Time Window Analysis for Automatic Speech Emotion Recognition

Boris Puterka [1], Juraj Kacur [2]
[1] Institute of Robotics and Cybernetics, Slovak University of Technology, Ilkovicova 3, Bratislava, Slovakia
[2] Institute of Multimedia ICT, Slovak University of Technology, Ilkovicova 3, Bratislava, Slovakia
boris.puterka@stuba.sk

***Abstract*—In this paper we present time analysis results of speech emotion recognition using convolutional neural network architecture and spectrograms as a speech features. Analyses were performed on model with two convolutional layers followed by pooling layer, and one fully-connected layer followed by dropout and softmax layer on the output. On this model we analyzed time characteristics of speech signal represented by spectrograms. The aim of our work was to find relation between duration of speech signal and the recognition rate of seven basic emotions. It was discovered that speech length is important and naturally the accuracy is growing with the length of analyzed window, however over approximately 1.2 seconds the growth becomes rather mild.**

***Keywords*—Speech Emotion Recognition; Spectrogram; NN; CNN**

I. Introduction

Deep learning has achieved huge success in wide spectrum of application. Convolutional Neural Networks (CNN) are often used in image processing where they can be seen as very successful approach [1]. In recent years, applications of CNNs in speech processing increased as well. Speech signal is a natural way to communicate with other people. There are many applications in area of Automatic Speech recognition (ASR), e.g. speaker identification, and it is inherent part in human-machine interaction. Speech Emotion Recognition (SER) has also its place in research and practical applications. Systems that can effectively recognize emotion from speech can be used in various security applications e.g. security of elderly inhabitants living alone. Other form of applications can be healthcare assistants to detect emotional state of speakers. For people with communication disorders like social-emotional agnosia or even autism, emotions are hard to understand, so such assistant may help them to rehabilitate. Sometimes it is really hard, even for healthy individuals to recognize emotions from speech, so multimodal approaches may have their role in SER tasks.

II. Related Work

There are various implementation of Feed-forward architectures e.g. CNN, in SER domain. Deep learning in both ASR and SER has replaced traditional recognition methods e.g. Hidden Markov Model (HMM) as in [3]. CNN are able to deal with inputs with high dimensions [1] and spectrograms are commonly used in SER as inputs to the CNN e.g. [2]. CNN architectures are relatively simple while applied in SER, with 2, maximum 3 convolutional layers. More focus is in signal processing part of input (which speech representation is better as input to CNN for SER). In [1] they used Mel scale filter banks on as an input feature having 40 filter bank channels. Others features such as Mel-frequency cepstral coefficients (MFCC) are also used in SER as in [4] but with Support Vector Machines (SVM) as classifier. Studies show that it is suitable to use 2D representation of speech signal as an input to CNN. However, there are various implementations with raw 1D representation of speech signal as well e.g. [5].

There is ongoing discussion on minimal speech duration to recognize emotions. In our work we focused on finding proper duration of speech signal to successfully recognize emotions from such signal.

The article is organized as follows: a summary of related work is given in section 2. Sections 3 and 4 introduce feature extraction and CNN classification tasks respectively. The database and accomplished experiments are in paragraph 5. The article is summarized in section 6.

III. Feature Extraction

Speech signal contains several rather distinctive sorts of information, e.g. lexical information, speaker identity, speaker actual physical, health and mental conditions, spoken language, dialect, educational and social background, etc. Event thought these pieces of information are well separated from human perception point of view they are combined together thorough a complex nonlinear mechanism of speech production. Thus it is rather a challenge to separate them depending on the kind of speech application. Usually only one kind of information is required at the same time, e.g. lexical information for speech recognition application, while the others are regarded as noise obscuring that one in the focus. Moreover, speech signals are corrupt by additive noises of arbitrary characteristics present in the environment, and finally they are distorted by convolutional noises introduced by the surrounding space and recording devices (echoes, room impulse responses, microphone impulse response, etc.).

Therefore a good speech features must extract such sort of information that is needed for particular deployment area and suppress the others. Most extraction methods mimic the human auditory system. Speech signal is known to be highly non stationary, i.e. changes its energy, time, frequency and statistical features in time. Therefore it is useless to describe a complex speech signal by a single static feature vector (FV). Thus speech signals are segmented into frames (mostly uniformly) that are both long enough to capture vital signal characteristic in the analyzed interval and short enough so that they can be regarded as static, i.e. different characteristic of speech are not mixed up too often. In the domain of speech recognition best results are achieved with frame lengths ranging from 10 to 30ms with 50% overlap of adjacent frames.

