IF YOU ARE HAPPY AND YOU KNOW IT, YOUR SPEECH WILL SURELY SHOW IT: A CNN BASED SPEECH EMOTION DETECTOR

Tilak Purohit *tilak.purohit@epfl.ch*
Manon Boissat *manon.boissat@epfl.ch*

Supervisor: Dr. Mathew Magimai.-Doss
Hosting Laboratory: Idiap Research Institute, Martigny, Switzerland
Project–2, CS433- Machine Learning 2021, EPFL

ABSTRACT

In this work, we address the task of Speech Emotion Recognition (SER). Inspired by the success of end-to-end systems to model acoustic information, we make use of a CNN based end-to-end framework for the classification of four emotion classes namely - angry, happy, neutral and sad. The framework use raw-speech for the classification task hence eliminating the step of extracting hand-crafted features. Further, inline with this task we investigate and address three specific research questions related to- 1) the efficacy of the model, 2) the training setup and 3) modelling the network derived information. The study was conducted using the standard IEMOCAP corpus.

***Index Terms*—** Speech Emotion Recognition, Convolution Neural Network, Representation learning

1. INTRODUCTION

There exists rich literature in the field of psychology which investigates the role of acoustics in human emotions. Blanton [1], in his article, writes- "the effect of emotions upon the voice is recognized by all people. Even the most primitive can recognize the tones of love and fear and anger; and this knowledge is shared by the animals. Dogs, horses, and many other animals can understand the meaning of the human voice. The language of the tones is the oldest and most universal of all our means of communication." With the remarkable progress made in the field of automatic speech recognition(ASR), the time has come where machines should now learn to recognize human emotions efficiently via speech.

There are several factors which makes Speech Emotion Recognition (SER) a challenging task, for example:

1. Human emotions exhibit fuzzy temporal boundaries which makes it difficult to predict the onset or the completion of an emotion.
2. Each individual has a different way of expressing emotions and there could be more than one emotion in an utterance.
3. Spoken language (ex: French, German, English) plays a key role in detecting emotions from speech.

Even with multiple challenges, SER still remains a highly active area of research, due to its several applications, from affective computing to human-computer interfaces.

An important and critical aspect of SER is feature extraction, traditional SER research was mainly devoted in the search for 'best' speech features that could discriminate between different emotions on turn-level [2, 3]. Usually 'brute-force' methods were used to select the most indicative acoustic features which could help in discriminating between the various emotions. With the advent of representation-learning, breakthrough results were obtained in the field of Speech recognition and signal processing. Convolutional neural network (CNN, or ConvNet) in particular became very handy for learning representations from the data (signal) [4], which made it easier to extract useful information when building a classifier. In this work, we exploit the representation learning prowess of 1D-CNNs to build an emotion detector by modelling directly the raw speech signal.

This paper builds upon the previous works on end-to-end acoustic modelling, where a segment of raw waveform is first modeled by convolution layers followed by a hidden layer and output classification layer (e.g. see Figure 1). This approach was originally proposed in the context of speech recognition [5] and has subsequently been extended to other tasks such as, speaker verification [6], gender recognition [7], depression detection [8]. In all of these tasks, the neural network is trained to classify the output classes using cross entropy error criterion by taking a speech segment of about 250 ms as input and further the output frame level probabilities are aggregated at utterance level to make the final decision. This network has shown to deliver competitive results for the tasks mentioned above and without the need for relying on the handcrafted features. The network learns the ideal filters during the training phase so as to extract the task specific information from the raw speech signal, for the optimal classification results.

Inspired from the studies based on the raw-waveform CNN-based architecture, for this project we adapted this network for SER task and further investigate the following questions:

1. whether such an approach is feasible for SER. More precisely, can speech emotion recognition be effectively achieved by training the end-to-end acoustic model to classify emotion by modeling a short-segment of speech at frame-level and aggregating the output frame-level probabilities?
2. How does the network behave when trained in a subject-dependent manner and subject-independent manner?
3. Can the representations/embedding derived via CNN be further improved for the classification task by capturing and modelling the temporal dynamics information by computing approximate first order temporal derivative (denoted as delta) and approximate second order temporal derivative (denoted as delta-delta)?

