past decade automatic speech recognition systems have been actively applied to a wide range of tasks in the field of modern human-machine interfaces, as well as in the field of security, where rapid situation assessment based on incoming voice data is required. This was made possible in large part by the increase in the number and availability of high-performance computing complexes and the active development of information and communications technologies.

However, modern market conditions and social and economic trends in society require new challenges. Areas of human activity such as medicine, marketing, security, control of personnel in hazardous industries or in transport, encourage researchers to find new and effective tools for automatically recognizing a person’s emotional state by their voice. Automatic emotion recognition is also necessary to elevate the human and computer interface systems used to a higher level. In addition, when solving this problem, it is possible to automatically determine the level of stress and fatigue, recognize depressive states, prevent fatigue, etc.

A significant positive effect of automatic voice-based emotion recognition introduction can be expected in industries where communication is predominantly through speech, without additional visual aids and where it is essential to reduce possible sources of risk to people and property through a continuous process of identification and monitoring of risk factors. These requirements relate primarily to aviation, [1] and other modes of transport.
It should be noted that, despite the importance of the problem identified, there is no general theory that discusses the relationship of the speaker’s emotions to the characteristics of the voice signal at present [2-4]. This fact largely determines the approach to be taken in solving the problem of automatic classification of emotions according to the characteristics of the voice acoustic signal: it is necessary to search for new algorithms and deep optimization of already existing ones for solving specific applications.

2. Research objective
Based on the above, this paper solves the problem of developing an effective algorithm for automatically recognizing the emotions of an announcer.

In the search for this problem solution among the many difficulties that arise, the fuzziness and ambiguity of the existing formulations of the concept of emotion, as well as theoretical models of their classification, should be distinguished. In this regard, when implementing the system of automatic recognition of the speaker’s emotions by voice, a set of archetypal emotions will stand out, which include joy, fear, anger, sadness, disgust, surprise and neutral state (calm) [5]. In this case, the purpose of automatic classification will be to determine the probability of assigning the emotional state of the announcer to each of the seven listed classes.

3. Theory
3.1. Development of a training dataset
Due to the specific nature of the task involved, it is inevitable that in the formation of the emotional body it is difficult to get a spontaneous emotional speech from the announcer. This makes it very difficult to accumulate a large number of marked sound recordings with an emotional color, which is rather rare in everyday human activity.

The generally applicable solution to this problem is the use as an emotional body of recordings of model bases formed with the participation of professional actors. It can be expected that the use of an automatic emotional classification system built using model bases will reduce the effectiveness of spontaneous speech. However, there is definitely a clear similarity in the expression of emotion in the conversation of arbitrary announcers [5, 6]. Therefore, the corpora of professional actors' records can be used effectively to create and initially evaluate systems of emotional state voice analysis [7]. The use of representative record bases that have proved to be reliable to other researchers will help to avoid obvious difficulties in dealing with spontaneous speech and reveal the relative effectiveness of the algorithms being developed [8].

An analysis of the existing emotional settings available to date showed that the following audio recording bases would meet the objectives of the study:
- The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [9];
- Surrey Audio-Visual Expressed Emotion (SAVEE) [10];
- Toronto emotional speech set (TESS) [11].

All bases contain seven types of emotional states records, but significantly differ in the form of emotions expression, the composition of the announcers and the pronunciation variants. Such a composite body of emotions would contribute to a more variable representation of samples for classification.

3.2. Preprocessing
At the first stage of data preparation for analysis, it is necessary to exclude from the audio files those moments of time in which the speaker does not speak (pauses). As the training is based on specially prepared recording bases a simplified procedure will be appropriate.

Comparing with the given threshold of \( \text{thr signal} \), where \( n \) is the discrete reference index, and \( N \). In turn, the signal is a filtered moving average discretized sequence of absolute values of the
audio signal being analyzed, where \( h(k) \) is the pulse characteristic of the filter, \( x(n) \) is the analyzed speech signal. The transfer function of the filter used is:

\[
H_{st}(z) = \frac{1}{W} \sum_{n=0}^{Z} z^{-\frac{k}{W}}.
\]  

(1)

It is (1) chosen to be \( W = 0.1 f_S \), where \( f_S \) is the signal sampling frequency.

