IMPROVED HMM FOR AUTOMATIC SPEECH EMOTION RECOGNITION

Ž. Nedeljković[1], M. Milošević[1], Ž. Đurović[1]

[1] School of Electrical Engineering / Signals & Systems Department, University of Belgrade, Belgrade, Serbia
nz135003p@student.etf.bg.ac.rs, mm125026p@student.etf.bg.ac.rs, zdjurovic@etf.bg.ac.rs

Abstract: **Attempting to improve the performance of an automatic speech emotion recognition systems has been a daily endeavour. Diverse classification methods have been used as a part of this undertaking. It appears that HMM, a structure that was intensely exploited in previous research, is being rarely explored in new studies. We believe that it is precisely the particular structure of HMM that can introduce an additional quality into modern automatic emotion recognition systems, if not into a single classifier setup, then as a part of the multiclassifier structure. The aim of this paper is to improve discrete HMM as a first step in the process of successfully introducing it into modern automatic emotion recognition systems. The main contribution relates to the use of a new method of vector quantisation based on the QQ-plot. The proposed algorithm was tested, in combination with a wide set of features, on various emotional speech databases.**
Keywords: **Emotion recognition, HMM, QQ-plot**

I. INTRODUCTION

Nowadays, artificial intelligence is increasingly present in most diverse applications, simultaneously imposing the need for further research and additional improvements. One very important aspect of sophisticated artificial intelligence systems that are supposed to function in the human environment is emotional awareness. However, the perception of emotional states is very often a difficult task even for a human, to whom it represents one of the innate forms of communication, so automatic emotion recognition is indisputably a very challenging task.

The field of automatic emotion recognition is a relatively young one, with a lot of elements being inherited from intensively analysed automatic speech recognition systems for the purpose of the speech emotion recognition. Hidden Markov Model (HMM), as a classification structure, is also inherited and often used in plenty of works with the early start of this field [1]. More frequent drawing of the conclusion that HMM generative structure is not the most appropriate for the purpose of emotion classification [2] has resulted in HMM being quite rarely considered in recent studies, with discriminative algorithms, such as Support Vector Machine (SVM) [3] and Deep Neural Network (DNN) [4], being predominantly used in the latest research. However, it is reasonable to believe that there are cases where precisely the particular structure of HMM can contribute to modern automatic speech emotion recognition systems, if not as a standalone classifier, then as a part of the multiclassifier structure.

The first step towards the successful incorporation of HMM into modern speech emotion recognition systems is considering the possibilities of HMM classifier improvement. Over the years, there have been plenty of proposals for the quality improvement of the basic structure of HMM model [5], [6], [7], as well as for creating a hybrid structure [8], [9], but none of these has been generally accepted as a replacement for the basic implementation [10]. Some of the reasons are insufficient robustness in comparison to the basic implementation, excessive complexity in comparison to the contribution, excessive distinction from the original idea, or it is just that the presentation of the proposed changes has not been sufficiently convincing in order to be accepted and further used by other researchers.

Starting from the basic implementation of the discrete HMM model as a solid base, we investigate possibilities for robust improvement without change of the basic structure. One very challenging task of discrete HMM is vector quantisation, where it is necessary to present the vectors of arbitrary dimensions with the finite set of symbols. Most frequently, this task is entrusted to a very simple algorithm based on K-means clustering or some variation, although there have also been proposals related to more significant changes regarding the vector quantisation step [11].

In this paper, we propose a completely new method of the vector quantisation based on the quantile-quantile (QQ) plot [12]. Also, we introduce a pre-training of HMM model aimed at reducing sensitivity to the choice of initial parameters. Furthermore, multiple training with model combining is developed in order to improve the robustness and the stability of the results. The effectiveness of the proposed algorithm, in the field of automatic emotion recognition, is tested in a robust manner, by engaging several emotional speech databases in combination with various feature extraction methods. For the comparison, the results obtained with standard HMM implementation are presented as well.

The remainder of the paper is organised as follows. Section 2 describes the emotion recognition system and introduces the proposed method. Section 3 presents and discusses the results. Section 4 concludes the paper.

II. SPEECH EMOTION RECOGNITION SYSTEM

The task of the automatic speech emotion recognition system is to provide the output information on the contained emotional state for the given input in the form of a speech signal. The speech emotion recognition system design includes the selection of two basic parts: a method for feature extraction from the speech signal and a method for classification based on extracted features. In order to achieve the more robust idea of the benefit of the proposed improvement, several designs have been used, all containing the same HMM classifier, and different feature extraction methods.

