AUTOMATIC SPEECH EMOTION RECOGNITION USING RECURRENT NEURAL NETWORKS WITH LOCAL ATTENTION

Seyedmahdad Mirsamadi[1], *Emad Barsoum*[2], *Cha Zhang*[2]

[1]Center for Robust Speech Systems, The University of Texas at Dallas, Richardson, TX 75080, USA
[2]Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

mirsamadi@utdallas.edu, ebarsoum@microsoft.com, chazhang@microsoft.com

ABSTRACT

Automatic emotion recognition from speech is a challenging task which relies heavily on the effectiveness of the speech features used for classification. In this work, we study the use of deep learning to automatically discover emotionally relevant features from speech. It is shown that using a deep recurrent neural network, we can learn both the short-time frame-level acoustic features that are emotionally relevant, as well as an appropriate temporal aggregation of those features into a compact utterance-level representation. Moreover, we propose a novel strategy for feature pooling over time which uses local attention in order to focus on specific regions of a speech signal that are more emotionally salient. The proposed solution is evaluated on the IEMOCAP corpus, and is shown to provide more accurate predictions compared to existing emotion recognition algorithms.

Index Terms— Emotion Recognition, Deep Recurrent Neural Networks, Attention mechanism

1. INTRODUCTION

The emotional state of human beings is an important factor in their interactions, influencing most channels of communication such as facial expressions, voice characteristics, and the linguistic content of verbal communications. Speech is one of the primary faucets for expressing emotions, and thus for a natural human-machine interface, it is important to recognize, interpret, and respond to the emotions expressed in speech. Emotions influence both the voice characteristics as well as linguistic content of speech. In this study, we focus on the acoustic characteristics of speech in order to recognize the underlying emotions.

A lot of research on speech emotion recognition (SER) has been focused on the search for speech features that are indicative of different emotions [1, 2]. While a variety of both short-term and long-term features have been proposed [3], it is still unclear which features are more informative about emotions. Traditionally, the most popular approach has been to extract a large number of statistical features at the utterance level, apply dimension reduction techniques to obtain a compact representation, and finally perform classification with a standard machine learning algorithm [4, 5, 10]. More specifically, the feature extraction consists of two stages. First, a number of acoustic features that are believed to be influenced by emotions are extracted from short frames of typically 20 to 50 msec. These are often referred to as Low-Level Descriptors (LLD). Next, different statistical aggregation functions (such as mean, max, variance, linear regression coefficients, etc.) are applied to each of the LLDs over the duration of the utterance, and the results are concatenated into a long feature vector at the utterance level. The role of these high-level statistical functions (HSF) is to roughly describe the temporal variations and contours of the different LLDs during the utterance. The assumption here is that emotional content lies in the temporal variations, rather than static values of short-term LLDs. Different classification methods have been used to categorize the obtained utterance-level features [3], with SVMs being one of the most popular choices in SER. Table 1 lists some examples of LLDs and HSFs commonly used for SER.

Table 1. Common low-level descriptors (LLDs) and high-level statistical functions (HSFs) for SER.

LLDs	pitch (F_0), voicing probability, energy, zero-crossing rate, Mel-filterbank features, MFCCs, formant locations/bandwidths, harmonics-to-noise ratio, jitter, etc.
HSFs	mean, variance, min, max, range, median, quartiles, higher order moments (skewness, kurtosis), linear regression coefficients, etc.

Recently, there has been growing interest to apply deep learning to automatically learn useful features from emotional speech data. The authors in [6] used a Deep Neural Network (DNN) on top of traditional utterance-level statistical features to improve the recognition accuracy compared to conventional classifiers such as Support Vector Machines (SVM). The works in [7] and [8] used deep feed-forward and recurrent neural networks (RNN) at the frame level to learn the short-term acoustic features, followed by traditional map-

ping to a sentence-level representation using extreme learning machines (ELM). In [9], the authors used both convolutional and recurrent layers to learn the mapping directly from time-domain speech signals to the continuous-valued circumplex model space of emotion.

