Multi-Resolution Feature Extraction Algorithm in Emotional Speech Recognition

Ales Zelenik[1], Zdravko Kacic[2]
[1]*NXP Semiconductors Gratkorn GmbH, Mikron-Weg 1, 8101 Gratkorn, Austria*
[2]*Faculty of Electrical Engineering and Computer Science, Smetanova 17, 2000 Maribor, Slovenia*
ales.zelenik@nxp.com

***Abstract*—In this paper a new approach for recognizing emotional speech from audio recordings is presented. In order to obtain the optimum processing window width for feature extraction and to achieve the highest level of recognition rates, a trade-off between time and frequency resolution must be made. At this point, we define a new procedure that combines the advantages of narrower and wider windows and takes advantage of dynamic adjustment of the time and frequency resolution of individual feature characteristics. To achieve higher recognition rates two major procedures are added to the multi-resolution feature-extraction concept, one being the exclusion of features calculated on different processing window widths and the other the idea to use only the parts of recordings with most explicit emotions.**

To confirm the benefits of the algorithm the audio recordings from the emotional speech database Interface along with four different classifiers were used in evaluation. The highest level of emotion recognition rate with multi-resolution approach exceeded the recognition rate of the best single-resolution approach by 3.5 % with the average improvement of 1.5 % in absolute terms.

***Index Terms*—Speech, emotion recognition, segmentation, multi-resolution.**

I. INTRODUCTION

Research in the field of emotions is one of the most confusing and still open research areas [1]. By a rough estimation, more than 90 different technical definitions for emotions were created in the twentieth century [1]. Slight variations in facial expressions, voice, and posture soon received different notions for emotions, which resulted in a long list of different emotions. Examinations of biological and psychological aspects of people led to the discovery of six primary emotions that became known as the big six or basic emotions [2]. These consist of joy, sadness, fear, surprise, anger, and disgust and are generally very well accepted among scientists that deal with emotions since they do not contain any mixtures of other emotions [3].

The research work was done using the emotional speech database [4] that was recorded as a part of the project Interface (IST-1999-10036 – Multimodal Analysis/ Synthesis System for Human Interaction to Virtual and Augmented Environments) because, in addition to recordings of English, Spanish, and French speakers, it also includes recordings in Slovenian language [5]. Slovenian part of the database was recorded in a professional recording studio by two native speakers of different gender. In addition to the recordings containing basic emotions, slow and fast neutral styles were added for reference [4], [5].

Results presented in the paper were obtained using all of the available emotional speech recordings found between indexes 36 and 135 which correspond to recordings of different durations. These 3200 different recordings with a total combined duration of five hours were randomly divided into three strictly separated sets of recordings in the ratio of 60:20:20, meaning three hours of material for training set and one hour per test and verification set with the average duration of a single recording being just over 5 s.

Recordings are then split into overlapping segments, where the ones that do not contain useful information are discarded. From each of the remaining segments parameters are extracted, their values smoothed and made speaker independent. Parameters are then fed into an evaluator which selects best parameters that will serve as emotion dependent features and will be used to distinguish between different emotion classes. By fixing the segment size some features are calculated over sub-optimal segment size which lowers the final recognition rate. The main contribution of the paper is in the selection of the segment size which determines the quality and usefulness of features extracted from it. Multi-resolution feature extraction algorithm uses features calculated from different sized segments retaining the values that contribute the most to the emotion recognition task.

The present article is divided into 6 sections. The first one is a short introduction into the emotional recognition system, explaining the setup and actions taken to achieve high recognition rates. The second one describes major advances in the field providing recognition rates and emotions used. Third section commences with the algorithms used to extract and improve features from speech signal for emotion recognition and the fourth one presents emotion recognition

Manuscript received May 11, 2015; accepted July 14, 2015.

