ISSN 1847-3946

mipro 2019

organizer upro

42nd

international convention

May 20 - 24, 2019, Opatija, Croatia

Lampadem tradere

mipro - innovative promotional partnership

mipro proceedings

MIPRO 2019

42nd International Convention

May 20 – 24, 2019
Opatija, Croatia

Proceedings

Conferences:

Microelectronics, Electronics and Electronic Technology /MEET

Distributed Computing in Data Science and Biomedical Engineering /DC

Telecommunications & Information /CTI

Computers in Education /CE

Computers in Technical Systems /CTS

Intelligent Systems /CIS

Information Systems Security /ISS

Business Intelligence Systems /miproBIS

Digital Economy and Digital Society / DE-DS

Engineering Education /EE

Software and Systems Engineering /SSE

Composability, Comprehensibility and Correctness of Working Software /4COWS

MIPRO Junior - Student Papers /SP

Edited by:
Karolj Skala

International Program Committee

Karolj Skala, General Chair (Croatia)
Enis Afgan (Croatia)
Miimu Airaksinen (Finland)
Saša Aksentijević (Croatia)
Slaviša Aleksić (Austria)
Slavko Amon (Slovenia)
Michael E. Auer (Austria)
Viktor Avbelj (Slovenia)
Dubravko Babić (Croatia)
Snježana Babić (Croatia)
Marko Banek (Croatia)
Mirta Baranović (Croatia)
Bartosz Bebel (Poland)
Ladjel Bellatreche (France)
Petar Biljanović (Croatia)
Adrian Boukalov (Finland)
Ljiljana Brkić (Croatia)
Marian Bubak (Poland)
Andrea Budin (Croatia)
Željko Butković (Croatia)
Željka Car (Croatia)
Jesus Carretero Pérez (Spain)
Matjaž Colnarič (Slovenia)
Lehel Csató (Romania)
Alfredo Cuzzocrea (Italy)
Marina Čičin-Šain (Croatia)
Dragan Čišić (Croatia)
Davor Davidović (Croatia)
Radoslav Delina (Slovakia)
Matjaž Depolli (Slovenia)
Saša Dešić (Croatia)
Todd Eavis (Canada)
João Paulo Fernandes (Portugal)
Maurizio Ferrari (Italy)
Tiziana Ferrari (Netherlands)
Nikola Filip Fijan (Croatia)
Renato Filjar (Croatia)
Tihana Galinac Grbac (Croatia)
Enrico Gallinucci (Italy)
Paolo Garza (Italy)
Dragan Gramberger (Croatia)
Matteo Golfarelli (Italy)
Stjepan Golubić (Croatia)
Montserrat Gonzalez (United Kingdom)
Simeon Grazio (Croatia)
Clemens Grelck (Netherlands)
Stjepan Groš (Croatia)
Niko Guid (Slovenia)
Marjan Gusev (Macedonia)
Jaak Henno (Estonia)
Bojan Hlača (Croatia)
Željko Hocenski (Croatia)
Vlasta Hudek (Croatia)
Darko Huljenić (Croatia)
Robert Inkret (Croatia)
Mile Ivanda (Croatia)
Hannu Jaakkola (Finland)
Matej Janjić (Croatia)
Darko Jardas (Croatia)
Vojko Jazbinšek (Slovenia)
Dragan Jevtić (Croatia)
Alen Jugović (Croatia)
Admela Jukan (Germany)
Oliver Jukić (Croatia)
Đani Juričić (Slovenia)
Aneta Karaivanova (Bulgaria)
Pekka Kess (Finland)
Tonimir Kišasondi (Croatia)
Zalika Klemenc-Ketiš (Slovenia)
Jan Kollár (Slovakia)
Pieter Koopman (Netherlands)
Štefan Korečko (Slovakia)
Marko Koričić (Croatia)
Gregor Kosec (Slovenia)
Goran Krajačić (Croatia)
Dieter Kranzlmüller (Germany)
Marjan Krašna (Slovenia)
Srećko Krile (Croatia)
Lene Krøl Andersen (Denmark)
Marko Lacković (Croatia)
Erich Leitgeb (Austria)
Maria Lindén (Sweden)
Tomislav Lipić (Croatia)
Dražen Lučić (Croatia)
Duško Lukač (Germany)
Ludek Matyska (Czech Republic)
Mladen Mauher (Croatia)
Igor Mekterović (Croatia)
Željka Mihajlović (Croatia)
Branko Mikac (Croatia)
Anđelko Milardović (Croatia)
Thor Moen (Norway)
Jadranko F. Novak (Croatia)
Dario Ogrizović (Croatia)
Ana Oprescu (Netherlands)
Predrag Pale (Croatia)
Mile Pavlić (Croatia)
Branimir Pejčinović (United States)
Ana Perić Hadžić (Croatia)
Dana Petcu (Romania)
Juraj Petrović (Croatia)
Damir Pintar (Croatia)
Rinus Plasmeijer (Netherlands)
Tonka Poplas Susič (Slovenia)
Zoltan Porkoláb (Hungary)
Andreja Pucihar (Slovenia)
Aleksandra Rashkovska Koceva (Slovenia)
Robert Repnik (Slovenia)
Slobodan Ribarić (Croatia)
Vittorio Rosato (Italy)
Rok Rupnik (Slovenia)
Dubravko Sabolić (Croatia)
Davor Salamon (Croatia)
João Saraiva (Portugal)
Jörg Schulze (Germany)
Zoran Skočir (Croatia)