One issue that appears to still puzzle researchers applying deep learning framework in SER, is how to effectively balance the short-term characterization at the frame level and long-term aggregation at the utterance level. Two bidirectional LSTM layers were used in [9] to transform short-term convolutional features directly into continuous arousal and valence output. However, works in [7] and [8] have both applied ELM for the utterance level aggregation, despite the fact that they already adopted a CTC-style recurrent network underneath [8]. The challenge lies in how speech emotion data are typically tagged. In most SER data sets, the emotion labels are given at the utterance level. However, an utterance often contains many short silence periods, and in many cases only a few words in the utterance are emotional, while the majority of the rest are emotionless. The silence periods can be addressed using a voice activity detector (VAD) [7], or by null label alignment [8], however, we are not aware of any work in the past that explicitly handles emotionally-irrelevant speech frames.

In this paper, we combine bidirectional LSTM with a novel pooling strategy using an attention mechanism which enables the network to focus on emotionally salient parts of a sentence. With the attention model, our network can simultaneously ignore silence frames and other parts of the utterance which do not carry emotional content. We conduct experiments on the the IEMOCAP corpus [12] by comparing various approaches, including frame-wise training, final-frame LSTM training, mean-pooling, and the proposed approach, weighted-pooling with local attention. Our preliminary results show that in general adding a pooling layer on top of the LSTM layers produces the better performance, and the weighted pooling with attention model further improves over mean-pooling by about 1-2% on IEMOCAP.

2. EMOTION RECOGNITION USING RECURRENT NEURAL NETWORKS

Most of the features listed in Table 1 can be inferred from a raw spectrogram representation of the speech signal. It is therefore reasonable to assume that given a fixed set of (differentiable) HSFs and sufficient data, similar short-term features can be learned from a raw spectral representation. Fig. 1(a) shows an example structure to learn short-term LLDs, using a few layers of dense nonlinear transformations. Note that the statistical functions in the context of neural networks function as pooling layers over the time dimension.

For the rest of the paper, we will focus mostly on learning both short term LLDs and long-term aggregation. We study the use of recurrent networks which can effectively remember relevant long-term context from the input features. The RNN output nodes in this case are expected to represent different long-term integrations over the frame-level LLDs. The challenge that arises with such a structure is how to train the network parameters, since the emotion labels are at the utterance level, which may not be blindly used at the frame-level. In the following, we discuss different approaches to address this issue.

2.1. Frame-wise training

The most naïve approach is to assign the overall emotion to each and every frame within the utterance, and train the RNN in a frame-wise manner by back-propagating cross-entropy errors from every frame (Fig. 1(b)). However, it is not reasonable to assume that every frame within an utterance represents the overall emotion. This is both because there are short pauses (silence frames) within the utterance, and because the overall emotion decision for a training example is often influenced only by a few words which strongly show the emotion, as opposed to the whole utterance. Second, since we are assuming the RNN outputs to be long-term aggregations over the input LLDs, we should not expect the outputs to have the desired long-term representation starting from the first frame. Rather, the RNN should be given enough past history (input context) until it can produce the correct representation.

2.2. Final-frame (many-to-one) training

An alternative to frame-wise training is to only pick the final RNN hidden representation at the last frame and pass it through the output softmax layer. The errors are then back-propagated to the beginning of the utterance. Fig. 1(c) shows such a structure, in which the final output at each direction is used, since the recurrent layer we adopted is bi-directional. Although this approach ensures that the RNN receives sufficient context before being expected to produce the desired representation, it still assumes that all parts of the utterance perfectly exhibit the overall emotion. As an example, if a sentence starts with a strong happy emotion but the emotion fades towards the end, the RNN output will start to diverge from the desired representation of happy as it encounters the non-emotional frames towards the end of the utterance. Therefore, relying only on the final frame of the sequence may not fully capture the intended emotion.