Speech signal characterization was carried out by using spectral features: Linear Prediction Cepstral Coefficients (LPCC) [10], Log Frequency Power Coefficients (LFPC) [13] and Mel Frequency Cepstral Coefficients (MFCC) [14]. The features of the speech signal and its deltas and double deltas were calculated on frames of 16ms duration and 9ms overlap.

For the purposes of the selected design evaluation, it is necessary to form sets of training and testing samples. Then the selected design analysis is conducted in two steps – training and testing. The training implies the extraction of features from the training samples and training the classifier, whereas the testing implies the extraction of features from the set of testing data and making the decision on the contained emotion based on the classifier output. Design validation rate is determined based on the number of correctly classified testing samples.

Experiments were conducted with three Slavic and one German language databases: GEES [15], PES [16], RUSLANA [17] and Berlin [18]. In the case of PES, RUSLANA and Berlin databases all sentences containing joy, anger, sadness, fear and a neutral emotional state were used. In the case of the GEES database, only long sentences were used for these emotional states.

A. HMM classifier – standard implementation

Many studies on automatic speech emotion recognition used HMM as a classifier [1], [13], [19], [20]. In several studies, the configuration of discrete ergodic HMM with four hidden states and 64-symbol codebook has been proposed as a good choice for performing the automatic speech emotion recognition task [1], [13]. The idea of HMM standard implementation is given below, and the details of implementation can be found in [10], [21].

HMM is a doubly embedded stochastic process. The first (hidden) stochastic process describes the transition between hidden states. The future state of the process depends only on its current state (it is independent of past changes). The first process can be observed only through the second stochastic process. The second process generates different observations depending on the current state of the first process and independent of previous states and observations. The model is described using three parameters: state transition probability matrix, observation probability matrix, and initial state probability vector. The Baum-Welch algorithm [10] is used to train the HMM model. The model probability matrixes are initialised by random values. Each emotional state is modelled by different HMM trained to fit training observation sequences extracted from training utterances set for that emotion. In the test phase, the forward procedure is used to evaluate the probability of the observation sequence extracted from the testing utterance. The emotion whose model is met with the highest probability is declared as the recognized emotional state. The discrete HMM model takes a series of symbols (observation sequence) as an input, so it is necessary to perform the vector quantisation of feature vectors. K-means clustering with 64 clusters is used for that purpose.

B. HMM improvements

Starting from the previously described basic implementation, the proposal of the improvements is given below. The proposed approach does not change the structure of the HMM model, but do affect the process of its formation with the aim to improve the robustness and accuracy.

The possibilities of QQ-plot application on the vector clustering task have been previously tested with promising results [12]. The possibilities of QQ-plot application for the vector quantisation in HMM are analysed in this work. The description of QQ-plot application for the purpose of distributing given samples into groups is described first, followed by the description of the application for the purpose of the vector quantisation. Also, several additional improvements are proposed, positively affecting the quality of HMM model formation.

QQ-plot grouping

QQ-plot [22] can be used for determination of goodness-of-fit of the given sample set and theoretical distribution. If the given sample set and the theoretical distribution have the same or linearly dependent probability density functions, the result is approximately linear QQ-plot, otherwise, there is a deviation from the straight line.

For the given feature vectors, the formation of QQ-plot was performed as described in [12]. The scalar

representations of vectors were formed first, by using the Fibonacci sequence of numbers. When forming scalar representations, it was necessary to define the priority of vector coordinates, which was done based on random permutation. Based on those scalar representations, QQ-plot was formed in relation to the normal theoretical distribution. Afterwards, a piecewise linear approximation of QQ-plot was determined, with the points belonging to the same linear segment being proclaimed as the representatives of the same group.

The vector quantisation approach selected in this paper implies the simultaneous separation of points in exactly two groups, so the formation of QQ-plot piecewise linear approximation was optimized accordingly. The problem came down to the determination of the optimum boundary between segments of piecewise linear approximation according to the criterion of the minimum squared distance between QQ-plot points and piecewise linear approximation. The assay was primarily done for several equidistant points of QQ-plot, thus locating the optimal point environment. Further on, the interval halving method was used to reach the final solution. The separation of the given points into groups was completed by forming the piecewise linear approximation of QQ-plot.

QQ-plot vector quantisation

The training of the vector quantizer was done by using the iterative procedure based on the previously described QQ-plot grouping. The aim was to divide the given set of samples into a given number of groups. Two parameters were joined to each group: centroid and dissipation. The centroid was the representative of the group and it was calculated as the mean value of all samples in the group. Dissipation was calculated as the summary deviation of the samples from the centroid. In each iteration of the algorithm, the group with the greatest dissipation was separated into two. The procedure was repeated until reaching the given number of groups.