The President of the Republic of Croatia KOLINDA GRABAR-KITAROVIĆ is a Patron of the 42nd International Convention MIPRO 2019

organized by

MIPRO Croatian Society

technical co-sponsorship

IEEE Region 8
IEEE Croatia Section
IEEE Croatia Section Computer Chapter
IEEE Croatia Section Electron Devices/Solid-State Circuits Joint Chapter
IEEE Croatia Section Education Chapter
IEEE Croatia Section Communications Chapter
EAI European Alliance of Innovation

under the auspices of

Ministry of Science and Education of the Republic of Croatia
Ministry of the Sea, Transport and Infrastructure of the Republic of Croatia
Ministry of Economy, Entrepreneurship and Crafts of the Republic of Croatia
Ministry of Public Administration of the Republic of Croatia
Ministry of Regional Development and EU Funds of the Republic of Croatia
Ministry of Environment and Energy of the Republic of Croatia
Central State Office for the Development of Digital Society
Primorje-Gorski Kotar County
City of Rijeka
City of Opatija
Croatian Regulatory Authority for Network Industries - HAKOM
Croatian Power Exchange - CROPEX

patrons

University of Zagreb, Croatia
University of Rijeka, Croatia
Juraj Dobrila University of Pula, Croatia
Ruđer Bošković Institute, Zagreb, Croatia
University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
University of Rijeka, Faculty of Maritime Studies, Croatia
University of Rijeka, Faculty of Engineering, Croatia
University of Rijeka, Faculty of Economics and Business, Croatia
Zagreb University of Applied Sciences, Croatia
Croatian Academy of Engineering - HATZ
Croatian Regulatory Authority for Network Industries - HAKOM
Ericsson Nikola Tesla, Zagreb, Croatia
T-Croatian Telecom, Zagreb, Croatia
Končar - Electrical Industries, Zagreb, Croatia
HEP - Croatian Electricity Company, Zagreb, Croatia
A1, Zagreb, Croatia

sponsors

HEP - Croatian Electricity Company Zagreb, Croatia
Ericsson Nikola Tesla, Zagreb, Croatia
T-Croatian Telecom, Zagreb, Croatia
Končar-Electrical Industries, Zagreb, Croatia
Infodom, Zagreb, Croatia
Storm Computers, Zagreb, Croatia
Transmitters and Communications Company, Zagreb, Croatia
A1, Zagreb, Croatia
Brintel, Zagreb, Croatia
Danieli Automation, Buttrio, Italy
Mjerne tehnologije, Zagreb, Croatia
Selmet Zagreb, Croatia
Institute SDT Ljubljana, Slovenia
Nomen Rijeka, Croatia

All papers are published in their original form

For Publisher:

Karolj Skala

Publisher:

Croatian Society for Information and Communication Technology,
Electronics and Microelectronics - **MIPRO**
Office: Kružna 8/II, P. O. Box 303, HR-51001 Rijeka, Croatia
Phone/Fax: (+385) 51 423 984

Printed by:

GRAFIK, Rijeka

ISSN 1847-3946

Emotion Classification Based on Convolutional Neural Network Using Speech Data

N. Vrebčević, I. Mijić, D. Petrinović
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
nikola.vrebcevic@fer.hr, igor.mijic@fer.hr, davor.petrinovic@fer.hr

