Temporal Discrete Cosine Transform for Speech Emotion Recognition

Branislav Popović
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
bpopovic@uns.ac.rs

Igor Stanković
European Center for Virtual Reality
Brest National Engineering School
Brest, France
stankovic@enib.fr

Stevan Ostrogonac
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
ostrogonac.stevan@uns.ac.rs

***Abstract*—Temporal Discrete Cosine Transform (TDCT) features have shown good performance in the speaker verification task, and in this paper we utilize them in speech emotion recognition. Tests were conducted on a Serbian emotional speech database, using Neural Networks (NN) as a classifier and Mel-Frequency Cepstral Coefficients (MFFC) as a reference feature set. Even though MFCC is one of the most employed techniques in emotion recognition, our results show that the TDCT features outperform MFCCs (with the first and second derivation) with any number of hidden nodes in the network, hence proving as an excellent starting feature set for recognizing emotions in South Slavic languages.**

Keywords— emotion recognition; speech; time discrete cosine transform; mel-frequency cepstral coefficients;

I. INTRODUCTION

Spoken communication between humans is intricately linked with linguistic information (verbal content) and paralinguistic information (e.g. emotional states and gestures) [1]. Humans are able to detect different emotions easily, e.g., by listening to the emotional speech signal. Their ability to recognize emotions is natural. Non-verbal communication, the manner in which the words are spoken, carries important information about the speaker intention [2]. Researchers in the field of cognitive sciences use various modalities in order to recognize emotions [3], [4]. However, from a psychological point of view, emotions are difficult to define. Usually, there is no unambiguous answer to what the correct emotion is [5]. Speech emotion recognition plays an important role in human-machine interaction. It is used in a number of different applications, e.g., in order to direct the response of a dialogue management system, based on a user emotional state, or to improve the performance of a continuous speech recognition system [6]. Human behavior is used as a natural reference for artificial systems. Speech is a promising modality for the recognition of human emotions [7]. This is a challenging task, both from engineering and psychological point of view. The combination of different features sometimes improves the emotion recognition performance and makes the systems more robust. In this paper, we propose the use of temporal discrete cosine transform (TDCT) features for the task of text-independent speech emotion recognition in Serbian, using feed-forward back-propagation neural network with one hidden layer and different number of hidden nodes. TDCT features were originally proposed in order to improve the accuracy of a speaker verification system [8]. Their reported performances were better in comparison to the MFCC features combined with their first and second order derivatives. Therefore, we expected similar performances of the TDCT features in the task of speech emotion recognition. We have compared the effectiveness of the proposed features versus more commonly used MFCC features.

A. MFCC features

MFCC features are frequently employed in the tasks of speech, speaker, genre and emotion recognition [9]. They are motivated by human hearing, calculated per frame and provide good individual accuracy in case of noiseless signals. MFCC feature vectors are commonly appended with their first and second order time derivative estimates, known as delta and delta-delta features. Even then, these features cover relatively small interval of approximately 100 ms. Hence, they are unable to capture longer term characteristics, needed in order to improve the accuracy of the above mentioned systems.

B. TDCT features

In the method proposed in [8], TDCT features are derived from the MFCCs. Each cepstral coefficient (n-th element of a MFCC feature vector) is considered as part of independent stream. Each stream is windowed in blocks of length B and

978-1-4799-1546-0/13/$31.00 ©2013 IEEE

each window is then transformed into discrete cosine transform (DCT) coefficients. Only the lowest K coefficients are retained per each stream, as they contain most of the energy. The coefficients calculated for different streams are stack together in order to form a long MK-dimensional vector (M is the dimensionality of the input vector, K is the number remaining DCT coefficients). The TDCT feature covers temporal information from a longer time context then the conventional delta and delta-delta coefficients. The next TDCT feature vector is calculated by advancing the block by one frame.

The paper is organized as follows. In Section 2, we describe our experimental setup. The database is described in brief, and the parameters of the algorithm used for training and classification of emotional states are revealed. In Section 3, the results are presented, confirming considerations from the previous sections. Paper concludes with Section 4, providing remarks and directions of future research.