As a result of the application of the threshold procedure at the output of the transformation, the desired signal \( s \) is obtained, the count of which will correspond to \( \hat{s}(n) \) count, taken when the condition \( \hat{s}(n) > thr \) is met.

The procedure described has been applied to all available samples. Informative features are isolated then.

3.3. Informative characteristics

An important feature of a speech signal is its statistical quasi-neutrality in a time region at a short intervals. For this reason, it is advantageous to analyze individual parts of the recording called frames when analyzing speech signals automatically. Within the frame the signal does not change significantly [12]. The size of the frame is determined by the size of the window, which moves along the investigated signal.

The tapering size is selected within the range of 20-100 m/s, since a smaller tapering size may result in a small number of discrete signal samples for spectral analysis and a larger number will lead to significant changes in the signal itself.

However even working with frameworks the sequence of frames contained in them contains much redundant information for analysis. For this reason, in order to classify emotions according to speech signals, it is necessary to identify informative features that can adequately represent the analyzed sample for classification and at the same time reduce the size of the input data vector.

In automatic dictator-independent speech recognition practice there is a high prevalence of Mel Frequency Cepstral Coefficient (MFCC) [13] now.

Based on the assumption that the mel scale more accurately simulates the sensitivity of the human ear, and the use of MFCC significantly reduces the space for individual characteristics of the speech signal, it can be assumed that in classifying emotions the use of mel-frequency cepstral coefficient as informative traits would be an effective solution.

The process of extraction from the MFCC input signal is as follows. After splitting the input signal \( s(n) \) into frames, we will have a \( s_i(n) \) signal for each \( i \)-frame, where \( n \) is the number of frames depending on the selected tapering length.

For each signal \( s_i(n) \) a discrete Fourier transform is performed (DFT):

\[
S_i(k) = \sum_{n=0}^{N-1} s_i(n)w(n)e^{-j2\pi kn/N},
\]  

(2)

where \( w(n) \) is the Hamming window (tapering) function used to reduce the leakage of the DFT at a finite interval; \( k \) is the DFT index in the frequency range. In digital signal processing fast DFT (FFT) is the most commonly used DF \( N = 2^j, j \in \mathbb{N} \) then.

The filter bank of \( M \) triangular filters of the form is determined then:

\[
H_m(k) = \begin{cases} 
0, & k < f(m-1); \\
\frac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \leq k \leq f(m); \\
\frac{f(m-1) - k}{f(m) - f(m-1)}, & f(m) \leq k \leq f(m+1); \\
0, & k > f(m+1).
\end{cases}
\]  

(3)
where \( m = \frac{1}{M} \) and expression (3) satisfies the \( \sum_{n=0}^{M-1} H_n(k) = 1 \) condition.

The next step is to move to a Mel Frequency Scale, which is linear below \( f = 1 \) kHz and logarithmic above, with an equal number of samples taken below and above 1 kHz [2]:

\[
B(f) = 1125 \ln(1 + f / 700).
\]

The reverse transition shall be carried out in the following manner:

\[
B^{-1}(b) = 700(e^{b/1125} - 1).
\]

Having information about the limits of the frequency range for the filter bank (3) \( f_{\text{min}} \) and \( f_{\text{max}} \) the following homogeneous base points shall be determined on the Mel frequency scale:

\[
f(m) = \left( \frac{N}{f_S} \right) B^{-1}\left( B(f_{\text{min}}) + m \frac{B(f_{\text{max}}) - B(f_{\text{min}})}{M + 1} \right).
\]

The logarithmic energy value of the spectrum component at the output of each filter is then calculated (3):

\[
P_m(m) = \ln\left( \sum_{k=0}^{M-1} S_k(k)^2 H_m(k) \right), \quad 0 \leq m < M.
\]