We investigate these questions using the IEMOCAP - a benchmark corpus in the field of SER. The rest of the report is organised as follows. Section 2 covers in brief the proposed method. Section 3

Click here to visit our git repository

presents the short description of the database and the experimental setup used, Section 4 describes the systems we are proposing in detail, Section 5 deals with the results and analysis. Section 6 concludes the paper.

2. PROPOSED METHODOLOGY

Figure 1 illustrates the proposed end-to-end framework adopted for speech emotion recognition task. The input to the neural network is a raw-speech waveform of duration W_{Seq} (about 250 ms), which is processed by N convolution layers followed by a multilayer perceptron (MLP) to output speech emotion class conditional probabilities. Similar to conventional short-term spectral processing, the speech segment is shifted by 10 ms to estimate class conditional probabilities for the next frame and so on. During the training phase, the neural network is trained with frame-level cross entropy error criterion. During the recognition phase, speech emotion class conditional probabilities estimated for each frame are summed and normalized by the number of frames to estimate utterance-level speech emotion class conditional probabilities. The decision is then taken by selecting the class with maximum probability.

To validate that the neural network is indeed learning information from 250 ms of speech that is indicative of speech emotion, we also investigate an approach where, an utterance level representation is obtained from frame level neural embeddings by computing functionals $(Funct)$ such as, mean, standard deviation, skewness and kurtosis, we call this static-representation. Further to investigate if these frame-level neural embeddings could be enhanced by providing them with the temporal dynamics information, we computed the first and second order temporal derivative, delta(D) and delta-delta(DD) respectively on frame-level embeddings. We use these frame-level D & DD embedding, containing temporal information to compute utterance level representation using $Funct$ in a similar manner as for static representations, but calling these as delta and delta-delta representation. We then use an SVM classifier to compare which representation either static or the one with temoral knowledge performs better for our classification results.

Finally, the above methods were applied and evaluated for two scenarios, speaker-independent setup and speaker-dependent setup for SER task.

3. DATABASE AND PROTOCOLS

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [9] is a benchmark dataset used for emotion studies and in particular for speech emotion recognition tasks. It consists of recordings from 10 skilled actors (five female and five male) who recorded 12 hours of audio-visual data organized in 5 different sessions, among which some were scripted and some improvised. The recording comprises of audio, video and face motion capture samples. Each sample of recording is called an utterance. The audio recording is sampled with a 16kHz samping rate and duration is between 3 and 15 seconds. Each utterance has a corresponding emotion label. There are ten different emotion labels in the corpus: anger, sad, neutral, happiness, frustrated, excited, fearful, surprised, disgusted and other. The emotion label was chosen by majority vote among three annotators. Furthermore, to be consistent with the previous studies, we decided to work with four emotions categories: anger, sad, happy and neutral. The detailed description of emotions distribution is provided in table 1.

We conducted speaker-independent experiments following the leave-one-session-out methodology for training. For testing the 'k'-th session, the model was trained on the remaining four sessions. For the speaker-dependent experiments the complete data was pooled and shuffled then k-fold cross-validation methodology was used for training and testing, where k=5. Following the literature, the performance is measured in terms of unweighted average recall (UAR).

Table 1: The proportion of the four emotions among the selected data of IEMOCAP-dataset

Class	Number of utterances	Proportion (%)
Anger	1103	19.94
Happy	1636	29.58
Neutral	1708	30.88
Sad	1084	19.6

4. PROPOSED SYSTEMS

4.1. CNN based system

Table 2 presents the architecture of the raw waveform modeling CNNs for emotion recognition, illustrated in Figure 1. We did not optimize the architecture of the neural network for the speech emotion recognition task. Rather, we chose this architecture from the previous work on depression detection [8]. This neural networks have four convolution layers followed by one hidden layer MLP and output layer consisting of four emotion classes. The output layer had softmax activation, while all the layers had ReLU activation. This CNN network, has the kernel width set to 30 samples (about 1.8 ms) in the first convolution layer kW_1. This network is referred to as a Sub-segmental(denoted as subseg) model.

For training the neural network, we split the train portion in each fold into training subset and cross validation subset in 80:20 ratio. The network was trained using cross-entropy loss with stochastic gradient descent. The learning rate was halved, in the range 10^1 to 10^6, between successive epochs whenever the validation-loss stopped reducing. We used Keras [10] deep learning library with tensorflow backend.

Table 2: CNN architectures. nf: number of filters, kW: kernel width, dW: kernel shift, MP: max-pooling.