II. Experimental Setup

Two sets of features were compared: the MFCC set versus the TDCT set. For the first set we chose 12 MFCCs + delta + delta-delta, resulting in a 36-dimensional feature vector per frame. MFCCs are calculated by using a filter-bank of 24 overlapping triangular windows, and the frequency range was set to [50, 3800] Hz. In order to derive the second set, we used the lowest 12 MFCCs as our input. The block size was set to $B = 8$, and the lowest $K = 3$ coefficients that contain most of the energy were retained, resulting in 36-dimensional feature vectors as well. The combination of B and K parameters used in this paper was previously reported as the best parameter setting [8]. Blocks were advanced by only one frame. Therefore, we had the same number of frames for both feature sets. In the tests we utilized the feed-forward back-propagation NN with one hidden layer and different number of hidden nodes. The training phase was always done in 100 epochs, or until the difference between the recognition rates of two consecutive epoch become too statistically insignificant (below 0.01). For the purpose of experiments, we used the long sentences from the Corpus of Emotional and Attitude Expressive Speech ("Govorna ekspresija emocija i stavova", GEES), presented in [10]. The database contains recordings from three male and three female actors, divided into 32 isolated words, 30 short semantically neutral sentences, 30 long semantically neutral sentences and one passage with 79 words for a single emotional state. It is phonetically balanced according to the phonetic statistics of the Serbian language. The linguistic information contained in the database is

TABLE I. MFCC vs. TDCT Feature Set, 20 Nodes

Feat. Set	**12 MFCC + delta + delta-delta**					**TDCT, $B = 8$, $K = 3$**				
Label	***Anger***	***Threat***	***Joy***	***Fear***	***Sadness***	***Anger***	***Threat***	***Joy***	***Fear***	***Sadness***
Anger	**20,00**	28,89	15,56	2,22	33,33	**26,11**	35,00	12,78	2,22	23,89
Threat	7,22	**53,89**	3,33	1,11	34,44	4,44	**66,11**	2,78	2,22	24,44
Joy	6,67	10,56	**45,56**	3,33	33,89	5,56	14,44	**46,67**	6,67	26,67
Fear	0,00	0,56	3,89	**22,78**	72,78	0,00	1,67	2,78	**26,67**	68,89
Sadness	0,00	0,00	0,00	0,00	**100,00**	0,00	0,00	0,00	0,00	**100,00**

TABLE II. MFCC vs. TDCT Feature Set, 40 Nodes

Feat. Set	**12 MFCC + delta + delta-delta**					**TDCT, $B = 8$, $K = 3$**				
Label	***Anger***	***Threat***	***Joy***	***Fear***	***Sadness***	***Anger***	***Threat***	***Joy***	***Fear***	***Sadness***
Anger	**32,22**	26,11	17,78	1,67	22,22	**40,56**	25,56	18,33	2,22	13,33
Threat	10,56	**51,11**	3,89	1,11	33,33	11,11	**68,89**	3,33	0,00	16,67
Joy	11,67	10,00	**49,44**	7,22	21,67	12,78	17,22	**47,78**	3,33	18,89
Fear	0,00	1,11	2,78	**32,22**	63,89	0,00	1,67	1,11	**39,44**	57,78
Sadness	0,00	0,00	0,00	0,00	**100,00**	0,00	0,00	0,00	0,00	**100,00**

TABLE III. MFCC vs. TDCT Feature Set, 60 Nodes

Feat. Set	**12 MFCC + delta + delta-delta**					**TDCT, $B = 8$, $K = 3$**				
Label	***Anger***	***Threat***	***Joy***	***Fear***	***Sadness***	***Anger***	***Threat***	***Joy***	***Fear***	***Sadness***
Anger	**44,44**	22,78	21,11	2,78	8,89	**55,00**	21,67	13,33	2,22	7,78
Threat	12,78	**54,44**	10,00	1,67	21,11	17,78	**61,67**	3,89	4,44	12,22
Joy	12,78	11,11	**53,89**	3,89	18,33	15,56	7,22	**58,33**	3,33	15,56
Fear	0,00	1,11	2,22	**36,11**	60,56	0,56	1,11	0,00	**42,78**	55,56
Sadness	0,00	0,00	0,00	0,00	**100,00**	0,00	0,00	0,00	0,00	**100,00**