***Abstract*—The human voice is the most frequently used mode of communication among people. It carries both linguistic and paralinguistic information. For an emotion classification task, it is important to process paralinguistic information because it describes the current affective state of a speaker. This affective information can be used for health care purposes, customer service enhancement and in the entertainment industry. Previous research in the field mostly relied on handcrafted features that are derived from speech signals and thus used for the construction of mainly statistical models. Today, by using new technologies, it is possible to design models that can both extract features and perform classification. This preliminary research explores the performance of a model that comprises a convolutional neural network for feature extraction and a deep neural network that performs emotion classification. The convolutional neural network consists of three convolutional layers that filter input spectrograms in time and frequency dimensions and two dense layers forming the deep part of the model. The unified neural network is trained and tested spectrograms of speech utterances from the Berlin database of emotional speech.**

***Keywords*—emotions, speech, emotion classification, convolutional neural network, deep learning**

I. Introduction

It is known that communication between people has two dimensions: verbal and non-verbal. Each dimension carries a portion of the information which can be examined both as a whole or independently. The verbal dimension will take into account only the lexical meaning of words that are produced while the non-verbal dimension comprises hand gestures, facial expressions, prosodic characteristics and emotion expressions [1]. In some situations verbal information by itself can be ambiguous and complete meaning of an utterance will be defined by a gesture, intonation, pitch fluctuations or any other component of non-verbal communication [1].

Emotions are an intriguing portion of non-verbal communication because they represent a biological response to an arbitrary event that is significant to a person involved in the event [2]. In other words, emotions unequivocally present a complex state of mind in a certain situation, where the state of mind is mainly defined by feelings, thoughts, and moods. On the other hand, emotions affect physiological state (e.g. heart rate, blood pressure, muscle movements, etc.) and thus provide a mechanism for direct mapping of an inner psychological state into a physiological state. Since physiological signals are measurable, by observing changes in such signals, it is possible to make conclusions about the presented emotions. The entire field of affective computing is based on that hypothesis and it tries to solve the problems of emotion recognition and prediction from different modalities such as facial expressions, electrodermal activity, heart rate variability, EEG and finally, speech, the modality which is focused on throughout this paper. Emotion recognition could be an useful enhancement in human-computer interaction (HCI) in a way that computers should modify their responses on user commands appropriately based on a estimated user's affective state. This improvement should take present HCI to the next level and make it more similar to the real-life interactions among people. Based on the survey in [3], affect-sensitive computer systems can find applications in the entertainment industry, intelligent vehicle systems, customer services such as call centres, health care, and research.

Speech production is a process in which a complex system of organs must perfectly collaborate to create series of sounds that are understandable and interpretable. The process results in a transformation of an abstract thought into a sequence of sounds that are spoken aloud and carry the meaning of the original thought. For this paper, the relevant subsection of the described complex chain of speech production is the mechanism of sound generation. Basically, Broca's area, the brain area responsible for speaking [4], uses the premotor cortex to send nerve impulses to all muscles involved in speech production, most prominent among them being the vocal chords that will either contract or relax. Simultaneously, the respiratory muscles will compress the lungs and a current of air will flow through the glottis. At that moment, if the vocal cords are tensioned, they will start to vibrate under the pressure of air (open and close very swiftly) and generate harmonical impulses of air which are manifested as voiced sounds. In contrast, if the cords are opened, the air flow will freely pass by and produce unvoiced sounds. Following the vocal cords, air flows through the rest of the vocal tract which ends with the lips and the nose. That part of the vocal tract can be represented by a system with a specific frequency response which defines the characteristics of the produced sound based on its shape, length, thermal capacity, the vocal tract wall elasticity and friction between the wall and the air. Specific

This research is founded by European Union, partly from European Regional Development Fund and Cohesion Fund for financial period 2014.-2020.

sounds are result of different, small perturbations in the air flow and the air pressure which can be described with linear vocal tract model excited with vocal cords.

An emotional state in speech is observable thorough different physical characteristics of sounds which are induced by changes of the vocal tract and its excitation. Different emotional states will cause modulation of nerve impulses which will result in minor oscillations of pitch, air flow and the physical properties of the vocal tract. Previously stated changes will affect vocal tract frequency response especially in higher frequencies regions [5], which is observable when statistics of frequency response is computed through sufficient period of time.

