Interspeech 2018
2-6 September 2018, Hyderabad, India

Towards Temporal Modelling of Categorical Speech Emotion Recognition

Wenjing Han[1], *Huabin Ruan*[2], *Xiaomin Chen*[3], *Zhixiang Wang*[1], *Haifeng Li*[3], *Björn Schuller*[3,4,5]

[1]Samsung Research Institute China Beijing (SRC-B), Beijing, China
[2]Protein Research Technology Center, Tsinghua University, Beijing, China
[3]School of Computing Science and Technology, Harbin Institute of Technology, Harbin, China
[4]GLAM - Group on Language, Audio & Music, Imperial College London, U.K.
[5]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

wenjing.han@samsung.com

Abstract

To model the categorical speech emotion recognition task in a temporal manner, the first challenge arising is how to transfer the categorical label for each utterance into a label sequence. To settle this, we make a hypothesis that an utterance is consisting of emotional and non-emotional segments, and these non-emotional segments correspond to silent regions, short pauses, transitions between phonemes, unvoiced phonemes, etc. With this hypothesis, we propose to treat an utterance's label sequence as a chain of two states: the emotional state denoting the emotional frame and *Null* denoting the non-emotional frame. Then, we exploit a recurrent neural network based connectionist temporal classification model to automatically label and align an utterance's emotional segments with emotional labels, while non-emotional segments with *Null*s. Experimental results on the IEMOCAP corpus validate our hypothesis and also demonstrate the effectiveness of our proposed method compared to the state-of-the-art algorithms.

Index Terms: speech emotion recognition, recurrent neural network, connectionist temporal classification, sequence-to-sequence

1. Introduction

In this paper, we consider the task of categorical speech emotion recognition (SER). As well known, the affective information conveyed by speech is inherently sequential, and thus architectures that explicitly perform temporal modelling, such as hidden Markov models (HMMs) [1, 2, 3] and recurrent neural networks (RNNs), e.g., with long short-term memory (LSTMs) [4, 5], would better exploit this information. Nowadays, there is a growing trend to apply LSTMs to model the categorical SER tasks [6, 7, 8, 9, 10, 11, 12, 13]. However, to do this, the first issue coming up to challenge is the mismatch between the short-term inputs at the frame level and the long-term outputs at the utterance level: On the one hand, acoustic features normally are extracted from short frames of typically 20 to 60 msec [14]; on the other hand, emotional labels are often provided for the whole utterances in many databases [15, 16]. This results in that many categorical SER-specific LSTMs are actually constructed following a *sequence-to-label* recipe with minor variations, where LSTMs can be interpreted as the utterance representation learners then followed by classification modules.

Here, we broadly categorize these previously adopted *sequence-to-label* LSTMs into three groups according to their representation learning strategies: 1) *final-pooling* LSTMs, which only pick the final hidden representation at the last frame of an utterance as the representation [6, 7, 8, 9, 13]; 2) *mean-pooling* LSTMs, which calculate the average of hidden representations of all inner frames of an utterance as the representation [7, 8, 9, 10, 11]; 3) *weighted-pooling* LSTMs, which compute a weighted sum as the representation, where the weights are normally determined based on an additional attention mechanism [8, 9, 10, 11, 17, 18]. It has been consistently demonstrated in previous works that *weighted-pooling* LSTMs outperform the other two groups of LSTMs and can achieve state-of-the-art performance. It should be, however, noted that all above mentioned LSTMs share a common underlying assumption that these pooling based utterance-level representations have the ability to capture variations from frame-level feature sequences and thus contain emotional content, even though these pooling operations lose much temporal information from successive frames inevitably [19].

Considering the temporal information is extremely important to SER tasks; one may naturally expect to utilize the inherently sequential attribution of LSTM to capture as much temporal information as possible. In an attempt to this, this work aims to model the categorical SER problem temporally by means of performing the LSTM under a completely *sequence-to-sequence* recipe. The most naïve approach is to assign the overall emotional label to each frame within the utterance, and train the LSTM in a *frame-wise* manner by back-propagating cross-entropy errors from every frame [9]. However, it is unreasonable to assume that every frame within an utterance represents the overall emotion. The most ideal approach should be that one could evade those emotionally irrelevant frames in an utterance and, meanwhile, align the overall emotional label to each emotionally relevant one. But how? Different from a spatial signal like images, speech is a kind of temporal signal. This implies that one cannot achieve acoustically understandable information at each time-distributed sampling point, unless one integrates enough sampling points one by one over time to form a speech segment. Moreover, the human auditory mechanism also determines that it takes longer for us to understand emotional content than it takes to understand linguistic content underlying speech [15]. All these suggest that it is almost impossible for a human to meaningfully annotate emotional labels at the frame level.