This research work was supported by Slovenian Technology Agency – TIA. Operation part financed by the European Union, European Social Fund. Operation implemented in the framework of the Operational Programme for Human Resources Development for the Period 2007-2013, Priority axis 1: Promoting entrepreneurship and adaptability, Main type of activity 1.1.: Experts and researchers for competitive enterprises.

algorithm and explains the effects of proposed improvements. In the conclusion, the findings are summed up and the last section lists references in which perspective readers can find additional information on individual procedures.

II. State of the Art

Automatic Emotion Recognition (AER) systems rely on the quality of chosen features extracted from segments of the speech signal of different size. The width of these segments, also called processing windows (PWs), determines two groups of features. When features are calculated using narrow PWs, that span between 20 and 100ms, they are called short-term segmental features [6]. To cover the whole recording the PWs are shifted and features are calculated each time until the recording is fully covered. To obtain more detailed data, the size of the shift should be smaller than the processing window width (PWW). Long-term features are calculated over longer segments like words, phrases, sentences or even over an entire recording.

Features used in AERs can be separated into groups. Features related to fundamental frequency (f_0) are included basically in every AER [7]. Second group consists of energy or magnitude coefficients and their derivates [8]. Third group relies on measuring durations of different utterances like pauses, syllables and also the presence of fillers and interjections [8]. Typical features known from speech recognition are also used for emotion recognition like mel-frequency cepstral coefficients (*MFCCs*) [9], linear prediction cepstral coefficients, log-frequency power coefficients, etc. [6], [7]. Fifth group includes voice quality features, like jitter and shimmer, harmonics-to-noise ratio that detects phonation types like creaky or breathy voice [8]. The benefit of different features is dependent on the PWW, so using a fixed PWW presents a compromise with the features used. Our method tries to overcome this compromise with simultaneously using features from different PWWs.

III. Analysis of Speech Recordings

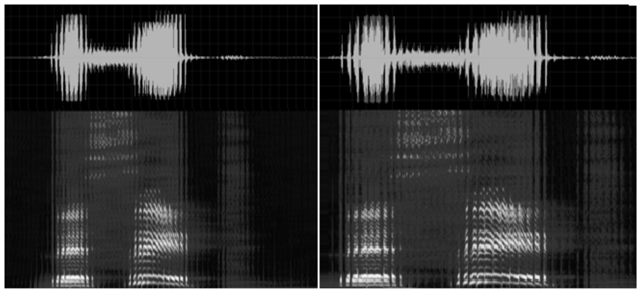
Emotional speech recognition was carried out using short-term features where the focus was on the gathering of information only from selected segments while consciously excluding information that is based on duration of individual elements of speech (duration of vocals, pauses, speech rate, etc.) in order to achieve lower computational complexity and higher robustness of recognition.

A. Selection of Processing Windows

Speech consists of a high percentage of shorter and longer pauses between words and phrases (Fig. 1(a)) [6]. In the proposed algorithm the energy based voice activity detector (VAD) was applied first (Fig. 1(b)) [10], removing noise segments from the recordings.

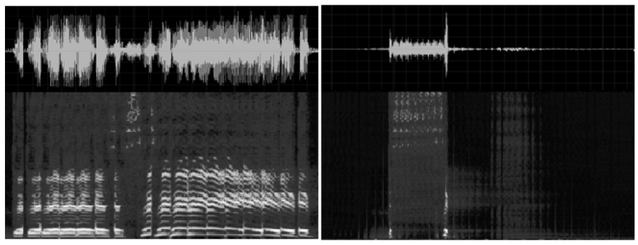
VAD is used in speech recognition where higher success rates are achieved if there is still some silence present before and after speech, as can be seen in (Fig. 1(b)) [10].

