Review of doctoral dissertation thesis

Title of the dissertation: **Classification of Emotions in Human Speech**

Author: Ing. Pavol Partila,
Faculty of Electrical Eng. and Computer Science, VŠB-TU Ostrava,

Reviewer: doc. Ing. Roman Jarina, PhD, University of Žilina, Slovakia

Motivation and relevancy of the subject

There is an active research worldwide in the given topic of emotion analysis. This topic as a part of affective computing has a great potential in behavioural modelling research. Affect and emotion analysis may help to better understand semantics of the spoken content or may be applied to adapt acoustic models in spontaneous speech recognition.

Comments on structure and content of the thesis

The manuscript is structured into 10 chapters including introductory chapter and conclusion. After the first introductory chapter, the author, in Chapter 2, briefly outlines the state of the art in emotion analysis, including theory of human emotion, list of the emotion-related speech features and common classification methods. He also mentions recent advances in development of emotional speech databases.

The thesis goals are outlined in chapter 3. The author chose the following goals: 1. Design a classifier for Speech Emotion Recognition (SER) based on Artificial Neural Nets (ANN). The most discriminative (or significant) speech features are to be selected by some optimization process. 2. Development of an emotional speech database in Czech. Performance of the developed system will be verified by comparison with present common SER systems.

Chapter 4 deals with speech feature extraction techniques. Author focuses to fundamental procedures for low-level speech feature extraction and pre-processing such as DC-offset removal, pre-emphasis, short-term energy computation, fundamental frequency computation based on autocorrelation and MFCC computation. Other few methods of feature extraction are only mentioned. From the feature space optimization, only PCA is very briefly mentioned.
I evaluate a content of this chapter as a weak and unbalanced. Instead of description of well known fundamentals procedures listed above, the author should pay much more attention to approaches of feature selection (even it is one of the goals stated in Chapter 3). In section 4.5, a list of 64 features offered by OpenSmile toolkit is presented, but it is not clear from the text if there is any own contribution to the selection process. (It seems the list of the features was taken as a whole by recommendation in ref. [64, 65]).

In Chapter 5, the following three classification methods are listed: Feed-Forward ANN with BP training, kNN, and SVM. I expected these methods would be elaborated in more depth. Especially, the author could pay more attention to SVM explanation, which is not trivial. The same for ANN; the topic of ANN (architectures, ways of training, etc.,) would need more space in the thesis, since this topic is directly related to fulfilment of the thesis goals.

Chapter 6 deals with the development of the baseline system that was trained and evaluated on freely available Berlin DB corpus. The author also proposed own approach based on parallel fusion of emotional models. Since the performance of the system was found unsatisfactory as stated by the experimental results section. 6.3, he decided to not use this approach in his further research.

The author in Chapter 7 describes the development of the new emotional speech databases in Czech. I appreciate the author suggested to utilize speech resources with spontaneous emotions since there is a lack of such labelled databases in general.

-- On the enclosed flash drive in Database/emoDBova directory, there are audio files (cut from movies) countering acted emotional speech. Therefore it has raised an issue: what kind of emotions (spontaneous or simulated) is in the developed database emoDBova. Note, it is not a correct practice to mix together both real and simulated emotions into one corpus. Please clarify this during the defence.

Next, the author investigates classification accuracy of the proposed three classifier (kNN, SVM, ANN) on the emoDBova corpus. The experimental results are followed by detailed discussion. Section 7.4 reviewed some other achievements of the author related to SER. The experimental results on stress vs. neutral state discrimination using speech data from the 112-emergency line are presented.

I have comments on the results obtained from experiments with ANN (FFBP-NN) – Tab 6.5. p.54, or Tab.7.6 p.65 – it is unrealistic to get precision of 100% for emotion classification because of ambiguous in emotion labelling. It seems the system is over-trained thus the results are not generally valid (i.e. it exactly captures the acoustic patterns in the database on which it was developed but it would not work on other databases).

A content of Chapter 8 may be considered as a core of scientific contribution of the dissertation. The doctoral candidate proposes here the multi-classifier based on fusion of outputs of three simple SER classifiers. The fusion is made on output level using three different approaches. The best results are obtained by the proposed Bayes belief integration. The author comments his achievements in comparison with other relevant works. (For Tab. 8.2 – see my comments in the paragraph above)

The thesis after concluding chapter, references and appendix, contains list of author's publications, and list of the research activities that he participated in.

Fulfilments of the thesis goals

The main scientific contribution of the thesis is laid upon (according to the 2nd goal, p. 31):
- The design of a novel classifier based on ANN
- Identification of the most significant features in speech affecting human emotions.

The goals as stated in Chapter 3, p.31 can be considered satisfied only partially. The objections are as follows:

- From the text of the thesis, I cannot evaluate the correctness of the ANN design due to lack of information provided. Information about the following important issues is missing: how the ANN type was selected; how the size of ANN was optimized or designed, and so on.
- In addition the author aimed to identify the most significant emotion-related speech features but I cannot see any part of the thesis related to this goal. He provides only list of the features taken from the previous work of other authors.

- The Ph.D. candidate also needs to clarify, what his role and scientific contribution is in the emotional Czech database development (there are 5 co-authors of the cited reference papers related to the database design [Uhr14, Uhr16] and the first author is D. Uhrin)

On the other hand, a scientific contribution can be considered in the following: a) comparison of kNN, ANN and SVM-based SER classifiers on three different speech databases, and b) proposal of the hybrid SER classification based on combination of outputs of these three independent SER classifiers. The author has experimentally proved that the proposed hybrid concept might outperform the SER classifier based on either kNN, SVM or ANN alone.

Conceptual and formal methods used

The research is experimentally oriented. The system design and tests are performed by open-source software tools. The methods used are scientific. Evaluation of the proposed approaches is based on conventional measures.

I find the thesis easily to follow. Although I am not going to evaluate the level of English (since my role is not to give comments on English grammar), readers may find quite a few incorrect wordings, inappropriate phrases or other grammar errors. The author should have done language correction before submitting the thesis.

Overall evaluation

The author has demonstrated his capability to perform research and bring scientific results. I positively acknowledge very active participation of PhD candidate in various international, national an institutional research projects and high publication activity in journals and at the conferences (indexed by WoS and Scopus). In spite of my negative comments on the fulfilment of the thesis goals I believe there is more author's own scientific contributions hidden behind the development of the SER system as seen from his dissertation thesis, therefore I do recommend the thesis for defence.

In Žilina 30/1/2017 Roman Jarina

Questions to the discussion

1. Why did you decide to use only 5 of 7 emotions of BerlinDB corpus? Why couldn't you compare your results with other systems ran on BerlinDB corpus?

2. How the level of veracity of emotions in the developed Czech database were computed? Please explain how you deal with lower veracity in some emotions (e.g. happiness) of emoDBova in the final labelling and then in the experiments with the proposed SER system (for example the confusable recordings could be withdrawn from the corpus as it was in the BerlinDB).

3. How did you scale the parameters of the kernel in SVM?