In order to investigate the role of a *sequence-to-sequence* fashion LSTM in learning temporal information from input utterances and meet the aim to automatically align emotional labels to emotionally relevant frames, we propose an approach with two phases. In the first phase, we make a hypothesis based on phonetic knowledge to help us point out the numbers of emotional and non-emotional sub-segments within an utterance. The hypothesis can be briefly described as that there are emotional frames contained in voiced phonemes, and non-emotional frames existing in unvoiced phonemes and between two phonemes. With this

10.21437/Interspeech.2018-1858

hypothesis, we re-define the overall label of an utterance to a specific label sequence consisting of a certain number of emotional and non-emotional states. In the second phase, we look for inspiration from connectionist temporal classification (CTC) models [20], which have been demonstrated to be effective in several temporal modelling tasks like handwriting recognition [21] and end-to-end automatic speech recognition (ASR) [22]. The core advantage of the CTC model is to remove the need for manually aligning labels to sub-segments of training samples, which is exactly what we desire. With the benefit of the CTC model, we are able to align those pre-set emotional and non-emotional states in the former phase to exact frames in an utterance automatically.

In this paper, we report a detailed and formal description of the proposed thorough *sequence-to-sequence* modelling solution for SER. To verify the effectiveness, we conduct experiments on the IEMOCAP corpus [23] by comparing various approaches, including a *final-pooling* LSTM, a *mean-pooling* LSTM, a *frame-wise* LSTM, the approach proposed by Lee *et al.* in [12], and our proposed approach. During the experiments, we seek to perform the re-defined label sequence with different implementations to support our hypothesis.

2. Related Works

An early use of LSTM in SER was conducted by Wöllmer *et al.* [24]. They, however, applied LSTM to predict the continuously labelled emotions over time in the valence-activation space rather than the categorical emotional labels used in our work. Late, Trigeorgis *et al.* [25] devised an end-to-end deep network that worked on raw time-domain signals, which appended a LSTM layer after several CNN layers to predict continuous, spontaneous and natural emotions.

In addition to the continuous SER, there were also works exploiting LSTMs in the field of categorical SER. Lee *et al.* [12] once made an attempt to use a CTC-style RNN underneath to encode the temporal information into the frame representations, despite they never declared the used RNN as a CTC based model explicitly. In essence, a LSTM model with CTC loss function transformed the sequence of frames in an utterance into a sequence of high-level representations. Then certain statistical functionals of these sub-utterance representations formed the utterance representation, and the classification task was carried out by a following extreme learning machine (ELM) [26]. Later, Chernykn *et al.* [13] reproduced and verified the effectiveness of Lee *et al.*'s method. Perhaps our presented work is most similar to Lee *et al.*'s, but there are at least two fundamental differences we would like to underscore. First, they restricted the non-emotional parts within an utterance only to the silent regions and neglected the other unvoiced sub-segments including short pauses or fricative phonemes, which we all take into account in this work. Second, they focused on learning the frame-level representations and constructed the utterance representation via statistical functionals, whereas we aim at decoding the utterance's emotion label directly from the LSTM's outputs. There is no extra classification module is needed in this work.

The goal of this work is to achieve higher temporal resolution of speech via the proposed method thus to achieve more accurate classification. To chase the same goal, some other works made attempts from different perspectives. Schuller *et al.* [27] automatically segmented speech-turns into chunks, and then mapped the information from the chunk level on the turn level by multi-instance learning. They reported better results on chunk level comparing to that on syllable and turn level. Another group of researchers, including Chao *et al.* [7, 8], Mirsamadi *et al.* [9], and Huang *et al.* [10, 11], leveraged the attention mechanism to evaluate the emotional level of each frame within the utterance. They consistently reported state-of-the-art performance on the categorical SER tasks. Parthasarathy *et al.* [28] worked on detecting the so-called emotional hotspots within the interaction based on the qualitative agreement method, however, they restricted the task to the continuous SER field.