TABLE IV. MFCC VS. TDCT FEATURE SET, 80 NODES

Feat. Set	12 MFCC + delta + delta-delta					TDCT, $B = 8$, $K = 3$				
Label	***Anger***	***Threat***	***Joy***	***Fear***	***Sadness***	***Anger***	***Threat***	***Joy***	***Fear***	***Sadness***
Anger	**48,33**	23,33	19,44	2,22	6,67	**55,00**	27,22	12,78	2,22	2,78
Threat	14,44	**51,11**	10,00	0,56	23,89	15,00	**68,33**	3,33	3,33	10,00
Joy	12,22	5,56	**58,33**	2,78	21,11	18,33	15,00	**55,56**	2,78	8,33
Fear	1,67	1,11	2,22	**34,44**	60,56	1,11	1,67	1,67	**45,56**	50,00
Sadness	0,00	0,00	0,00	0,00	**100,00**	0,00	0,00	0,00	0,00	**100,00**

TABLE V. MFCC VS. TDCT FEATURE SET, 100 NODES

Feat. Set	12 MFCC + delta + delta-delta					TDCT, $B = 8$, $K = 3$				
Label	***Anger***	***Threat***	***Joy***	***Fear***	***Sadness***	***Anger***	***Threat***	***Joy***	***Fear***	***Sadness***
Anger	**60,56**	15,56	13,33	3,89	6,67	**58,33**	26,11	9,44	2,22	3,89
Threat	28,33	**43,89**	8,89	1,67	17,22	17,78	**70,00**	2,22	4,44	5,56
Joy	21,67	7,22	**48,89**	7,78	14,44	13,33	13,89	**57,78**	2,78	12,22
Fear	1,11	0,56	1,67	**45,56**	51,11	0,56	2,78	2,22	**50,00**	44,44
Sadness	0,00	0,00	0,00	0,00	**100,00**	0,00	0,00	0,00	0,00	**100,00**

insignificant for the task of speech emotion recognition. The utterances are labeled by four primary emotional states, i.e., *happiness*, *anger*, *fear*, and *sadness*, and two attitude expressions, i.e. *commands* and *threats*, using *neutral* speech as referent. In our research, we have utilized four basic emotional states, as well as threats, while commands were removed from the classification process, due to the low accuracy, initially obtained by using both feature sets.

Each speech signal is pre-emphasized and the whole database was windowed using 30 ms Hamming windows, with 20 ms overlape between the adjacent frames. Classification was conducted and features were extracted per each frame. After classifying each frame, the results per frame were grouped together, so that all frames from a same database file are together determining the ”winning” emotion - the most frequent detected emotion among those frames will be the selected emotion for the whole file. A total number of 900 database files (6 speakers, 5 emotional labels, 30 long sentences per each speaker) were employed.

Due to the size of the database, our input matrix had over 400,000 rows (features per frame), which was too memory consuming to be trained in a single NN. Therefore, our input dataset was cut into blocks of 10,000 rows (where the frames that enter a block are chosen randomly) and the final output result is an average of all block-results. This method is proven to give results similar to those trained in a single network [11]. Furthermore, 60% of our data was used for training, 20% for validation, and 20% for testing, without any data overlapping between these three sets. Finally, the tests were conducted with different number of hidden nodes in the hidden layer, and the results are shown in the following section.

III. RESULTS

Four basic emotions and one attitude from the database were tested: anger, joy, fear, sadness and threat. The results in the following figures are presented in a 5x5 confusion matrix, the main diagonal represent correctly classified data (e.g. anger classified as anger, joy classified as joy), while all the other fields sum up to the overall error.