In the final step of MFCC calculating, a discrete cosine transform (DCT) is performed for \( P(m) \):

\[
c(l) = \sum_{m=0}^{M-1} P_m(m) \cos\left( \pi l \left(m + \frac{1}{2} \right) / M \right), \quad 0 \leq l < M.
\]

The DCT (8) is necessary for of MFCC calculation because the pulse characteristics of the synthesized filters (3) are mutually intersected and the energy at the filter output is substantially correlated. The DCT makes it possible to eliminate the resulting correlations.

After receiving \( c(l) \), the \( c(0) \) coefficient is discarded because it does not carry information about the speaker’s speech and sets a constant displacement.

On figure 1 there is the standardized estimate of calculated MFCC \( c_1(1) - c_1(13) \) for processed audio recordings of the same narrator, but with different emotional colors. The audio recordings are obtained with a sampling frequency \( f_S = 22050 \) Hz, the frame length is 512 times (about 23 m/s).

![MFCC for happy](image1.png)

![MFCC for angry](image2.png)

Figure 1. Calculated values MFCC \( c_1(1) - c_1(13) \). a is for happy; b is for angry
In recognizing emotions, it is also necessary to consider an acoustic characteristic that could be responsible for the perception of speech intonation, which is linked to the emotional expressive characteristics of speech.

According to existing studies, this characteristic may be the base tone frequency of $F_0$ [3, 4]. The Base Frequency (BF) is associated with the vibration of the vocal cords. If the ligaments vibrate quickly, the BF will be higher, and the tone of speech will be higher. This may indicate an agitated speaker. However, the monotonous speech will have a constant meaning of the BF. In addition, the BF heavily depends on the gender of the announcer. For male voices, $F_0 = 70$–200 Hz, and for female voices, up to 400 Hz.

Despite its informativeness, the precise calculation of BF remains to date a very trivial task. Difficulties arise when $F_0$ is separated from associated harmonics [14]. On the noisy audio recordings of $F_0$, it is also difficult to detect, due to the disappearance of peak values at lower frequencies.

In this study, it is proposed to use a multi-stage algorithm to track the trajectory of the BF in the audio recording of the speaker’s speech, working in the frequency area using a threshold procedure to extract maximum in the energy spectrum. Here, the efficiency of searching for spectral maximums in a limited frequency range is increased by the application of the parabolic interpolation method [15]. This algorithm is implemented as a function of the piptrack librosa 0.6.3 library for the Python 3.7 programming language.

This method differs from a relatively simple implementation, but may not work correctly if a second and other method is present in the range considered base tone harmonica.

Figure 2 shows the result of the BF determination algorithm under consideration for two types of emotional speech of the same narrator.

![Figure 2. BF definition result for a speech signal of different emotional coloration](image)

As shown in Figures 1 and 2, the visual analysis of the MFCC and $F_0$ values does not clearly separate the two types of emotions displayed. In such a situation, it is reasonable to use intelligent machine learning algorithms to construct a classifier model capable of finding hidden regularities in the data of investigated objects features.

3.4. Informative features processing

The calculated values of the MFCC are a two-dimensional structure (Figure 1), of size $m \times n$, where $m$ is the Mel frequency scale (the number of coefficients used) and $n$ is the number of signal frames. It is reasonable to assume that the informative value is not only the value of a given coefficient, but also its position in two-dimensional space. In this case, the data matrix of $M_{m \times n}$ coefficients can be viewed as a monochrome image, with each pixel playing a certain MFCC value.

As [16] practice shows, the most effective type of classifier when working with two-dimensional data sets (images) at the moment is deep convolutional neural networks (DCNN).