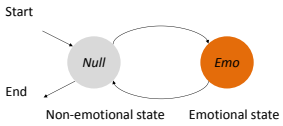

Figure 1: *Markov chain for label sequence of each utterance.*

3. LSTM-CTC based SER Modelling

In conventional algorithms for categorical SER, formally, a training sample can be represented as $\{\mathbf{x}, y\}$, where $\mathbf{x} = (x_1, x_2, ..., x_T) \in \mathcal{X}$ is an utterance with T frames and y is its corresponding categorical label, i.e., *Angry, Happy, Neutral, Sad* in our used database (cf., Section 4.1). However, since the label y is annotated for the whole utterance, it does not mean all frames in $\mathbf{x}$ should be mapped to y. From this view, a strategy that can selectively map $\mathbf{x}$'s emotionally relevant frames to ys and, meanwhile map other emotional irrelevant frames to non-emotional labels, is our goal.

3.1. The Hypothesis and Label Sequence Construction

To approach the above goal, in this section, we first define an extra label *Null* to denote the non-emotional state beyond the emotional ones, i.e., {*Angry, Happy, Neutral, Sad*}. Then, we propose to extend the overall label y to a sequence $\mathbf{y}^{seg} = (Null, y, Null, ..., y, Null)$ ($|\mathbf{y}^{seg}| \leq T$) under the guidance of a hypothesis. One should note that, here, the elements *Null*s and ys in sequence $\mathbf{y}^{seg}$ are all at the segment (i.e., a set of frames) level rather than the frame level. This is because it is rarely possible for us to specify a label at frame level hand-craftedly.

From the view of phonetics, each utterance can be regarded as a sequence of phonemes. So, when one pronounces an utterance, one pronounces the involved phonemes one by one in fact. In phonetics, it is believed that when one pronounces two neighbouring phonemes, there often exists joint frames that can be a very short pause belonging to neither phoneme or can be a transition belonging to either phoneme [29]. If these joint frames are emotional or not? Additionally, affricate, aspirate or fricative phonemes are unvoiced in the duration of the pronunciation (i.e., vocal folds do not vibrate). We wonder if their corresponding frames are emotional or non-emotional as well.

Given the above, we make the following hypothesis: There are non-emotional frames existing between two phonemes (e.g., frames involved in silent regions, short pauses, or transitions) or even in the unvoiced phonemes themselves (e.g., frames involved in affricate, aspirate and fricative phonemes). According to this hypothesis, we can depict $\mathbf{y}^{seg}$ as a sequence of states intuitively following a Markov chain shown in Fig. 1. This means that the proposed label sequence for each utterance starts from a non-emotional segment, and goes through emotional segments and non-emotional segments in alternation, and finally ends at a non-emotional segment. In particular, silent regions, short pauses, transitions, and unvoiced phonemes in the utterance correspond to *Null* states in $\mathbf{y}^{seg}$, and voiced phonemes correspond to y states in $\mathbf{y}^{seg}$.

However, our final goal is to further determine which frames

of $\mathbf{x}$ should be annotated as states (i.e., labels) ys or *Null*s. Therefore, the above operation of extending an utterance $\mathbf{x}$'s label from y to $\mathbf{y}^{seg}$ is just the first step, which we formally denote as 1) $y \rightarrow \mathbf{y}^{seg}$. Since $\mathbf{y}^{seg}$ passes by y and *Null* alternatively, we can further remove all $Null$s from $\mathbf{y}^{seg}$ for simplification. Then, $\mathbf{y}^{seg}$ becomes $\mathbf{y}^{seg} = (y, y, ..., y)$ in what follows. Given so, we only need to assign the number of ys contained in $\mathbf{y}^{seg}$ as a certain value, i.e., the number of voiced phonemes in an utterance under the guidance of our hypothesis. To do this, we leverage the publicly released text annotations within the used IEMOCAP corpus and the popular CMU pronouncing dictionary [30] to generate the phoneme sequence for every utterance, and count the number of voiced phonemes in each one. In case there is no available text annotations provided for the used corpus, an extra ASR module can help to generate the phoneme sequences. Now, for an utterance $\mathbf{x}$, we are able to specify its $\mathbf{y}^{seg}$. However, we still can not recognize the exact frame-boundaries between two segments. So here comes the second step: 2) $\mathbf{y}^{seg} \rightarrow \mathbf{y}^{frame}$, to assign each frame belonging to a certain segment with the corresponding label. In this step, we propose to apply a LSTM-based CTC model (LSTM-CTC) to align $\mathbf{y}^{seg}$ to $\mathbf{y}^{frame}$ automatically.