Auditory system is based on frequency analyses, and in the case of stationary signals at certain intervals the phases among frequencies proved not to be very discriminative. Therefore mostly spectral magnitudes are extracted [6]. Usually prior to frequency calculation using short time DFT (1) each time frame is element wise multiplied by a proper window $w(n)$ e.g. Hamming to suppress the so called spectral aliasing.

$$X_n(f) = \left|\sum_{m=0}^{N-1} w(m)x(nS+m)e^{\frac{-2\pi mf}{N}}\right| \tag{1}$$

N is the length of a frame, and S is the shift between adjacent frames (less than a frame length) listed in samples. Appling (1) to signals leads to the construction of spectrograms that present standard time-frequency division of the non-stationary signal which is vital for further analyses or classification.

Each frame captures acoustic information of a processed signal which is vital for sound perception. Even though a 20ms long frame can provide some clue about a processed phoneme (vowel, consonant, fricative, plosive, voiced, etc.) it certainly cannot distinguish among several emotions. Emotions are known to be coded and spread along longer time intervals capturing all sorts of characteristics in time. Thus block of a spectrogram containing consecutive short time spectra can be considered as a proper feature vector (FV) (2).

$$FV_b = [X_b(f)^T, X_{b+1}(f)^T, \dots, X_{b+B-1}(f)^T] \tag{2}$$

where B is the length of a block, and T denotes transposition; thus FV is a matrix of a dimension $F_{Nyquist}$ x B. It is advantageous to keep this 2D time-frequency structure for the proper classification and not to form a simple feature vector, where this structure would fall apart. Obviously the length of such a block (B) is a vital parameter influencing robustness, accuracy, training times, network complexity, etc. Roughly speaking the longer the block the more accurate results could be expected. On the other hand, there must be an upper limit as no emotion lasts forever and the longer block the fewer training samples. Finally, too long block can enormously increase the feature vector length. Thus it is important to find a proper tradeoff.

IV. Classifiaction via Convolutional Neureal Networks

Using features as stated in (2) the problem of unequal length of speech sequences reduces to fixed length samples. However these samples pose certain structure and time frequency variability, which the classification system must cope with.

There are several classification methods that can be easily applied to this problem e.g. KNN, GMM, NN, SVM, with higher or lower success rate depending on the amount and structure of data. Nevertheless, in case of limited data having high dimensional FVs, parametric methods using proper discriminative training can provide better results. Recently, Neural Networks deployed across different domains recorded best results (speech and face recognition, etc.). Furthermore the concept of NN provide wide spectrum of network structures and training strategies than can handle different sorts of signals having specific characteristic. The basic one is Feedforward network that possess the property of being universal approximator i.e. such network can with arbitrary accuracy approximate any function under some mild conditions. Furthermore using ReLu activation function [7] and modified training algorithms enabled the construction of networks with many layers (deep learning) solving complex tasks. However it lacks the ability to model time series with long history, variable input sizes, and is inefficient to cope with shift or rotations like variability of the input features. For time series Recurrent (RNN) and LTSM networks using variants of basic BPTT algorithm can be successfully applied. Mainly in cases of image and speech processing, CNNs can successfully handle some sorts of variability e.g. spatial shift. Both CNN and RNN can be extended by adding FNN or even all abovementioned structures can be combined together. An extensive overview on NN is given e.g. in [8].

Because of the construction of FV as stated in (2) where the location variability both in time and frequency is eminent, the natural choice is CNN. Thus in the following a brief description of CNN is given, for more detailed one see [9]. CNNs contain at least one but usually more convolutional layers that performs convolution (3) of the input signal *x(n)* with a given impulse response *h(n)* of length *P*.

$$y(m) = \sum_{i=0}^{P} h(i)x(m-i) \tag{3}$$

Such impulse response is subject to a training process too and can be viewed as a standard layer with shared weights among neurons in the convolutional layer (C) that each processes a shifted input. Usually such a layer is followed by so called pooling layer (P) that selects neurons in certain range (down sampling is taking place) from a convolutional layer based on certain criteria such as maximal value, minimal value, average value etc. and presenting its modified output to the next layer for further processing. It can be another convolutional layer or a standard, fully connected one (FC). As there can be more vital traits subjected to shifts there are often more filters needed to detect all of them, which are then placed in parallel planes in a given convolutional layer. Usually, after convolution or even pooling layer data normalization (N) can be applied. For a classical structure of CNN that was used also in our experiments see Fig. 1. Such network is trained as normal FNN, e.g. using variety of gradient descend and BP algorithms except convolutional layers, where the weight update is based only on the selected neuron in the pooling layer (only its local gradient is considered). It should be noted that CNN can have many structures (how C, P, FC, N layers are used and ordered). Depending on the input *x* the filter *h(n)* can be multidimensional i.e. 2D for images. This is the case here