The majority of emotion recognition studies use features that rely on extracting information from lower frequencies (f_0, *MFCCs*, *tristimulus*, energy of individual frequency bands, etc.), which seldom extend above 1.5 kHz [4]. With that in mind, a very strict removal of speech signal is proposed in order to preserve only the most relevant segments of speech where lower frequency bands prevail in the speech signal. The algorithm that provided the best results compares the amount of energy of lower frequencies to the energy of higher frequencies. If the amount of energy in a frame is considerably higher (a characteristic of voiced speech), then the frame is kept (Fig. 2(a)), otherwise it is discarded (Fig. 2(b)). Discarded frames must be analysed in a way that no substantial information is discarded. As it can be seen in Fig. 2(b), only irrelevant information is discarded in this case.



(a) (b)

Fig. 1. Time and frequency responses of Slovenian word "deset" (eng.: ten) before (length of the segment – 3650 ms) a); and after VAD (length of the segment – 2350 ms) b) was applied.



(a) (b)

Fig. 2. Strict removal of non-voiced segments: a) retained parts (1100 ms), b) removed parts (2550 ms).

The result of a strict removal is a signal that represents only 30 % of the total material, which greatly reduces the calculation complexity in further procedures and also diminishes the effect of false positives in the final recognition results.

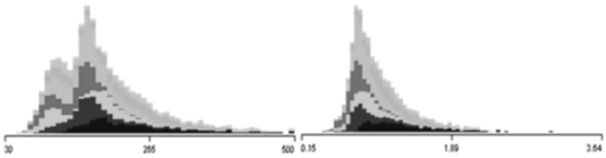
B. Feature Extraction

Feature extraction was carried out in time and frequency domains and, depending on the number of values calculated from each PW, scalar or vector features were generated. In our proposed algorithm different PWWs were used, spanning from 32 ms up to 1024 ms. From each PW, 549 different parameters of emotional speech that presented potential candidates for emotion dependent features were calculated. The total number of supported parameters presents a major upgrade of the basic system where 210 different parameter values were used [10].

Before the evaluation of parameters two additional procedures were carried out. To remove big differences in sequential parameter values a sliding window is used. This is achieved by replacing the centre value of the sliding window with the median value of parameter values inside it. The sliding window width should be found empirically; if the window used is too wide, the resulting differences in feature values could be smoothed out, but if it is too narrow, the differences will not be filtered out. In our case the best results were obtained with window width encompassing 7 values. The second procedure helped to achieve parameter values ($y_{i,j}$) that are independent from the speaker by dividing the value of each parameter ($x_{i,j}$) by the average value of this parameter from the same speaker (1)

$$y_{i,j} = \frac{x_{i,j}}{ABS\left(x_i\right)}. \tag{1}$$

The aim of this procedure is putting the same feature values obtained from different speakers closer together in the *n*-dimensional feature space, which would in turn mean achieving higher recognition rates [11]. Example of the impacts of the two procedures is seen in Fig. 3 by observing the value of f_0 distribution changes.



(a) (b)

Fig. 3. Values of f_0 before a) and after b) pre-processing is applied.

Fig. 3(a) features two distinct peaks, which is expected since male (85 Hz–180 Hz) and female (165 Hz–255 Hz) characteristic values of f_0 are different [6], [12]. After pre-processing the two peaks are no longer present (Fig. 3(b)), confirming that the correct normalization values were used. Different shades in Fig. 3 present values from different emotional classes. From different distributions on the graph it can be seen that the sole value of f_0 provides a lot of information about the emotion used. Other approaches, like splitting the recordings to male and female speakers [7], were also considered, but discarded because there are many other parameters that are not solely gender related [6].

C. Parameter Evaluation and Subset Selection

From the multitude of potential features a subset has to be selected, one that best distinguishes between the individual emotion classes [13]. For this task, a data mining tool "Waikato Environment for Knowledge Analysis" (WEKA) was used [14]. For removal of a large number of irrelevant parameters a two-step approach is proposed. In the first step a simple algorithm that still gives an estimate of the parameter contribution like Information Gain or Gain Ratio shall be used. With the use of Information Gain calculation we kept only the top 200 features for every PWW. In the second step the use of more sophisticated algorithms that evaluate different subsets is appropriate. In our case the CfsSubsetEval that *"evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them"* [15] was chosen with the addition of the GreedyStepwise algorithm that uses the hill climbing approach of gradually replacing a single parameter to find better solutions [15]. The process is repeated until improvements are present.