3.2. LSTM-CTC based Temporal Modelling

LSTMs [4, 5] have been widely used for modelling time series. However, it has not been possible to apply LSTMs directly to sequence labelling. The problem is that the standard neural network objective functions are defined separately for each point in the training sequence; in other words, LSTMs can only be trained to make a series of independent label classifications. This means that the training data must be pre-segmented, and that the network outputs must be post-processed to give the final label sequence. Fortunately, the LSTM-CTC proposed in [20] has been proven to be able to remove the need for pre-segmented training data and post-processed output, and model all aspects of the sequence within a single network architecture. That is exactly what we want for modelling a *sequence-to-sequence* SER. Its basic idea is to interpret the LSTM outputs as probability distribution over all possible label sequences, conditioned on a given input sequence.

In what follows, we formally describe the LSTM-CTC based SER modelling. Given an input utterance $\mathbf{x} = (x_1, x_2, ..., x_T)$, a LSTM neural network can be denoted as a map $\mathcal{N}_{\mathcal{W}} : (\mathbb{R}^m)^T \mapsto (\mathbb{R}^n)^T$, where m and n are the numbers of input and hidden units, $\mathcal{W}$ is the hyper-parameter, and T is the frame number of $\mathbf{x}$. Then, the corresponding hidden layer output $\mathbf{h} = (h_1, h_2, \ldots, h_T), h_k \in R^n, k = 1, 2, \ldots, T$ is mapped as:

$$\mathbf{h} = \mathcal{N}_{\mathcal{W}}(\mathbf{x}). \tag{1}$$

Next, we introduce the CTC strategy [20, 31] to update the parameters of the above LSTM, and refer to this LSTM as LSTM-CTC. LSTM-CTC has a softmax output layer with one more unit than the number of labels in L, where L is the set of labels. In our case, $L = \{Angry, Happy, Neutral, Sad\}$. The activations of the first $|L|$ units are interpreted as the probabilities of observing the corresponding emotional labels at particular times. The activation of the extra unit is the probability of observing a non-emotional label *Null* or no label.

More formally, let $s_t = (s_t^1, s_t^2, ..., s_t^{|L|}, s_t^{|L|+1})$ be the sequence of the softmax layer's outputs at time t:

$$s_t = softmax(W_s \cdot h_t + b_s), \tag{2}$$

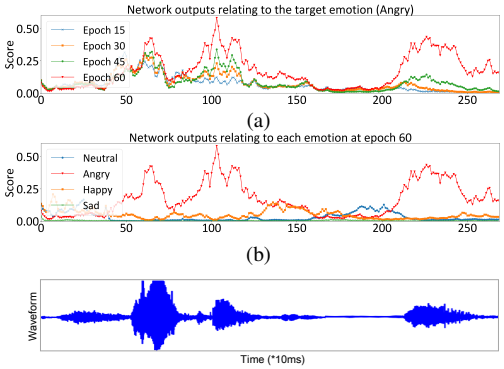

Figure 2: *Network outputs of a training example. (a): the outputs relating to the target emotion during the training stage; (b): the outputs at epoch 60 relating to four emotions; (c): the raw waveform.*

where W_s is the weight matrix from the hidden layer to the softmax layer, b_s is the bias, and h_t is the output of the hidden layer at time t. $s_t^m (m = 1, 2, ..., |L| + 1)$ denotes the activation of the output unit m at time t. Then, s_t^m is interpreted as the probability of observing label m at time t, which defines a distribution over the set L'^T of the sequences with length T, where $L' = L \bigcup \{Null\}$:

$$p(\pi|\mathbf{x}) = \prod_{t=1}^{T} s_t^{\pi_t}, \forall \pi \in L'^T. \tag{3}$$

The next step is to define a many-to-one map $\mathcal{B}$: $L'^T \mapsto L^{\leq T}$, $L^{\leq T}$ is the set of possible labelling (i.e., the set of sequences of length less than or equal to T over the original label set L). We do this by simply removing all *Null*s and repeated labels from any π in L'^T (e.g., $\mathcal{B}(Null\ Null\ Angry\ Null\ Angry\ Angry\ Null\ Null) = \mathcal{B}(Null\ Angry\ Angry\ Null\ Null\ Null\ Angry\ Null) = (Angry\ Angry)$). Finally, we use $\mathcal{B}$ to define the conditional probability of a given labelling $\mathbf{l} \in L^{\leq T}$ as the sum of the probabilities of all the πs corresponding to it:

$$p(\mathbf{l}|\mathbf{x}) = \sum_{\pi \in \mathcal{B}^{-1}(1)} p(\pi|\mathbf{x}). \tag{4}$$