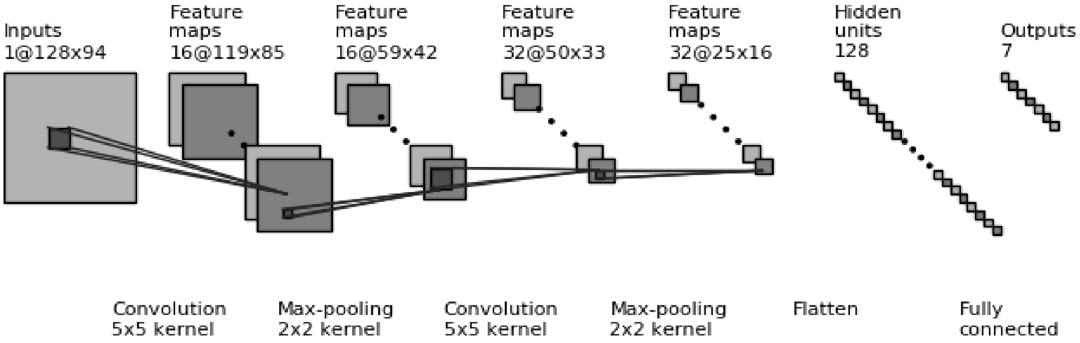

Figure 1. Architecture of CNN used in our experiments. Note that the input size varies in our experiments. This particular inputs size relates to 94 time frames which is equivalent to 1s long speech signal. Overall architecture, such as number of convolutional layers, number of feature maps in each layer, size of kernels and number of fully-connected layers is the same

having 2 dimensions, i.e. time and frequency. Then 2D convolution is calculated as (4)

$$y(m,n) = \sum_{i=0}^{P} \sum_{j=0}^{L} h(i,j)x(m-i,n-j) \quad (4)$$

V. Experiments

In our experiment we used Berlin Database of Emotional Speech [10]. Database consists of 535 utterances, each with sampling frequency$F_s = 16$ kHz. Actors (5 men, 5 women) expressed 10 different texts (sentences) in 7 emotional categories: anger, boredom, disgust, fear, happiness, neutral and sad.

We applied Hamming window prior STFT on raw speech signal segmented into 256 long frames having 50% overlap of adjacent frames. In order to find eligible duration of speech signal to successfully recognize emotions, we trained and tested our network with 0.25s, 0.5s, 0.75s, 1s, 1.25s and 1.5s long segments of speech signal respectively. With 16kHz sampling frequency, 256 samples long window and 50% overlap, 1 feature frame represents 16ms long speech signal. As an input to our model we divided spectrograms into blocks with desired length over time using shift of 10 feature frames. Every sequence was labeled with utterance class, since data were labeled on utterance-level.

We used CNN with two convolutional layers. Each layer has squared kernels of size 5×5 and ReLu was used as an activation function at the output of convolution layers. After each convolutional layer, we used max-pooling layer with kernel size 2×2. After convolutional layers one fully-connected layer with 128 hidden neurons and ReLu activation function was used. It was followed by dropout layer with retention probability $p = 0.5$. At the output of the network there is an output layer with softmax activation function to provide probabilistic like classification.

Mini-batch was set up to size of 256 FV and the learning rate used in the training was 0.01. We used categorical cross-entropy as loss function and Adam algorithm for weight optimization. We trained model in respect to class (category) weights since the categories were unbalanced.

Performance of the architecture was evaluated with 10 fold cross-validation. We applied early stopping with respect to validation loss. 80% of data were used for training, remaining 20% for testing. We used 20% of training data as validation during training. Results of our experiments are shown in Tab. I. Note that the test accuracy is an average of weighted accuracies from 10 fold cross-validation with standard deviation in third column.