Emotion classification and prediction tasks are usually performed by modelling patterns in arbitrary data. Most widely used models are statistical models (e.g. *Hidden Markov Model*, *Gaussian Mixture Model*), machine learning models (mainly *Support Vector Machine* - SVM) and deep learning models [6]. Statistical and machine learning models use handcrafted features as information describing specific physical properties of speech and they have advantage of low complexity and theoretical interpretability. On the other hand, deep learning models have the ability of automatically learning relevant features from raw data that best suit the given task and use them to perform classification. Popular deep learning models are inherited from the domain of computer vision and they usually consist of multiple consecutive subnetworks, in most cases some combination of convolutional neural networks and deep neural networks, and recently recurrent networks.

In this paper through Section II will introduce related work with similar architecture to the proposed model. Details about the explored model are given in the Section III. Final results and discussion are presented in IV and V.

II. Related work

Emotion recognition from speech data is to the present day still a challenging research problem. One of the reasons researchers like to experiment using the speech modality is because speech is a natural way of sharing information [6]. In recent years, the field of deep learning exploded and provided tools for designing complex models from raw data. These models have very high number of parameters and when trained on sufficient amount of data, they outperform conventional machine learning models.

Authors in [7] used convolutional neural network (CNN) that have been trained and evaluated on multiple emotional speech databases. The CNN model was inherited from AlexNet [8] and evaluated using the *Leave-One-Speaker-Out* (LOSO) cross validation method. Input features for the model were spectrograms computed on 1.5 s long speech fragments using 20 ms wide analysis window. To increase the amount of training data, authors computed spectrograms from speech signals with different sampling frequencies: 16, 15 and 14 kHz. The model was defined by five convolutional layers, with max pooling at the first and the last layer, followed by two dense layers with ReLU activation and 50% dropout and softmax layer at the output. Finally they reported unweighted recall averaged over each LOSO cross validation fold. For *Database of German Emotional Speech* (EmoDB) unweighted averaged recall (UAR) was 0.71 for the seven class classification problem, the eNTERFACE corpus produced an UAR of 0.66 for six classes and SUSAS 0.57 for five classes.

III. Motivation and methodology

In this preliminary research, the main goal is to design a stable and simple model by using deep learning methods and algorithms. The model should serve as a reference for future models that will emerge from upcoming research by authors in the domain of affective computing, mainly based on an exploration of the emotion recognition using the speech modality. In Section II a model has been introduced which is, by its architecture and parameters, the closest one to the model which is explored in this paper. In that manner, another important aim of this research will be a comparison between model used in [7] and model CNN model in this research. Additionally, a performace of the CNN model will be compared with the performance of an SVM model which is frequently used in affective computing [3], [6] and the AlexNet model [8] that is usually referenced in the computer vision tasks.

According to the survey in [6], every speech classification or prediction task consists of: defining the features that contain the highest amount of affective information from speech; subsequent design of the best possible model for a given task; and finally, careful preparation of the raw speech data for model training and evaluation. This section will describe these key points in detail.

A. Dataset

For the purpose of this paper, emotional speech data from the publicly available EmoDB corpus [9] was used. It comprises utterances from 10 speakers (5 female and 5 male). Each speaker had to pronounce predefined sentences while acting one emotion at a time from a set of seven basic emotions (fear, disgust, happiness, boredom, sadness, anger and neutral). The corpus includes 535 studio recorded utterances with a sampling rate 16 kHz.

B. Features

A widely accepted way of extracting features from utterances is by using some set of predefined features, e.g. *Computational Paralinguistics Challenge* (ComParE) [10] or *The Geneva Minimalistic Acoustic Parameter Set* (GeMAPS) [11]. In general, these features describe the statistical properties of specific physical occurrences present in a sample of speech. Each physical property should be calculated on short segment (frame of analysis) in which speech is considered to have

characteristics of stationary process. Finally, the statistics will be computed over a predefined number of frames, where results may vary depending on the number of frames.

Two types of features were computed for development of the baseline model and the CNN model. First type of features are functionals defined by the GeMAPS [11] feature set and they describe statistical properties of speech in terms of low level descriptors (LLDs) which are representing different physical characteristics of an utterance. Functionals were computed by using openSMILE feature extractor [12]. The other type of features are spectrograms which are carrying complete information about speech in frequency domain, in a given time segment.