D. Characteristics of Selected Feature Sets

For the initial experiments feature sets with 25 features were selected for each PWW. Analysis of features obtained for three different PWWs (32 ms–128 ms) has shown that the majority of the selected features rely on f_0, *MFCC*s and magnitude or energy coefficient values. From the additional features implemented, apart from the known ones, a feature named *Lowpass* and its variants appeared in all PWWs.

The *Lowpass* feature relates to the energy distribution in lower frequencies. Its values consist of adding the square of the amplitudes of the first *n* frequency bins. When choosing the limit that will determine the number *n*, we have to take into account the width of the PW that further determines the frequency resolution and the order of FFT. *N* can also be obtained empirically, where in our case the f_0 proved to be a good candidate. Value of f_0 changes in relation to the speaker, emotion, gender, etc. and its values also change in adjacent windows, allowing us to dynamically adapt to measuring the amount of energy up to a certain point (f_0). The calculation is shown in (2) where E_n represents the energy of the *n*-th frequency bin and *k* is an additional coefficient that represents the chosen ratio between the limit and the actual value of f_0

$$lp = \frac{\sum_{n=1}^{N} E_n^k}{f_0^{\,k}}. \tag{2}$$

Number of frequency bins that will be taken into calculation, based on the desired upper limit of the energy summation, can be calculated from (3)

$$N = \left\lfloor k \frac{N_{FFT}\ f_0}{f_{vz}} \right\rfloor. \tag{3}$$

Fig. 4 shows the distribution of *Lowpass* feature values for individual emotion classes. From Fig. 4 it can be also seen that a linear separation of two emotional classes, with only a small proportion of errors, can be made just by using the values of this single feature.

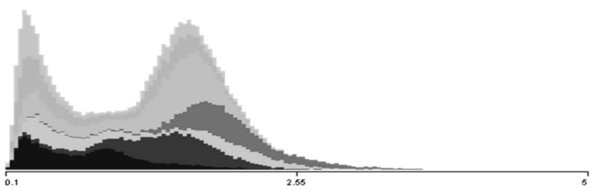


Fig. 4. Distribution of Lowpass feature values.

IV. MULTI-RESOLUTION FEATURE EXTRACTION APPROACH

To prove the benefits of the proposed approach different classifiers were tested. Best recognition results were achieved by the following four classifiers: Multi-Layer Perceptron (MLP), Random Trees (RF), Gaussian Mixture Models (GMM) and with the use of K-Nearest Neighbours classifier (KNN) [9], [13], [14]. Since the classifiers provide a result for each PW, the final emotion was determined as the one that has the most occurrences in a single recording.

A. Benchmark

Two different research papers were already presented on the Slovenian part of the Interface dataset. The first one included a rule-based emotion-dependent feature extraction method where best results were achieved with RSZ-2 feature extraction technique using 144 features and MLP for classification [11]. Recordings were split into two separate sets in the ratio of 80:20 and the highest emotion RR was 62 % [16]. The second research focused on expressive speech classification problems where three expressive classes, namely positive, negative and neutral were used [10]. The highest RRs were achieved using MLP and PWWs of 32 ms and 128 ms, where a RR of 62 % was achieved. For the classification a subset of 22 and 17 features was used [10].

Initial results, using 25 different features, two different MLP configurations, which differed in the number of hidden neurons (25-1 or 75-3) and distinguished between 7 emotional classes (basic emotions and neutral speech), produced similar results also in our case.

TABLE I. INITIAL RESULTS USING 25 FEATURES (%).

Width\class.	MLP-1	MLP-3
32 ms	46.6	54.0
64 ms	48.4	62.0
128 ms	58.7	66.5
Average	51.2	60.8

Table I shows that the average results were improved by 9.6 % in MLP-3 configuration, which used 75 hidden neurons distributed between 3 hidden layers with the highest recognition rate at 66.5 %.