In this context, it is proposed to use DCNN to extract the vector of more informative features from the two-dimensional matrix $M_{m \times n}$ of the Mel-frequency cepstral coefficients, as well as to reduce the dimension of the data. For this purpose, Figure 3 presents an outline of the DCNN.
The $M_{m\times n}$ MFCC matrix enters the network (Input Layer on Figure 3). The number of coefficients $ci(1) - ci(13)$ is given for $m = 13$, and the number of frames $n = 16$. The four Conv2d 1-Conv2d 4 coalescence layers have a two-dimensional bundle operation with a kernel size of 3. The number of output filters used is 16, 32, 64, 128 respectively. The ReLu activation function is used for all matching layers [17].

In the Maxpooling 2D layer (Figure 3), a $2 \times 2$ window is used for maximum integration. To regularize the learning process use the Dropout layer, which discards every second block of input data. The subsequent layers are compacted and the output layer Dense 4 is a softmax classifier with 7 output neurons for 7 types of emotions.

The neural network was created and trained on the prepared test dataset with the help of the keras 2.3.1 library for Python 3.7. At the same time, random regions of input speech signals, measuring $16 \times 512$ samples were used for the training. MFCC values were normalized. After training, the Dense 4 output layer was removed from the network architecture. As a result, a one-dimensional vector of 32 coefficients for the calculated MFCC values is formed at the Dense 4 output.

At the same time, the calculation $n = 16$ BF values for each frame in the frequency range of 50 - 300 Hz is performed. Thus, the required vector of informative features to classify emotions according to a speech signal will contain $32 + 16$ are 48 elements.

4. Experimental results

4.1. Classifier model
Logistic regression [18] is used as a model of the multi-class classifier of emotions according to the speaker’s speech. Logistic regression (LR) is less demanding on computational resources than other models of machine learning classifiers, is well understood in theory, and also shows better results.
when working with sparse data. The data are highly sparse for the informative attributes generated by the DCNN. In addition, the LR results return the probability of the sample belonging to a given class [18]. This makes it possible to assess the spectrum of emotions in a comprehensive manner.

A training set of 70% and a test set of 30% were formed to teach the classifier from the data set that did not participate in the DCNN training. For the samples included in the datasets, the proposed informative features were calculated.

The topics are prepared as shown in Figure 4. The samples of classified signals are successively «cut» on $K$ of equal sections, length 0.4 cm. For each obtained section, the vector of informative features is calculated. Furthermore, feature values for all sites are averaged to arrive at the final value for the current sample.

**Figure 4.** Preparation of features for the classifier

Standardized estimate calculated for each derived topic (9):

$$z_{\theta} = \frac{\theta - \mu}{\sigma_{\theta}},$$

where $\theta$ is an informative feature; $\mu$ is the mean of this topic for all sample objects; $\sigma_{\theta}$ – is the standard deviation of the feature for all sample objects.

Implementation and training of the LR model was carried out with the *scikit-learn* 0.22.1 library for *Python* 3.7. As a hyperparameter of the LR model, the regularization value of the $L_2 = 1.0$ is defined. In Figure 5 there is a model error matrix obtained for test data.

**Figure 5.** Error matrix of emotion classifier model

### 5. Results and discussion

Table 1 presents the quality metrics of LR classifier model resulting.
Table 1. Quality metrics for a classifier model.

| № | Emotion type | precision | recall | f1-score | Number of samples |
|---|--------------|-----------|--------|----------|------------------|
| 1 | neutral      | 0.97      | 0.95   | 0.96     | 178              |
| 2 | happy        | 0.98      | 0.95   | 0.97     | 194              |
| 3 | sad          | 0.93      | 0.94   | 0.94     | 209              |
| 4 | angry        | 0.97      | 0.96   | 0.97     | 196              |
| 5 | fearful      | 0.97      | 0.96   | 0.97     | 216              |
| 6 | disgusted    | 0.92      | 0.96   | 0.94     | 179              |
| 7 | surprised    | 0.95      | 0.97   | 0.96     | 187              |
|    | accuracy     |           |        | 0.96     | 1359             |

The precision metric shows the proportion of objects that are classified by classifiers to a given class and that actually belong to that class. Recall (completeness) is the proportion of objects of this class from all objects of this class. The metric f1-score is an aggregated measure of the quality of classification according to precision and recall. Accuracy is the proportion of correct answers for all classes.