Moving to our case, given training set $\mathcal{S}$ with training samples $\{\mathbf{x}, \mathbf{y}^{seg}\}$s, the total probability of any label sequence $\mathbf{y}^{seg}$ can then be calculated by summing the probabilities of its different alignments in $\mathbf{y}^{frame}$s. We can train the network by maximizing the log probability of all $\{\mathbf{x}, \mathbf{y}^{seg}\}$s in $\mathcal{S}$:

$$\mathcal{J} = \sum_{(\mathbf{x}, \mathbf{y}^{seg}) \in \mathcal{S}} -ln(p(\mathbf{y}^{seg}|\mathbf{x})). \tag{5}$$

For more detailed information regarding the training algorithm, please refer to [20]. Figure 2 illustrates the changing of the outputs relating to the target emotion during the training stage (Figure 2(a)), the outputs at epoch 60 relating to four emotions (Figure 2(b)), together with the corresponding waveform of the example (Figure 2(c)). If we interpret the output value as how emotional every frame has been decided to be, we can see from Figure 2(a) that the silence frames within the signal are automatically aligned as very low-level emotional. And it is worth noting that the speech frames with higher energy do not necessarily lead to higher outputs, which suggests our proposed method does not focus on energy only, and it is capable of considering the emotional content of different portions of speech.

Table 1: *Low-level descriptors (LLDs) for each frame. (Δ and $\Delta\Delta$ represent the first and second order differences.)*

Energy and Spectral related (211)	
PLPCC(5)+Δ+$\Delta\Delta$	MFCC-sma(15)+Δ+$\Delta\Delta$
LogEnergy(1)+Δ+$\Delta\Delta$	SpectralRollOff(4)+Δ
LSPFrequency(8)+Δ	Chroma(12)
LengthL1norm(2)+Δ	LPCCoeff(11)
LogRelF0(2)+Δ	Amplitude(3)+Δ
SpectralEnergy(2)+Δ	SpectralSlope(2)+Δ
ZCR(1)+Δ	Loudness(1)+Δ
RASTA-filtered(26)+Δ	RMSEnergy(1)+Δ
Spectral(Flux,Centroid,Entropy,	Hammarberg(1)+Δ
Variance,Skewness,Kurtosis,	LPGain(1)
Harmonicity,Flatness)(8)+Δ	AlphaRatio(1)+Δ
Voicing related (27)	
F0(2)+Δ LogHNR(1)+Δ	JitterLocal(1)+Δ
FormantFreqLPC(6)	ShimmerLocal(1)+Δ
FormantBandwidthLPC(6)	FormantFrameIntensity(1)
VoicingfinalUnclipped(1)+Δ	JitterDDP(1)+Δ

Moreover, as shown in the Figure 2(b), after several iterations, when the network becomes stable, it can learn the characteristics of the target emotion well.

4. Experimental Evaluation

4.1. Database and Features

We run experiments on the publicly available and highly popular IEMOCAP database [23] with speaker-independent 10-fold cross validation. It contains about 12 hours of audio-video data organized in 5 sessions, in each of which 1 actress and 1 actor are involved in. We leverage audio data from 5 emotional categories *Angry*, *Excited*, *Happy*, *Neutral*, *Sad*, and further merge the data of *Excited* and *Happy* to form a new *Happy* category as several related works did [32, 33]. Finally, the leveraged data contains 5,531 utterances totally (i.e., 1,103 for *Angry*, 1,636 for *Happy*, 1,708 for *Neutral*, 1,084 for *Sad*). We extract 238 low-level descriptors (LLDs) as in [34] at frame level considering the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [14] and the INTERSPEECH Challenges [35] feature sets, by openSMILE [36]. Table 1 gives the details.