In Tables I-V, the results are presented for the MFCC and TDCT feature sets, after 100 epochs, with 20, 40, 60, 80 and 100 nodes in the hidden layer, respectively. In all of our experiments, sadness was the best recognizable emotion, both per frame and per file. This could be explained by the fact that this emotional state was the least expressive among all the other emotional states in our database. Similar results were reported for the same database, by using other types of features and classifiers [12]. The confusion could be observed between emotions fear and sadness, as well as anger and threat, especially in cases where we used lower number of hidden nodes. This could also be explained by the acoustic nature of this emotional and attitude expressive states. For higher number of hidden nodes, the confusion between the states of anger and threat exist in both directions.

TABLE VI. MFCC & TDCT FEATURE SET, 100 NODES

Feat. Set	12 MFCC + delta + delta-delta + TDCT, $B = 8$, $K = 3$				
Label	***Anger***	***Threat***	***Joy***	***Fear***	***Sadness***
Anger	**53,89**	18,89	20,56	3,33	3,33
Threat	26,11	**50,00**	10,56	3,33	10,00
Joy	27,22	11,67	**45,56**	5,56	10,00
Fear	2,78	2,78	3,89	**37,78**	52,78
Sadness	0,00	0,00	0,00	0,00	**100,00**

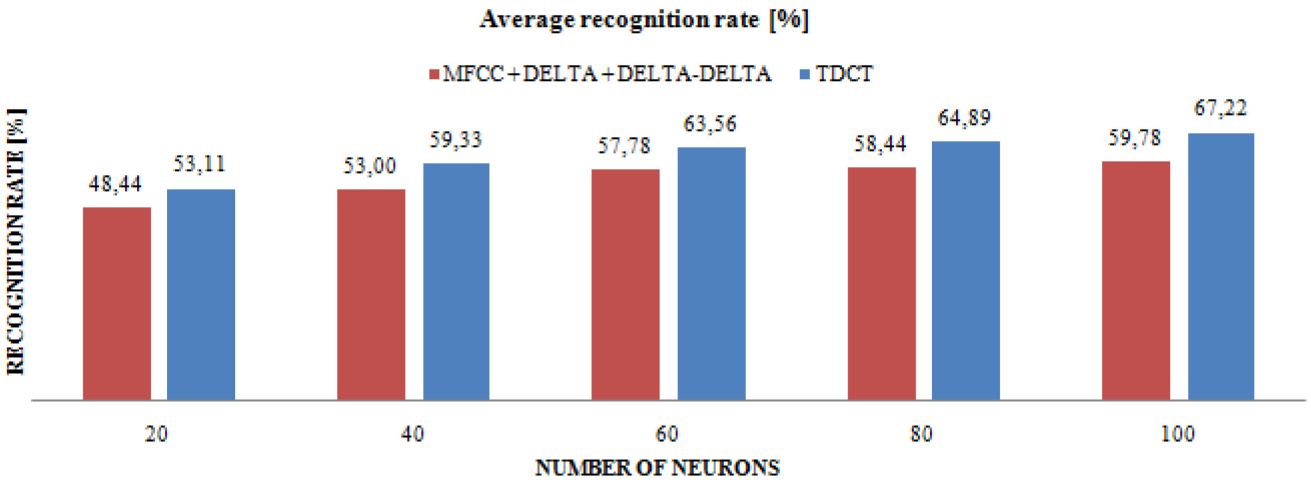

Fig. 1. Final overlook at the results

In Table VI, the results are given for the joint MFCC and TDCT feature vector, containing 12 MFCCs with their first and second derivatives, and TDCT coefficients, derived from the 12 MFCCs. The results are comparable with the results presented in Table V. The dimensionality has been doubled for the same number of training epochs and equal number of hidden nodes (72- instead of 36-dimensional feature vectors). This could explain the fact that the results were better in case where we used a unique set of features, unlike the results for speaker verification reported in [8]. The other difference is that our TDCT feature set was derived using only the first 12 MFCCs, without their first and second derivatives.