B. *Proposed Improvements*

Higher accuracies were obtained with the use of proposed improvements explained before. First two columns in table II contain results achieved with the additions of smoothing the adjacent feature values with a median filter and creation of speaker independent feature values. In the last two columns the additional effects from removal of non-voiced segments algorithm are shown.

TABLE II. IMPROVED RESULTS USING 25 FEATURES (%)

Width\class.	MLP-1	MLP-3	MLP-1	MLP-3
32 ms	49.9	57.5	63.9	74.1
64 ms	61.4	71.6	69.2	74.6
128 ms	62.5	72.2	62.6	71.9
Average	57.9	67.1	65.2	73.5

Recognition rates for PWWs between 32 ms and 128 ms have risen on the average by 6.5 % in comparison with the results from Table I. With the removal of non-voiced speech the recognition rates went up by a further 6.8 %.

A strong limitation was imposed by limiting the algorithms to only a subset of 25 different features. The removal of this limitation led to larger feature subsets that contained between 29 and 55 different features (variations depending on the PWW). Apart from the results with a full subset of features, Table III additionally contains the results of wider PWs.

TABLE III. RESULTS USING A FULL SUBSET OF FEATURES (%).

Width\class.	MLP-1	STDEV	MLP-3	STDEV
32 ms	68.5	26.3	78.1	20.8
64 ms	70.6	21.7	77.7	17.9
128 ms	68.7	29.5	82.1	13.5
Average	69.3	25.8	79.3	17.4
256 ms	75.7	19.8	84.1	11.4
512 ms	75.8	19.2	83.8	9.8
1024 ms	80.3	10.3	81.5	9.2
Average	77.3	16.4	83.0	10.1

Using a full subset of features added 5 % to the recognition rates on the average but even more importantly, the wider PWs proved more successful than using narrower ones. The average difference was 5.8 %. The highest recognition rate was 84.1 % and was achieved with the PWW of 256 ms. A similar recognition rate was achieved with the 512 ms PWW where the standard deviation (STDEV) was even lower.

Apart from MLP three other classifiers were tested. Their highest recognition rates were slightly lower (best configuration is mentioned in the brackets):

– RF: 83.0 % (50 different decision trees used);
– GMM: 82.7 % (32 GMMs used);
– KNN: 81.0 % ($k = 5$).

In addition to the above mentioned widths, the 16 ms and 2048 ms versions were also tested but produced lower recognition rates and were discarded from future processing.

C. *Multi-resolution Extraction Technique*

Transition to the frequency domain is achieved by using Fast Fourier Transformation (FFT). Unfortunately the usage of a finite PWW adds errors to the final frequency spectrum. Errors can be minimized by using specific windows like Blackman-Harris window, which is used in our case. Window width determines the time-frequency resolution of the signal. If one goes up, the other goes down. Depending on the features a compromise must be made when using a fixed window width. A multi-resolution approach (MRA) tries to overcome the limitations and extract each feature at the time-frequency resolution that best suits it. For MRA feature extraction, a main processing window is constructed in a way that each off the higher levels contains twice as many narrower windows than are present on the lower levels. To obtain the same number of PWs a 2-point or 4-point reduction of feature values is made by using an average value instead of a group of values on that level. All these values are then put into the feature selection process where the best subsets among them are selected independently of the level at which they were extracted at. PWWs were divided into two groups that represented the narrower part (32 ms–128 ms) and the wider part (256 ms–1024 ms) of used PWW. Results can be seen in Table IV.