It can be observed that selected features are, directly or indirectly, describing statistical properties of observed data and the statistics will depend on a span of a time window in which it is computed. To explore dependency of a model performance and width of the analysis window, previously stated features were calculated on speech fragments of length 720 ms and 1280 ms. Utterances were fragmented into fragments of aforementioned lengths and overlap between consecutive fragments in the case of length 720 ms was 480 ms (2/3 overlap), and for the length of 1280 ms, overlap was 560 ms (1/2 overlap). For the fragments that were 720 ms long, spectrograms were calculated by a using Hamming time window of 40 ms width FFT with frame advance of 10 ms, while for the fragments of length 1280 ms, spectrograms were calculated by using 36 ms Hamming window and 512-point FFT with the window advance of 5 ms. Resulting spectrograms had dimensions 256×72, and 256×256, respectively. In Fig. 1 is illustrated the process of spectrogram calculation for speech fragments lenght of 1280 ms.

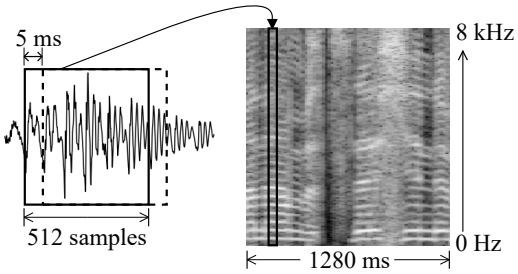

Fig. 1: Illustrated method of spectrogram calculation with parameters of analysis frame and resulting spectrograms dimensions

C. Data Augmentation

For the purpose of training a complex convolutional neural network, the data was expanded by using bootstraping techniques. More precisely, original utterances were downsampled to 14 kHz and 15 kHz (the method was adopted from [7]), a reverberation effect was applied, utterances were reversed in time and additive white and ambient noise was combined with everything previously stated. White and ambient noise were added with *signal-to-noise ratio* (SNR) of 20 dB. The white noise was derived from three different distributions: Gaussian, Laplace and uniform. The ambient noise was taken from the MUSAN dataset [13] and sounds of plane takeoff, keyboard typing and a motorcycle passing by were used. Alltogether, six noise types. The number of annotated speech fragments for each level of augmentation is displayed in Table I. The total number of original and downsampled fragments without noise is 46890 for the 720 ms fragment length and 14328 for the 1280 ms fragment length. When three types of white and ambient noise are added, the total number of fragments for clean and noisy utterances becomes 328230 and 100296 for the 720 ms and 1280 ms fragments, respectively. Finally, when the reverberation effect and the time reversed samples are added to the original and downsampled fragments and when the noise is applied, the database comprises the original clean fragments, the clean fragments with effects, the noisy fragments and the noisy fragments while the total number of fragments becomes 474110 and 144872 for the 720 ms and 1280 ms long fragments, respectively.

TABLE I: Amount of anotated speech fragments for different levels of augmentation

Augmentation type	**Fragment length** [ms]	**Utterances Number**
downsampled	720 (2/3 overlap)	46890
downsampled	1280 (1/2 overlap)	14328
noise	720 (2/3 overlap)	328230
noise	1280 (1/2 overlap)	100296
noise & effects	720 (2/3 overlap)	474110
noise & effects	1280 (1/2 overlap)	144872

downsampled - original and downsampled fragmets
noise - downsampled augmentation type extended with white and ambient noise
noise & effects - combination of downsampled augmentation type extended with reverberation and reverse effect with and without white and ambient noise

D. Baseline SVM Model

As stated above, a simple SVM model has been trained to serve as a reference for the more complex CNN models. Presuming nonlinear feature interactions, a radial basis function (RBF) kernel was used as an extension of the model, to ensure separability between data points in a high dimensional space. The model hyper parameters C and γ were tuned by using a grid search algorithm and by narrowing the range of values that C and γ should have. For the parameter C, the final range of values was $[3, 5]$ and for γ was $[0.005, 0.012]$.

The model was validated through 4-fold cross validation where each fold comprised train and test partition. Partitions in each fold were stratified by gender. The train partition was built from all utterances of a randomly selected four male and four female speakers. The remaining speakers (one male and one female) were included in the test partition. Such partition

organisation accomplishes a type of gender-stratified-speaker-independent cross validation evaluation.