2.3. Mean-pooling over time

Instead of computing the cross-entropy error at all frames or the last frame, it is possible to perform a mean-pooling over time on the RNN outputs, and pass the result to the final softmax layer (Fig. 1(d)). This assumes there are sufficient *correct* RNN outputs within the utterance to dominate the average value. It will be shown in section 3 that

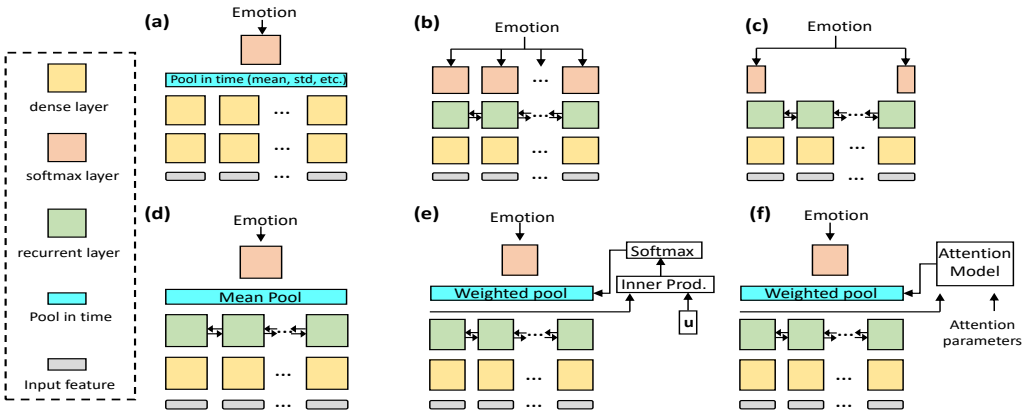

Fig. 1. Architectures for applying DNN/RNN for SER. (a) Learning LLDs using fixed temporal aggregation. (b) frame-wise training. (c) final-frame (many-to-one) training. (d) Mean-pooling in time. (e) Weighted pooling with logistic regression attention model. (f) general attention model.

this simple mean pooling strategy provides considerably better results compared to frame-wise and final-frame training. However, this approach still suffers from the problems discussed above, namely the presence of silence frames and non-emotional speech frames within the utterance. Including these frames in the overall mean pooling will distort the desired representation for the emotion.

2.4. Weighted-pooling with local attention

Inspired by the idea of attention mechanisms in neural machine translation [11], we introduce a novel weighted-pooling strategy to focus on specific parts of an utterance which contain strong emotional characteristics. Instead of mean pooling over time, we compute a weighted sum where the weights are determined based on an additional set of parameters in an attention model. Using a simple logistic regression as the attention model, the solution can be formulated as follows.

As shown in Fig. 1(e), at each time frame t, the inner product between the attention parameter vector $\mathbf{u}$ and the RNN output $\mathbf{y}_t$ is computed, and interpreted as a score for the contribution of that frame to the final utterance-level representation of the emotion. A softmax function is applied to the results to obtain a set of final weights for the frames which sum to unity:

$$\alpha_t = \frac{\exp(\mathbf{u}^T \mathbf{y}_t)}{\sum_{\tau=1}^{T} \exp(\mathbf{u}^T \mathbf{y}_\tau)}. \quad (1)$$

The obtained weights are used in a weighted average in time to get the utterance-level representation:

$$\mathbf{z} = \sum_{t=1}^{T} \alpha_t \mathbf{y}_t. \quad (2)$$

The pooled result is finally passed to the output softmax layer of the network to get posterior probabilities for each emotional class. The parameters of both the attention model ($\mathbf{u}$ in Eq. (1)) and the RNN are trained together by back-propagation. Note that the weighted pooling described here was based on a simple logistic regression attention model. However, given sufficient data, it is possible to use more sophisticated (i.e. deeper) models for attention (Fig. 1(f)).