TABLE IV. MULTI-RESOLUTION APPROACH RESULTS (%)

Width\class.	MLP-1	STDEV	MLP-3	STDEV
32–128 ms	76.1	22.9	83.4	14.0
32 ms–128 ms E	79.9	14.6	85.5	9.6
256 ms–1024 ms	86.1	7.5	87.6	6.8
256 ms–1024 ms E	86.7	7.5	86.5	7.0

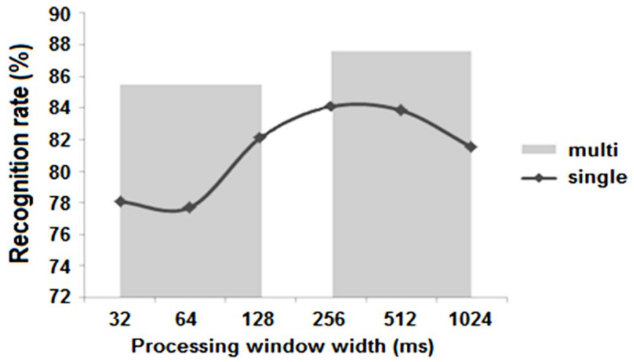


Fig. 5. Comparison of highest recognition rates between multi-resolution and single-resolution approach.

TABLE V. DETAILED EMOTIONAL RECOGNITION RATES COMPARISONS (%).

Classifier	AVG	STDEV	ANG	DIS	FEA	SAD	NEU	JOY	SUR
MRA-MLP	87.6	6.8	88.8	81.3	90.0	92.8	98.3	82.5	79.7
MLP	84.1	11.4	88.1	77.8	91.3	94.1	97.0	67.8	72.5
MRA-RF	84.0	10.7	88.4	81.3	84.4	93.1	98.9	67.8	74.4
RF	83.0	11.2	88.1	77.2	84.1	92.2	99.2	71.6	68.4
MRA-GMM	83.4	9.9	88.4	75.3	85.0	95.0	94.4	70.0	75.6
GMM	82.7	12.9	87.5	76.3	88.4	92.8	93.6	56.6	83.4
MRA-KNN	81.7	12.5	83.1	72.2	85.3	94.7	99.1	64.7	73.1
KNN	81.0	13.5	83.1	69.4	86.6	93.4	99.2	61.3	73.8
MRA-AVERAGE	84.2	10.0	87.2	77.5	86.2	93.9	97.7	71.3	75.7
AVERAGE	82.7	12.3	86.7	75.2	87.6	93.1	97.3	64.3	74.6
DIFFERENCE	1.5	-2.3	0.5	2.3	-1.4	0.8	0.4	7.0	1.1

The MRA gained another 5.5 % using the narrower part and 6.7 % using the wider part of the PWWs on the average. By examining the two feature subsets selected for the MRA, we noticed that some features, like f_0, are present on all three levels. To achieve a higher recognition rate a feature was kept only in its best rated level, whereas its values on other levels were replaced with additional features that were not included in the original subset. The results of this change are marked in Table IV with a letter ‘E’ and are on the average higher by 1.4 %. Benefits of the MRA can be seen in higher recognition rates, as shown in Fig. 5.

In Table V detailed results are gathered for all 4 classifiers. Comparing the combined results for all classifiers it can be seen that the MRA provides better recognition rates for each of the classifiers. Examining the details and looking into recognition results for every single emotion, the MRA prevails in all but one emotion. In the case of fear, better results were obtained by using a single-resolution approach with the average difference being at -1.4 %. The largest difference of +7.0 % in favour of MRA was achieved with the recognition of joy, that was actually the worst recognised emotion, making the improvement even more beneficial. Standard deviation of emotion classes improved for each of the four used classifiers with the average improvement of 2.3 %, using MRA.

V. CONCLUSIONS

Despite the large selection of machine learning algorithms they are still very dependent on the quality of the input feature values. Therefore, the main focus of research remains finding a quality subset of features [13] through which emotions in speech can be properly identified, regardless of the recording mode, language used and independently of other cultural, psychological, and sociological factors that vary between individuals [4].