E. CNN Model

The convolutional neural network comprises convolutional and dense layers. Three sequentially stacked convolutional layers define the first portion of the network. In the following description of model's architecture, strides were applied in both time and frequency dimensions of spectrograms. The first layer has 96 rectangular filters with size 7×7 and stride 3, then the second layer has 256 filters with size 5×5 and stride 1 and the third layer has 256 filters with size 3×3 and stride 1. After each convolutional layer max-pooling was applied with kernel 2×2 and stride 2. The aforementioned convolutional layers are used for custom feature extraction where the neural network aims to learn its own features throughout training. The second portion of the model consists of two dense layers with 1024 neurons which perform the classification. The final decision is made on a final dense layer with seven output neurons corresponding to seven emotion classes. The outline of the described architecture is presented in Fig. 2 with annotations of the data dimensions after each of the convolutional layers. ReLU activation has been used at three points in the model, after the third convolutional layer and after each of the dense layers. The high number of model parameters (7.626.848) along with the low number of speakers and spoken utterances in the dataset can result in poor generalisation of the model, with overfitting to specific speakers' utterances from the training partition. To accomplish better generalisation, several regularisation techniques were applied. Dropout regularisation [14] was used before and after the first dense layer and after the second dense layer. For convolutional layers, batch normalisation is used as a regularisation technique [15]. Finally, the RMSprop optimiser is used to optimize the softmax cross entropy loss function and update the model's parameters in the training phase. All the hyperparameters stated above were manually picked in the process of fine tuning of the model. Layers were jointly trained from a random state where weights were initialized by using Glorot random uniform initialization [16]. Additionally, step learning rate decay was applied which lowered learning rate by factor 0.9 every 10 epochs and an early stopping criterion was used to finish training at an optimal moment.

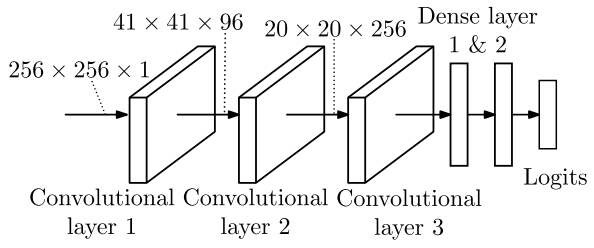

Fig. 2: Convolutional neural network architecture

For training and evaluation of the CNN model, the dataset was partitioned in a similar way as for the baseline SVM model, except for a validation partition created for the purposes of tracking the training performance of the model. In the end, training partitions were defined with three random male and female speakers, and remaining speakers were randomly organised in validation and test partition. Same as in the case of the baseline SVM model, the partitions were gender-stratified and evaluation was speaker-independent.

IV. Results

Table II summarises performances of all the trained models. As it was stated previously each model was trained with features that were calculated on speech fragments of different lengths. Additionally, it was explored how different degrees of data augmentation will affect the performance. The first type of augmentation included only ambient and additive white noise, while the second type was extended with reverberation and reverse effects. Finally, the performance of the proposed CNN model (denoted with *minimal* in the Table II) was compared with an AlexNet type model (the number of model filters and nodes in dense layers was reduced relative to the original architecture ([8]) in order to fit in the memory of used hardware). This smiplified AlexNet model is denoted *large* in the Table II. In comparison to the minimal CNN model, the AlexNet has two extra convolutional layers and its total number of trainable parameters without regularisation is 7.634.272. In average, the models reached 58% of accuracy in the training phase before their validation loss metrics started diverging (and training was stopped) . Depending on the model's architecture, the best test performance on the clean data set was achieved for the minimal model CNN-4 (32.28%) and the performance on the augmented data set was the best for large models CNN-11 (34.03%) and CNN-12 (33.17%) where original data set was augmented with additive noise and combination of additive noise and aforementioned effects, respectively. In general, the minimal model has better performance on the clean data set, in contrast to the large model which has higher accuracy in the test phase on augmented data set.

V. Discussion

Results show that a CNN model with higher complexity will have a greater ability to generalise when compared to minimal CNN models. On the other side, SVM models in this research outperform any of the given CNN models. This kind of behaviour is unclear since SVM models are trained on a small portion of handcrafted features in comparison to the CNN models which are able to extract a large number of custom features that are describing a pattern present in the data set. Depending on the length of the utterance fragments, features computed for an SVM model can contain slightly different information amounts and thus the separability of emotions in high dimensional space can increase for longer duration fragments which is observed in the test performance of an SVM model trained on features computed from fragments of

TABLE II: Comparison between the performance of the CNN model described in Section III (minimal) and customized AlexNet model (large). Models were trained with the datasets that have different augmentation levels (relative to the base dataset described in Section III)