The comparison of the average recognition rates for the MFCC and TDCT feature sets is given in Fig. 1. Better recognition rates were obtained by using larger numbers of hidden nodes. For 100 nodes, and 100 epochs, the results were still improving and the network was not overtrained. For any number of hidden nodes, the recognition rate obtained by using 36-dimensional TDCT feature set, derived from 12 MFCC features, was significantly better than the recognition rate obtained by using 12 MFCCs with their first and second derivatives, as expected.

IV. Conclusion

The results presented in this paper indicate superior classification capabilities of the TDCT feature set in comparison with the reference MFCC feature set, for different number of nodes in the hidden NN layer, in the task of speech emotion recognition. We believe that the higher recognition accuracy could be obtained by using TDCT coefficients in combination with other previously suggested features [10], instead of MFCC features from which they were derived, although additional experiments are necessary in order to verify this claim. Future work will be oriented toward exploration of different feature selection techniques, in order to obtain the optimal low-dimensional set of features (TDCT coefficients in combination with several other features). Different number of hidden layers, as well as other classification techniques should be explored, including deep architectures, for which we need additional data.

Acknowledgment

The results presented in this paper were supported in part by the Ministry for Education, Science and Technological Development of the Republic of Serbia, within the project "Development of Dialogue Systems for Serbian and Other South Slavic Languages".

References

[1] A. Tawari, and M. Trivedi, "Speech emotion analysis: exploring the role of context," IEEE Trans. Multimedia, vol. 12, no. 2, pp. 502-509, 2010.

[2] S.G. Koolagudi, and K.S. Rao, "Emotion recognition from speech: a review," Int. J. Speech Technol., vol. 15, no. 2, pp. 99-117, 2012.

[3] P. Baranyi and A. Csapo, "Definition and Synergies of Cognitive Infocommunications," Acta Polytechnica Hungarica, vol. 9, no. 1, pp. 67–83, 2012.

[4] G. Sallai, "The Cradle of Cognitive Infocommunications," Acta Polytechnica Hungarica, vol. 9, no. 1, pp. 171–181, 2012.

[5] C. M. Lee, and S. Narayanan, "Towards detecting emotions in spoken dialogs," IEEE Trans. Speech Audio Processing, vol. 13, no. 12, pp. 293-303, 2005.

[6] M. Gnjatović, M. Bojanić, and B. Popović, "An adaptive recovery strategy for handling miscommunication in human-machine interaction," Proc. of 18th Telecom. Forum (TELFOR), pp. 1121-1124, 2010.

[7] V. N. Degaonkar, and S. D. Apte, "Emotion modeling from speech signal based on wavelet packet transform," Int. J. Speech Technol., vol. 16, no. 1, pp. 1-5, 2013.

[8] T. Kinnunen, C. W. E. Koh, L. Wang, H. Li1, and E. S. Chng, "Temporal discrete cosine transform: towards longer term temporal features for speaker verification," Proc. 5th Int. Symposium on Chinese Spoken Language Processing (ISCSLP), LNAI 4274, pp. 547-558, 2006.

[9] M. Gilke , P. Kachare, R. Kothalikar, V. P. Rodrigues, and M. Pednekar, "MFCC-based vocal emotion recognition using ANN," Proc. Int. Conf. Electronics Engineering and Informatics (IPCSIT), vol. 49, pp. 150-154, 2012.

[10] S. T. Jovičić, Z. Kasić, M. Djordjević, and M. Rajković, "Serbian emotional speech database: design, processing and evaluation," Proc. Int. Conf. Speech and Computer (SPECOM), pp. 77-81, 2004.

[11] I. Stanković, M. Karnjanadecha, and V. Delić, "Improvement of Thai speech emotion recognition using face feature analysis," Int. Review on Computers and Software (IRECOS), vol. 7, no. 5, pp. 2003-2015, 2012.

[12] V. Delić, M. Bojanić, M. Gnjatović, M. Sečujski, and S.T. Jovičić: "Discrimination capability of prosodic and spectral features for emotional speech recognition," Electronics and Electrical Engineering, vol. 18, no. 9, pp. 51-54, 2012.