Using MRA on voiced segments of emotional speech recordings obtained from project Interface [4] with other improvements proposed in this article, provided the highest recognition rate at 87.6 % among 7 emotional classes. This presented a major improvement over the previous results (62 %) obtained from the Slovenian part of the Interface database and also topped the best subjective human evaluators that achieved recognition rates between 60 % and 80 % with the best being at 83 % [10], [16]. The highest recognition rate with MRA was also higher by 3.5 % in absolute terms with the average improvement being 1,5 % compared to single-resolution approach with the same improvements.

REFERENCES

[1] R. Plutchik, “The nature of emotions”, *American Scientist*, vol. 89, no. 4, p. 344, 2001. [Online]. Available: http://dx.doi.org/10.1511/2001.4.344

[2] P. Ekman, E. R. Sorenson, W. Friesen, “Pan-cultural elements in facial displays of emotion”, *Science*, vol. 164, no. 3875, pp. 86–88, 1969. [Online]. Available: http://dx.doi.org/10.1126/science.164.3875.86

[3] J. Prinz, “Which emotions are basic?”, *Emotion, Evolution, and Rationality*, pp. 69–88, 2004. [Online]. Available: http://dx.doi.org/10.1093/acprof:oso/9780198528975.003.0004

[4] D. Ververidis, C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods”, *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2006.04.003

[5] V. Hozjan, Z. Kacic, “Interface databases: design and collection of a multilingual emotional speech database”, in *Proc. European Language Resources Association (LREC 2002)*, Grand Canarias, 2002, vol. 6, pp. 2024–2028.

[6] J. Pribil, A. Pribilova, J. Matousek, “Comparison of formant features of male and female emotional speech in Czech and Slovak”, *ElAEE*, vol. 19, no. 8, 2013.

[7] S. A. Firoz, S. A. Raji, A. P. Babu, “Automatic emotion recognition from speech using artificial neural networks with gender-dependent databases”, in *Int. Conf. Advances in Computing, Control, and Telecommunication Technologies*, 2009.

[8] I. Joyline Christina, A. Milton, “Analysis of all pole model to recognize emotions from speech signal”, *2012 Int. Conf. Computing, Electronics and Electrical Technologies (ICCEET)*, 2012. [Online]. Available: http://dx.doi.org/10.1109/ICCEET.2012.6203899

[9] S. G. Koolagudi, N. Kumar, K. S. Rao, “Speech emotion recognition using segmental level prosodic analysis”, in *Int. Conf. Devices and Communications (ICDeCom)*, 2011. [Online]. Available: http://dx.doi.org/10.1109/icdecom.2011.5738536

[10] A. Zelenik, B. Kotnik, Z. Kacic, A. Chowdhury, “Novel expressive speech classification algorithm based on multi-level extraction techniques”, in *5th Int. Conf. Pervasive Computing and Applications*, 2010. [Online]. Available: http://dx.doi.org/10.1109/ICPCA.2010.5704138

[11] V. Hozjan, Z. Kacic, “A rule-based emotion-dependent feature extraction method for emotion analysis from speech”, *The Journal of the Acoustical Society of America*, vol. 119, no. 5, p. 3109, 2006. [Online]. Available: http://dx.doi.org/10.1121/1.2188647

[12] I. R. Titze, D. W. Martin, “Principles of voice production”, *The Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 11–48, 1998. [Online]. Available: http://dx.doi.org/10.1121/1.424266

[13] R. Lileikyte, L. Telksnys, “Quality measurement of speech recognition features in context of nearest neighbour classifier”, *ElAEE*, vol. 118, no. 2, 2012.

[14] S. Srivastava, “Weka: a tool for data preprocessing, classification, ensemble, clustering and association rule mining”, *IJCA*, vol. 88, no. 10, pp. 26–29, 2014. [Online]. Available: http://dx.doi.org/10.5120/15389-3809

[15] M. A. Hall, “Correlation-based feature subset selection for machine learning”, PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[16] V. Hozjan, “Podatkovno vodeno modeliranje emocij z nevronskimi mrežami v emocionalni sintezi govora”, PhD thesis, Maribor, 2003.