4.2. Setup and Results

As a baseline SER system, we use a feedforward deep neutral network (DNN) containing 4 hidden layers each of which contains 256 units, and use the 88 dimensional statistical GeMAPS features [14] as input. In the LSTM-CTC based system, the 238 dimensional LLDs are directly used for input. The network contains 1 hidden layer with 256 bidirectional LSTM cells (128 forward nodes and 128 backward nodes), and 1 output softmax layer with 5 units corresponding to 4 emotional labels {*Angry*, *Happy*, *Neutral*, *Sad*} along with 1 non-emotional label {*Null*}, respectively. The final emotional label of an utterance is decoded as the one with highest frequency in the output sequence. Additionally, for comparison, we implement three other LSTM based SER systems, namely a *frame-wise* system, a *final-pooling* system, and a *mean-pooling* system. All of them adopt the same structured LSTM networks as the LSTM-CTC network above, except that the sizes of the output softmax layers are 4 and the loss functions are cross-entropy based. Moreover, we also re-implement the methods presented in [12] but on our used data partition. Further, in the LSTM-CTC based system, to evaluate whether our hypothesis is in fact supported, we implement the $y \rightarrow \mathbf{y}^{seg}$ by considering assigning the number of emotional labels $\mathbf{y}$s in $\mathbf{y}^{seg}$ as 1) the number of words in $\mathbf{x}$, 2) the number of voiced phonemes in $\mathbf{x}$ following the hypothesis (cf., Section 3.1), and 3) the double number of voiced phonemes in $\mathbf{x}$. Both the unweighted average recall (UAR) and weighted average recall (WAR) are used for performance evaluation.

Table 2: *Comparison of UARs and WARs on IEMOCAP using different methods. (WN: the number of words in an utterance, PN: the number of voiced phonemes in an utterance, and ELM: extreme learning machine.)*

Methods for Comparison	**UAR [%]**	**WAR [%]**
DNN	62.4	61.6
frame-wise LSTM	63.2	61.1
final-pooling LSTM	53.0	53.1
mean-pooling LSTM	63.8	62.5
LSTM in [12]	59.7	57.4
LSTM + ELM in [12]	63.5	62.4
Proposed Method		
LSTM-CTC, $\lvert\mathbf{y}^{seg}\rvert$ = WN	63.7	62.9
LSTM-CTC, $\lvert\mathbf{y}^{seg}\rvert$ = PN	**65.7**	**64.2**
LSTM-CTC, $\lvert\mathbf{y}^{seg}\rvert$ = PN*2	65.0	63.1

Table 2 summarizes the performance comparison between different methods. As it can been seen, the best UAR 65.7% and WAR 64.2% are reached by our proposed LSTM-CTC based method, particularly, while we assume the utterance $\mathbf{x}$ consists of PN (abbr. of the voiced phoneme number) emotional segments, i.e., $\lvert\mathbf{y}^{seg}\rvert$ = PN. This implementation outperforms other LSTM-CTC based ones where $\lvert\mathbf{y}^{seg}\rvert$ = WN (abbr. of the word number) and $\lvert\mathbf{y}^{seg}\rvert$ = PN*2, which indicates that the best performance is achieved when splitting an utterance into emotional segments at phoneme level whereas shorter or longer segmenting provide lower accuracies. This supports our hypothesis consequently. Moreover, our proposed method surpasses the baseline DNN absolutely by 3.3% on UAR and 2.6% on WAR, and significantly outperforms other *sequence-to-label* LSTMs at well (t-test, $p < 0.05$). As shown, the *mean-pooling* LSTM achieves relatively higher accuracies than the *frame-wise* and *final-pooling* LSTMs, probably because more emotionally relevant information is captured by the global mean pooling. We also investigate the comparison between the proposed method and [12], which implemented a different CTC-style system. As observed from the table, our method achieves significantly higher performance than it (t-test, $p < 0.05$) and even outperforms the combination of LSTM and ELM further proposed in [12].

5. Conclusions

We present an effective LSTM-CTC based modelling approach toward the categorical SER task, which appears as a more thorough temporal recipe compared to other state-of-the-art works. After making a hypothesis to guide the extension from an utterance label to a label sequence, a LSTM model with CTC loss function is trained to align emotionally relevant segments within the utterance with emotional labels. In the prediction stage, the utterance label is decoded as the most frequent label in the output sequence. Within the whole approach, the only one operation left to manual work is to specify how many emotional segments are contained in an utterance under the guidance of the proposed hypothesis. Experimental results suggest that while we specify each utterance consists of PN (abbr. of the voiced phoneme number) emotional segments, the proposed approach achieves the state-of-the-art performance on the IEMOCAP corpus with a UAR of 65.7% and a WAR of 64.2%. Future work involves the exploration of better strategies to extend utterance labels to label sequences and more effective *sequence-to-sequence* algorithms for categorical SER-specific modelling.