Model		**Layer**	**Conv**			**MP**
			nf	**kW**	**dW**	
RawCNN	subseg	1	128	30	10	2
		2	256	10	5	3
		3	512	4	2	-
		4	512	3	1	-

4.2. Neural-embedding based systems

As mentioned in Section 2, apart from the CNN-based system, where the frame-level speech emotion class probabilities are averaged and the class with maximum probability is chosen as the output. We also setup systems where the neural embeddings derived from the subseg network were modelled to form a utterance level embedding using functionals ($Funct_{(.)}$), the different functionals we computed were: mean (m), standard deviation (sd), skewness (sk) and kurtosis (k). In first setup we directly use the 10 dimentional neural

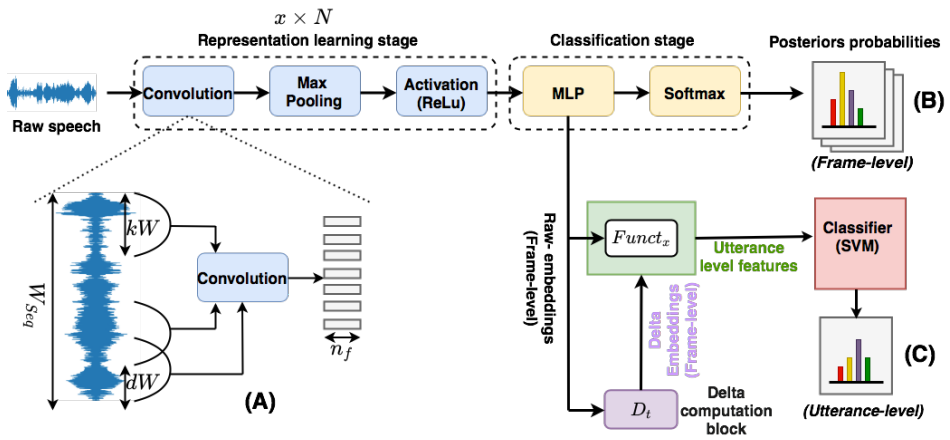


Fig. 1: Illustration of the proposed speech emotion recognition method. (A)illustrates the processing in the first convolution layer. kW denotes kernel width, dW denotes kernel shift and n_f denotes number of convolution filters; (B) denotes the approach of aggregating frame-level probabilities for speech emotion classification; and (C) denotes the approach of speech emotion classification using utterance level representations obtained from frame level neural embeddings.

embeddings (S-embeddings) to create its utterance-level representation using $Funct_{(.)}$ whereas, in the second setup we modify the S-embeddings by computing its approximate first order and the second order derivatives, D-embeddings and DD-embeddings respectively and then create utterance-level representation.

Delta and Delta-delta features: A common method for extracting information about transition dynamics is to determine the first difference of signal features, known as the delta of a feature and the second difference in known as delta-delta of a feature. A trivial observation/interpretation of the delta and delta-delta features is that they approximate first and second derivatives of the signal. For speech recognition tasks where the features are spectrum based, example: mel-frequency cepstral coefficients (MFCCs), the delta and delta-delta features computed on spectrum provides better information than the static features [11]. One of the question we put forward was to probe if adding transition dynamics information to the static neural-embedding would help in classifying the emotions from speech better. We implemented Eq 1, to compute the delta coefficient based on the HTK-toolkit [12] regression formulation .

$$D_t^d = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta}^d - c_{t-\theta}^d)}{2\sum_{\theta=1}^{\Theta} \theta^2} \; \forall d \in \{1, \cdots 10\} \; \forall t \in \{1, \cdots N\} \tag{1}$$

where:

- D_t^d is the delta coefficient calculated at time from t of dimension embedding dimension d ($N \times 10$)
- Θ is the window considered for the delta calculation. θ goes from 1 to Θ.
- c_t^d corresponds to the frame-level embedding d time frameo t. Where N are the total frames in an utterance.

For our study we set $\Theta = 2$ that is the temporal information of 2 preceding and 2 successive frames were considered for the delta computation.