Model	**Architecture**	**Fragment length** [ms]	**Data augmentation type**	**Train accuracy** [%]	**Test accuracy** [%]
SVM-1	—	720	—	—	52.01
SVM-2	—	1280	—	—	64.33
CNN-1	minimal	720	downsampled	50.18	30.64
CNN-2	minimal	720	noise	58.76	29.29
CNN-3	minimal	720	noise & effects	55.59	30.22
CNN-4	minimal	1280	downsampled	56.27	32.28
CNN-5	minimal	1280	noise	59.22	28.96
CNN-6	minimal	1280	noise & effects	75.35	28.81
CNN-7	large	720	downsampled	45.72	29.15
CNN-8	large	720	noise	59.05	31.32
CNN-9	large	720	noise & effects	58.87	30.64
CNN-10	large	1280	downsampled	57.51	29.55
CNN-11	large	1280	noise	62.71	34.03
CNN-12	large	1280	noise & effects	61.71	33.17

downsampled - original and downsampled fragmets
noise - downsampled augmentation type extended with white and ambient noise
noise & effects - combination of downsampled augmentation type extended with reverberation and reverse effect with and without white and ambient noise

720 ms and 1280 ms. The same behaviour is not observed in the case of the minimal CNN model where better evaluation results are achieved with shorter fragments. Regarding the augmentation level, it is observable that the test accuracy is higher for an augmented data sets in comparison to the clean sets which indicates that the EmoDB is still a small data set which does not contain a sufficient number of inter-speaker and intra-speaker variations of emotional speech and thus complex models do not have wide enough variety (and number) of examples of an emotional speech to generalise well between speakers and emotions. In between types of augmentation, better results were achieved for an augmentation with the additive noise only while the combination of an additive noise and effects is slightly inferior. It should be further explored why the effects degrade affective information since reverberation effects occur in real life situations while time reversing should not change the frequency content of an utterance or its statistics.

VI. Conclusion

The trained models are defined by the most widely used architecture in the domain of computer vision aided with deep learning. Results show that there is still room for significant improvement both with an architecture upgrade with recurrent layers and better understanding of the entanglement between data and the specific model architectures. Additionally, data sets with higher number of speakers and utterances per speaker should result in better generalisation of an CNN model, and would negate the need for data augmentation (which in the case of speech isn't as straightforward as in computer vision applications). Finally, further research should aim on devising specific model architecture used for the task of emotion classification in speech modality since models in this paper were constructed by using best practices from the domain of computer vision that might not be the optimal for the speech modality. Further work should continue to explore the model design to achieve similar or better performance than the SVM model which is relatively high when it is taken into account that the classification problem in this paper contained seven independent classes.

VII. Acknowledgements

This work was partly supported by: Croatian Science Foundation under the project number IP-2014-09-2625 and DOK-2018-01-2976; DATACROSS project under number KK.01.1.1.01.009. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Croatian Science Foundation.

References

[1] M. Argyle, "Non-verbal Communication and Language," *Royal Institute of Philosophy Lectures*, vol. 10, no. 1976, pp. 63–78, 1976.

[2] K. R. Scherer and M. R. Zentner, "Emotional Effects of Music: Production Eules." *Music and emotion: Theory and research*, pp. 361–392, 2001.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[4] R. M. Lazar and J. P. Mohr, "Revisiting the contributions of Paul Broca to the study of aphasia," *Neuropsychology Review*, vol. 21, no. 3, pp. 236–239, 2011.

[5] C. E. Williams and K. N. Stevens, "Emotions and Speech: Some Acoustical Correlates," *The Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238–1250, 1972.

[6] 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.
[7] N. Weiskirchen, R. Bock, and A. Wendemuth, "Recognition of Emotional Speech with Convolutional Neural Networks by Means of Spectral Estimates," *Conference on Affective Computing and Intelligent Interaction Workshops and Demos*, no. October, pp. 50–55, 2017.
[8] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, vol. 25, p. 04015009, 2012.
[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *Interspeech*, pp. 1517—-1520, 2005.
[10] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, no. August, pp. 148–152, 2013.
[11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.
[12] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, no. May, pp. 835–838, 2013.
[13] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.
[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.
[15] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," 2015.
[16] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proceedings of 13th International Conference on Artificial Inteligence and Statistics*, vol. 9, pp. 249–256, 2010.