The separation of groups was done by using the previously described approach based on QQ-plot. Since the procedure of defining the groups by using QQ-plot included the step referring to the transformation of vectors into the scalar form, a piece of information was inevitably lost. In order to achieve a more robust result, the QQ-plot grouping was repeated for various permutations of coordinate priorities. The number of repeats in this paper was 30. The newly-formed groups with the least dissipation were kept as the final ones.

After the given number of groups were formed, the group centroids were kept as the vector quantisation etalons. The vector quantisation procedure itself was conducted in a conventional manner, where an arbitrary vector was quantised by representing with the symbol of the nearest centroid.

Additional HMM improvements

The most important additional improvement was introducing the pre-training step, which included the training of one initial HMM model based on all training samples (for all emotions). The parameters of the formed HMM model served as initial (rather than random) ones in forming the models for individual emotions, thus positively affecting convergence in forming those models, and more importantly, contributing to the improvement, as well as reducing the dissipation of final results.

There is a restriction related to the Baum-Welch iterative algorithm for HMM model training in terms of convergence towards a local optimal model. The choice of initial parameters directly influences the local optimum towards which the algorithm is converging. Setting the random values as the initial parameters results in having no control at all over the direction of convergence. With the introduction of pre-training, the models of individual emotions started training by using meaningful values of parameters, and the convergence was performed according to the distinctiveness of training samples for the particular emotion, which gave a greater chance for the convergence towards the globally optimal model. The coincidence still occurred through the initial selection of parameters of the initial HMM model, but since the initial model was trained with much more samples and since additional training was conducted for the models of individual emotions, there was much less impact of the coincidence on the final result.

The next improvement referred to the introduction of multiple training of the individual emotions models. In order to increase the chance of finding the globally optimal model, the training procedure (including pre-training) was repeated several times. In this paper, the repetition of the training was performed three times. Finally, the combination of the models of individual emotions which best fits the training samples was chosen, thus additionally increasing the chance for obtaining higher quality results.

Introducing the selection procedure for the optimal combination of HMM models resulted in creating the possibilities for inclusion of the corrective training as well. The corrective training implies the correction of HMM model parameters (more precisely, observation probability matrix) aiming to reduce the classification errors over the training set [10]. Since the corrective training performs correction of parameters in relation to individual samples, generalizing characteristics of the model are being disturbed, which might have a negative effect on the quality of the model. The introduction of HMM models combining procedure made it reasonable to introduce the corrective training in the entire procedure as well, without the fear of reducing the final model quality. In order to ensure the generalizing

characteristics, the models obtained by the corrective training were subject to another iteration of the Baum-Welch algorithm, with all training samples for the given model. Thus obtained corrected models appeared as additional models in the selection of the optimal combination.

III. RESULTS AND DISCUSSION

The experiments were organized with the aim of quality evaluation of the suggested method for HMM model formation. In order to achieve the higher quality evaluation, the tests were conducted by using several databases, engaging various features, but with the same improved HMM implementation. The experiments were repeated for the standard HMM implementation as well.

The tests were performed in the speaker independent setup, which implies that all sentences from one speaker were used for testing and all sentences from all the other speakers were used for training. This means that the voice of the test speaker was unfamiliar to the classification system. Training and testing of the classification system were repeated several times, each time with a different speaker left out for testing. The final results were based on all test runs.

Table 1. Classification rate of all experiments

Classifier	Feature	Database			
		GEES	PES	RUSL	Berlin
Standard HMM	LPCC	57.24	52.50	30.31	55.22
	LFPC	65.18	49.75	31.73	62.50
	MFCC	53.45	44.50	30.27	59.73
Improved HMM	LPCC	**59.22**	**53.50**	**33.67**	**55.39**
	LFPC	**66.22**	**51.50**	**33.28**	**63.24**
	MFCC	**56.33**	**47.50**	**36.39**	**60.05**

Table 1 shows the percentage of successful classifications of all tests for standard and improved HMM implementation. Moreover, the comparison of the obtained numerical values was made for each database-feature pair, and better results are bolded in the table. Examining the table, it is clear that the improved HMM achieved better results regardless of the chosen feature or database, thus robustly showing that the suggested improvements, without the change in the structure, positively affected the quality of formed HMM model.

The suggested changes extended the training duration, but in a controlled manner, meaning that it was possible to define the balance between the duration of the training and the desired quality of the final results. Namely, the repeating of the algorithm sections was done in several steps of the suggested algorithm, thus minimizing the effect of coincidence, where it was reasonable to expect better performance with the increased number of repetitions.

The extension of the training duration should not be the problem from the aspect of practical applicability, considering the fact that the training is conducted once during the model formation. From the aspect of practical applicability, the testing procedure, unchanged in any way, is more important.