6. References

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Readings in Speech Recognition*, vol. 77, no. 2, pp. 267–296, 1989.

[2] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. of IEEE International Conference on Multimedia and Expo*, San Jose, California, USA, 2003, pp. I–401–4 vol.1.

[3] L. Li, Y. Zhao, D. Jiang, and Y. Zhang, "Hybrid deep neural network–hidden Markov model (DNN-HMM) based speech emotion recognition," in *Proc. of Affective Computing and Intelligent Interaction*, Geneva, Switzerland, 2013, pp. 312–317.

[4] A. Graves, *Long Short Term Memory*. Springer Berlin Heidelberg, 2012.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, vol. 4, pp. 3104–3112, 2014.

[6] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Proc. of International Conference on Affective Computing and Intelligent Interaction*, San Antonio, Texas, USA, 2017, pp. 190–195.

[7] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based encoding method for emotion recognition in video," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, 2016, pp. 2752–2756.

[8] ——, "Audio visual emotion recognition with temporal alignment and perception attention," *arxiv:1603.08321*, 2016.

[9] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, USA, 2017.

[10] C. W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition," in *Proc. of INTERSPEECH*, San Francisco, USA, 2016, pp. 1387–1391.

[11] ——, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *Proc. of IEEE International Conference on Multimedia and Expo*, Hong Kong, China, 2017, pp. 583–588.

[12] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. of INTERSPEECH*, Dresden, Germany, 2015.

[13] C. Vladimir and S. Grigoriy, "Emotion recognition from speech with recurrent neural networks," *arxiv:1701.0807*, 2017.

[14] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, and S. S. Narayanan, "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.

[15] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, 2011.

[16] W. Han, H. Li, H. Ruan, and L. Ma, "Review on speech emotion recognition," *Journal of Software*, vol. 25, no. 1, 2014.

[17] M. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *Computer Science*, 2015.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. International Conference on Learning Representations*, vol. abs/1409.0473, 2015.

[19] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild." Sydney, Australia: Proc. of ACM International Conference on Multimodal Interaction, 2013, pp. 517–524.

[20] A. Graves, S. Fernández, and F. Gomez, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of International Conference on Machine Learning*, Pittsburgh, Pennsylvania, USA, 2006, pp. 369–376.

[21] T. Bluche, H. Ney, J. Louradour, and C. Kermorvant, "Framewise and CTC training of neural networks for handwriting recognition," in *Proc. of International Conference on Document Analysis and Recognition*, Nancy, France, 2015, pp. 81–85.

[22] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, and G. Diamos, "Deep speech 2: End-to-end speech recognition in English and Mandarin," *Computer Science*, 2015.

[23] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources & Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[24] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes – towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. of INTERSPEECH*, Brisbane, Australia, 2008, pp. 597–600.

[25] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, 2016.

[26] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[27] B. Schuller, B. Vlasenko, R. Minguez, G. Rigoll, and A. Wendemuth, "Comparing one and two-stage acoustic modeling in the recognition of emotion in speech," in *Proc. of ASRU 2007*, Kyoto, Japan, 2007, pp. 597–600.

[28] S. Parthasarathy and C. Busso, "Defining emotionally salient regions using qualitative agreement method," in *Proc. of INTERSPEECH*, San Francisco, USA, 2016, pp. 3598–3602.

[29] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*. Wiley-Blackwell, 1995.

[30] "The CMU pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict?in=C+M+U+ Dictionary.

[31] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," *Neurocomputing Algorithms Architectures & Applications*, vol. 68, pp. 227–236, 1990.

[32] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Proc. of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1263–1267.

[33] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Proc. of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2741–2745.

[34] X. Xia, J. Liu, W. Han, X. Zhu, H. Sahli, and D. Jiang, "Speech emotion recognition based on global features and dcnn," in *Proc. of International workshop on Affective Social Multimedia Computing (ASMMC), a satellite workshop of INTERSPEECH*, Stockholm, Sweden, 2017.

[35] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity and native language," in *Proc. of INTERSPEECH*, San Francisco, USA, 2016, pp. 2001–2005.

[36] F. Eyben, F. Weninger, and F. Gross, "Recent developments in OpenSMILE, the munich open-source multimedia feature extractor," in *Proc. of ACM International Conference on Multimedia*, Barcelona, Catalunya, Spain, 2013, pp. 835–838.