After the computation of 10 dimensional frame level embeddings-S-embedding, D-embeddings and DD-embeddings, we converted them to utterance based representations by using $Funct_{m,sd,sk,k}$ this representation was of dimension 40, 10 corresponding to each each functional namely, mean, standard deviation, skewness and kurtosis. We used these utterance level representation for the classification task, SVM with linear kernel was used as a classifier, sklearn [13] implementation was used for SVM. Apart from the individual representations (static and delta) used for classification, and for better understanding of the features, we augmented the feature space by concatenating the different representations, for example we combined static representation with delta representations we also combined static, delta and delta-delta representation for one experiment. This complete setup with combination experiments were repeated for speaker-dependent and speaker-independent settings as mentioned in Section 3.

Table 3: Performance of different systems measured in terms of UAR.

IEMOCAP

Systems	Classifier	UAR
Proposed systems - Speaker Dependent(SD)		
Raw-CNN	Softmax	66.7
$Funct_{m,sd,sk,k}$(S-EMBEDDINGS)	SVM	67.3
$Funct_{m,sd,sk,k}$(D-EMBEDDINGS)	SVM	61.7
$Funct_{m,sd,sk,k}$((S+D)-EMBEDDINGS)	SVM	66.9
$Funct_{m,sd,sk,k}$((S+D+DD)-EMBEDDINGS)	SVM	66.2
Proposed systems - Speaker Independent(SIn)		
Raw-CNN	Softmax	57.4
$Funct_{m,sd,sk,k}$(S-EMBEDDINGS)	SVM	56.7
$Funct_{m,sd,sk,k}$(D-EMBEDDINGS)	SVM	54.9
$Funct_{m,sd,sk,k}$((S+D)-EMBEDDINGS)	SVM	56.6
$Funct_{m,sd,sk,k}$((S+D+DD)-EMBEDDINGS)	SVM	55.5

5. RESULTS AND DISCUSSION

It can be observed from the results reported in table 3, that the raw-speech CNN-based end-to-end framework proposed here, has the potential for the SER task. It is worth mentioning that the results reported are comparable and on par with the frameworks which uses handcrafted/knowledge based features [14–16].

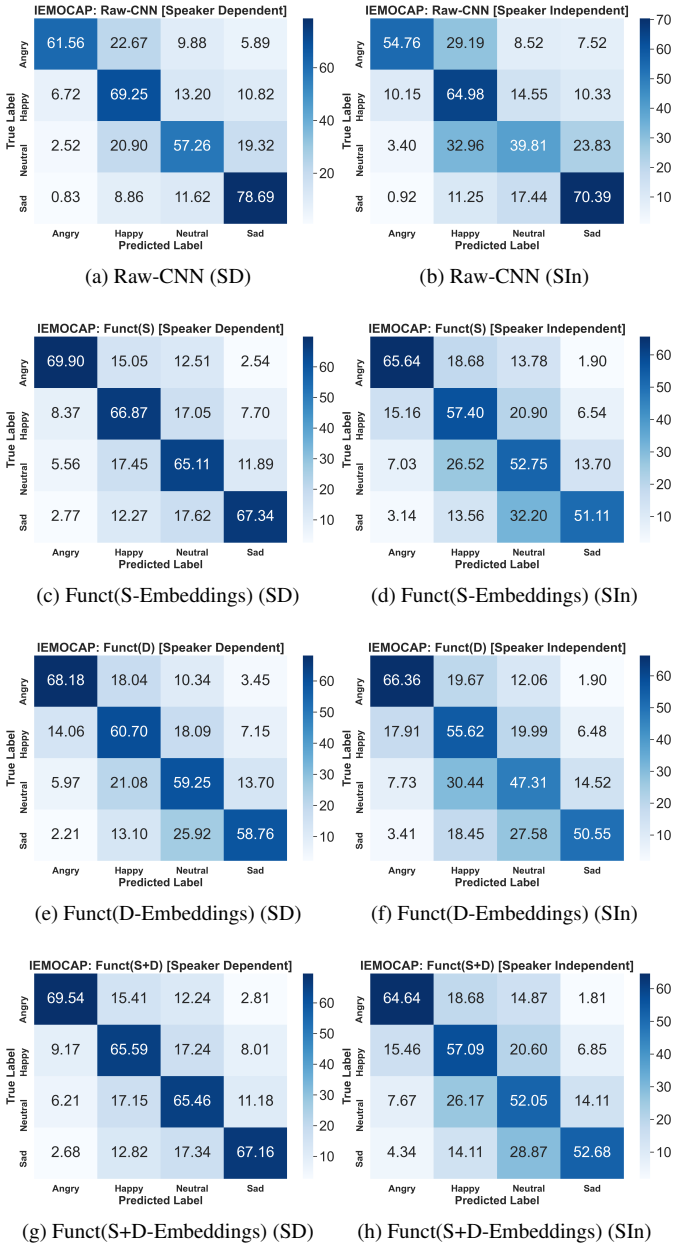
We can see the speaker dependent systems performing better than the systems trained on speaker independent setup. This may not