As can be seen from Table 1, the constructed model of the classifier makes it possible to effectively recognise seven archetypical emotions according to the speaker’s speech regardless of sex and age, since the prepared data set contains records from different actors [9-11]. The correct response rate for the entire test sample is accuracy equal to 0.96. The quality of the classification according to a number of criteria achieved exceeds the results obtained in the works of other authors and published in the press [19, 20].

6. Summary and conclusions

The inventive method and algorithm for automatically classifying a speaker’s emotional state according to a speech signal show high results of quality recognition on body emotional data of sound recordings of formed model bases with professional actors participation. This approach to the classification of emotions can be the basis for the creation of automatic systems for control of a person’s psychoemotional state by his spontaneous speech. In doing so, particular attention should be paid to the preprocessing stage for the removal of noise from signals and the selection of speech fragments on sound recordings.

The positive effect of using the described method can be expected in areas of human activity where English is used for information exchange with a limited number of words and phraseology, and where emotional expression in the voice is rare. An example of such a language is aviation English, which is used in air transport.

Reference

[1] International Civil Aviation Organization 2010 Aviation Safety Management Manual: document 9859AN/474. – 2nd edition (Monreal: ICAO) p 150
[2] Huang X, Acero A and Hon H 2001 Spoken Language Processing: A guide to theory, algorithm and system development (New Jersey: Prentice Hall) p 1008
[3] Markel G and Grey A 1980 Linear prediction of speech (Berlin: Springer) p 620
[4] Gerhard D 2003 Pitch Extraction and Fundamental Frequency: History and Current Technique (Technical Report) p 22
[5] Ekman P 1971 Nebraska Symposium on Motivation 19 pp 207–283
[6] Cornelius R 1996 The science of emotion: Research and tradition in the psychology of emotions (New Jersey: Prentice Hall) p 272
[7] Klikushin Yu, Kobenko V, Koshekov K, Belosludtsev O and Koshekov A 2017 Proc. 2016 Dynamics of Systems, Mechanisms and Machines 7819026
[8] Davydov A, Kiselev V and Kochetkov D 2011 International conference «Dialogue 2011» pp 178–185
[9] Livingstone S and Russo F 2018 PLoS ONE 13(5) e0196391
[10] Haq S and Jackson P 2009 Proc. Int. Conf. on Auditory-Visual Speech Processing p 53–58
[11] Garofolo J, Lamel L, Fisher W, Fiscus J, Pallet D, Dahlgren N and Zue V 1993 TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1 (Philadelphia: Linguistic Data Consortium)
[12] Koshekov K, Klikushin Yu, Kobenko V, Sof’ina N, Savostin A, Kashevkin A 2016 Russian Journal of Nondestructive Testing 52 pp 280–286
[13] Davis S and Mermelstein P 1980 IEEE Transactions on ASSP 28(4) pp 357–366
[14] Zahorian S and Hongbing H 2008 The Journal of the Acoustical Society of America 123(6) pp 4559–4571
[15] Sinusoidal Peak Interpolation. URL: https://ccrma.stanford.edu/~jos/sasp/Sinusoidal_Peak_Interpolation.html
[16] Nicolenko S, Kaduryn A and Arhangelskaya 2018 Deep learning (St. Petersburg: Peterburg) p 480
[17] LeCun Yan, Bengio Y and Hinton G 2015 Nature 521 pp 436–444
[18] Mohri M, Rostamizadeh A and Talwalkar A 2012 Foundations of Machine Learning (Massachusetts: MIT Press) p 504
[19] Morrison D, Wang R and De Silva L 2007 Speech Communication 49 pp 98–112
[20] El Ayadi M, Kamel M and Karray F 2011 Pattern Recognition 44(3) pp 572–587