TABLE I. Comparison of different lengths of speech signal in SER

Analyzed sequence length [s]	*Test Accuracy [%]*	*Std. [%]*
0.25	59.34	±1.05
0.5	67.47	±1.24
0.75	72.14	±1.68
1	73.95	±0.65
1.25	76.48	±0.87
1.5	76.87	±1.24

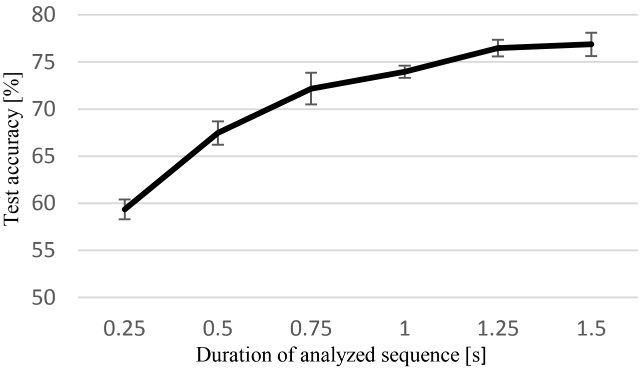

Figure 2. Relation between length of a sequence and the accuracy of the system. We can observe that accuracy of the recognition will stop increasing significantly over more than 1.25s long sequences.

Results show, that in order to recognize emotions from speech signal represented as spectrogram, it is appropriate to investigate 1.5s long interval of the speech signal. On average, 1.5s long analyzed speech signal was the best among tested lengths of sequences. We can observe that increasing the length of sequences had huge impact on performance of our model. Fig. 2 demonstrates the trend that model trained on longer sequences performed better than trained on shorter ones. Also, we can observe that there is only small increase in the accuracy between 1.25s and 1.5s long sequences.

TABLE II. DEPENDENCY BETWEEN LENGTH OF ANALYZED SEQUENCE AND THE SIZE OF DATASET

Length of sequence [s]	*Train*	*Test*	*Validation*	*Total*
0.25	9750	3048	2438	15236
0.5	8781	2745	2196	13722
0.75	7784	2433	1947	12164
1	6812	2130	1704	10646
1.25	5840	1826	1461	9127
1.5	4788	1497	1198	7483

With sequences longer than 1.5s, there can be a problem because some of utterances in database are not longer than 1.5s. It would cause reduction of dataset and we would not be able to correctly compare our results. It is worth to mention that the number of training and testing data decreased with the increased data dimension over time. It means that with greater time duration of a signal input we reduce a dataset. Tab. II shows dependency between size of dataset and the length of analyzed sequence. We can observe that there is more than twice as big dataset for sequences of the length of 0.25s than for sequences of length 1.5s.

The system was evaluated on sequence-level. Based on practical use, we can utilize utterance-level classification. We obtain several sequences from each utterance and evaluate them using our trained model. The output class (category) for the utterance will be the one that occurs most in the sequence-level classification (majority criteria). With this approach we can get better results in emotion recognition.

VI. CONCLUSIONS

Recognizing emotions from speech using Neural Networks gained huge attention in last few years. We performed time analysis of speech signal in a speech emotion recognition task using spectrogram as an input into Convolutional Neural Network. We analyzed performance of the proposed system using speech sequences of the length from 0.25s to 1.5s. The aim of this work was to find performance dependency of CNN using spectrograms as speech features and the time length of speech signal that is required for successful emotion recognize. It can be seen, that with increasing length of analyzed speech, the accuracy of system increases. The recorded difference in accuracy between 0.25s and 1.5s long sequences was more that 17%. On the other hand the increase of accuracy between 1.25s and 1.5s long sequences was less than 4%. In our future work we want to focus more on time and frequency analysis of speech signal and also on finding good set up of CNN architectures for SER.

ACKNOWLEDGMENT

The authors hereby declare that the article was founded by grants VEGA 1/0867/17 and VEGA 1/0800/16.

REFERENCES

[1] Fayek, H. M., et al., Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks (2017), http://dx.doi.org/10.1016/j.neunet.2017.02.013

[2] W. Lim, D. Jang and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, 2016, pp. 1-4.

[3] B. Schuller, G. Rigoll and M. Lang, "Hidden Markov model-based speech emotion recognition," Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on, 2003, pp. I-401-4 vol.1.

[4] P. P. Dahake, K. Shaw and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and Support Vector Machine," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, 2016, pp. 1080-1084.

[5] P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning," 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, 2017, pp. 137-140.

[6] L. Rabiner, B. Juan, Fundamentals of speech recognition, ISBN 0-13-015157-2, Prentice Hall PTR, New Yersey.

[7] X. Glorot, A. Bordes and Y. Bengio, "Deep Sparse Rectifier Neural Networks", In proceedings of AISTATS 14, pp .315-323, 2011.

[8] J. Schmidhuber, "Deep learning in neural networks: An overview", Neural Networks, vol. 61, pp. 85–117, 2015

[9] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, "A Convolutional Neural Network Cascade for Face Detection", In proceedings of Computer Vision and Pattern Recognition, pp. 5325-5334, 2015

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss: A Database of German Emotional Speech, Proc. Interspeech 2005