IV. CONCLUSION

The main focus of this paper was the improvement of HMM classifier, with the ultimate goal of introducing HMM into the design of modern automatic emotion recognition systems. Suggested changes were described in the paper, followed by robust testing with several databases and various speech features in speaker independent setup. The obtained results were compared with the results obtained by using standard HMM implementation, leading to the unambiguous conclusion that the introduction of suggested changes positively affected the quality of the HMM structure.

The unchanged structure of the model, as well as the unchanged testing procedure, are beneficial from the aspect of introducing the suggested algorithm into the existing systems, designed in compliance with the standard HMM implementation.

The future studies will review the possibilities of the additional extension of changes in HMM model, where the room for improvement could be found in the creation of scalar representations of vectors for the QQ-plot construction, towards elimination, or at least reduction, of the need for repeating of QQ-plot grouping step.

Another direction of future studies will be the design of the system which incorporates improved HMM structure into multiclassifier system, alongside the classifiers of diverse characteristics, all with the aim of achieving the more precise and robust classification of emotional states.

REFERENCES

[1] M. El Ayadi, M.S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011.

[2] R.B. Lanjewar and D.S. Chaudhari, “Comparative analysis of speech emotion recognition system using different classifiers on Berlin emotional speech database,” *International Journal of Electrical and Electronics Engineering Research*, vol. 3, no. 5, pp. 145-156, 2011.

[3] U. Jain, K. Nathani, N. Ruban, A.N. Joseph Raj, Z. Zhuang, and V.G.V. Mahesh, “Cubic SVM classifier based feature extraction and emotion detection from

speech signals," *2018 International Conference on Sensor Networks and Signal Processing*, pp. 386-391, 2018.
[4] S.K. Pandey, H.S. Shekhawat, and S.R.M. Prasanna, "Deep learning techniques for speech emotion recognition: A Review," *2019 29th International Conference Radioelektronika*, pp. 1-6, 2019.
[5] Ž. Nedeljković and Ž. Đurović, "Automatic emotion recognition from speech using hidden Markov models," *Proceedings of 59th Conference on Electrical, Electronic and Computing Engineering*, pp. AU1.6.1-5, 2015.
[6] H. Farsi and R. Saleh, "Implementation and optimization of speech recognition system based on hidden Markov model using genetic algorithm," *2014 Iranian Conference on Intelligent Systems*, pp. 1-5, 2004.
[7] Z. Han, S. Lun, and J. Wang, "Speech emotion recognition system based on integrating feature and improved HMM," *Proceedings of the 2012 International Conference on Computer Application and System Modeling*, pp. 571-574, 2012.
[8] X. Mao, L. Chen, and L. Fu, "Multi-level speech emotion recognition based on HMM and ANN," *2009 WRI World Congress on Computer Science and Information Engineering*, pp. 225-229, 2009.
[9] L. Li *et al.,* "Hybrid deep neural network-hidden Markov model (DNN-HMM) based speech emotion recognition," *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 312-317, 2013.
[10] L.R. Rabiner and B.H. Juang, *Fundamentals of speech recognition*, New Jersey: Prentice Hall, 1993.
[11] J.M. Koo, H.S. Lee, and C.K. Un, "An improved VQ codebook design algorithm for HMM," *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 357-360, 1992.
[12] Ž. Nedeljković and Ž. Đurović, "QQ-plot based clustering," *Proceedings of 6th International Conference on Electrical, Electronic and Computing Engineering* (forthcoming), 2019.
[13] T.L. Nwe, S.W. Foo, and L.C.D. Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603-623, 2003.
[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentence," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.
[15] S.T. Jovičić, Z. Kašić, M. Đorđević, and M. Rajković, "Serbian emotional speech database: design, processing and evaluation," *Proceedings of the 9th International Conference Speech and Computer*, pp. 77-81, 2004.
[16] J. Cichosz, "Database of Polish emotional speech," Retrieved October 16th, 2015, from http://www.eletel.p.lodz.pl/med/eng, 2008.
[17] V. Makarova and V.A. Petrushin, "RUSLANA: A database of Russian emotional utterances," *Proceedings of the 7th International Conference on Spoken Language Processing*, pp. 2041–2044, 2002.
[18] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss, "A database of German emotional speech," *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 1517-1520, 2005.
[19] Y.L. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, pp. 4898-4901, 2005.
[20] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 552-557, 2009.
[21] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.
[22] M.B. Wilk and R. Gnanadesikan, "Probability plotting methods for the analysis of data," *Biometrika*, vol. 55, no. 1, pp. 1-17, 1968.