be very surprising since it is know in the literature that speaker information present during training help attain better results. In our case the difference in speaker dependent setup as compared to speaker independent is of approx. 10% gain for every experimental study. From the confusion matrix shown in fig 2 we can observe a gain in emotion class recognition for all the labels but the neutral class is the one which gains the most with the speaker dependent settings for all the experiments. Also, the framework is very capable in distinguishing the emotions like happy and sad which can be observed from fig 2.

Contrary to our hypothesis, we do not see any improvements by using delta representations, this tells that the static neural embeddings are robust enough for the SER task. But the delta representation do not perform too inferior on its own, and when concatenated with static embedding it gives similar results like the static embedding on its own, which suggests that deltas do not provide any complementary information to the static features for better classification. Even the second derivative the delta-delta representation is not much useful for the SER task.

6. CONCLUSION

We studied speech emotion recognition task using IEMOCAP corpus, and asked three specific research questions inline with this task. From the research and experiments conducted, we were successfully able to answer all the three questions. We showed that our end-to-end raw speech framework is capable of carrying out the SER task and is comparable to the networks using handcrafted/knowledge based features. We further demonstrated empirical results showing speaker-dependent and speaker-independent settings do have an affect on the performance of the system. Lastly, our attempt to enhance the neural embedding by providing temporal dynamic information for improving the performance did not gave better results which was contrary to our hypothesis. For future work we wish to optimize the network settings and for the embedding based systems use different classifier like random forest and experiments with different kernels of SVM. We would also like to extend this study to different corpora.



(a) Raw-CNN (SD) (b) Raw-CNN (SIn)

(c) Funct(S-Embeddings) (SD) (d) Funct(S-Embeddings) (SIn)

(e) Funct(D-Embeddings) (SD) (f) Funct(D-Embeddings) (SIn)

(g) Funct(S+D-Embeddings) (SD) (h) Funct(S+D-Embeddings) (SIn)

Fig. 2: Confusion matrices of eight selected systems: (a), (b) are based on frame level classification, where as (c) to (h) are utterance based; (c), (d): static embeddings ; (e), (f): detla embeddings ; (g), (h): concatenation of static and delta embeddings; left column subplots are speaker dependent (SD) experiments similarly right column are for speaker independent (SIn)

7. REFERENCES

[1] Smiley Blanton, "The voice and the emotions," *Quarterly Journal of Speech*, vol. 1, no. 2, pp. 154–172, 1915.

[2] Björn Schuller et al., "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *Proc. of Interspeech*, 2007.

[3] Shashidhar G Koolagudi and K Sreenivasa Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," 2014.

[5] D. Palaz, R. Collobert, and M. Magimai.-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proc. of Interspeech*, 2013.

[6] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, "Towards directly modeling raw speech signal for speaker verification using cnns," in *Proc. of ICASSP*, 2018.

[7] S. H. Kabil, H. Muckenhirn, and M. Magimai-Doss, "On learning to identify genders from raw speech signal using cnns.," in *Proc. of Interspeech*, 2018, pp. 287–291.

[8] S. P. Dubagunta, B. Vlasenko, and M. Magimai.-Doss, "Learning voice source related information for depression detection," in *Proceedings of ICASSP*, 2019.

[9] C. Busso and other, "IEMOCAP: Interactive emotional dyadic

motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[10] Francois Chollet et al., "Keras," 2015.

[11] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[12] Steve Young et al., "The htk book," *Cambridge university engineering department*, vol. 3, no. 175, pp. 12, 2002.

[13] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] Shahin Amiriparian et al., "On the impact of word error rate on acoustic-linguistic speech emotion recognition: An update for the deep learning era," 2021.

[15] Serdar Yildirim, Yasin Kaya, and Fatih Kılıç, "A modified feature selection method based on metaheuristic algorithms for speech emotion recognition," *Applied Acoustics*, vol. 173, pp. 107721, 2021.

[16] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227–2231.