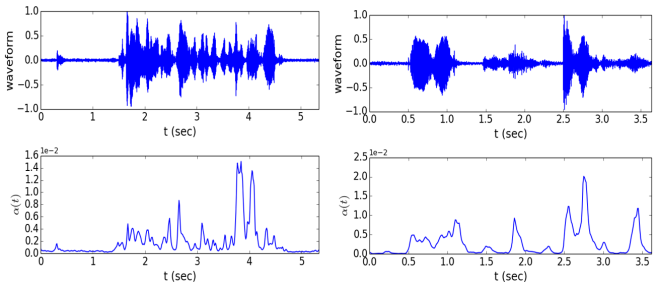

Fig. 2. Local attention weights for two test examples. Top: the raw waveform; bottom: the attention weight $\alpha(t)$ over time.

Fig. 2 illustrates the obtained attention weights (α_t) together with the corresponding waveforms for two different test examples. The obtained weights indicate that the introduced attention-based pooling achieves two desirable properties necessary for an RNN-based dynamic classfication of emotions. First, the silence frames within the signals are automatically assigned very small weights and effectively ignored in the pooling operation, without the need for any external mechanism such as VAD. Moreover, the speech frames are also assigned different weights based on how emotional they have been decided to be. So the attention model does not focus on energy only, and it is capable of considering the emotional content of different portions of speech.

3. EXPERIMENTS

To assess the performance of the introduced RNN-based SER architectures, we perform speaker-independent SER experiments using IEMOCAP dataset [12]. The corpus is organized in 5 sessions, in each of which two actors are involved in scripted scenarios or improvisations designed to elicit specific emotions. We use audio signals from four emotional categories of *happy*, *sad*, *neutral*, and *angry*. Four sessions of the corpus are used for training, and the remaining session used for testing. The experiments apply both raw spectral features (257-dimensional magnitude FFT vectors), as well as hand-crafted LLDs commonly used for SER, consisting of fundamental frequency (F_0), voicing probability, frame energy, zero-crossing rate, and 12 Mel-frequency Cepstral Coefficients (MFCC). Together with their first order derivatives, this makes 32-dimensional LLDs for each frame. Both of these frame-level features are extracted from 25 msec segments at a rate of 100 frames/sec, and normalized by the global mean and standard deviations of neutral speech features in the training set.

As a baseline SER system, we use a SVM classifier with Radial Basis Function (RBF) kernel on utterance-level features obtained by applying fixed statistical functions to the hand-crafted LLDs (mean, std, min, max, range, extremum positions, skewness, kurtosis, and linear regression coefficients). The train data is imbalanced with respect to the emotional classes, so we use a cost-sensitive training strategy in which the cost of each example is scaled according to the number of examples in that category. Since the test sets are also imbalanced, we report both the overall accuracy on test examples (weighted accuracy, WA) as well as average recall over the different emotional categories (unweighted accuracy, UA). We use Rectified Linear (ReLU) dense layers with 512 nodes for LLD learning, and Bi-directional Long Short-Term Memory (BLSTM) recurrent layers with 128 memory cells for learning the temporal aggregation, with 50% dropout on all layers during training to prevent over-fitting.

Table 2 compares the classification performance of learned and hand-crafted LLDs with different fixed HSFs for temporal aggregation. The learned LLDs with a softmax classifier provide better accuracy in most cases compared to conventional emotion LLDs with a SVM. Also, while the SVM approach necessarily needs a large number of HSFs to reach its peak performance, the DNN solution is less sensitive to the number and diversity of the used HSFs. The results in Table 3 with hand-crafted LLDs focus on learning the temporal aggregation task with recurrent layers. Frame-wise and final-frame training provide lower accuracies because they assume all frames carry the overall emotion and they include the silence frames. Mean-pooling in time can in principle have the same problems, but in practice provides significantly higher accuracies, since for short and carefully segmented IEMOCAP examples, the intended emotion is sufficiently dominant in a global mean pool. The proposed attention-based weighted pooling strategy outperforms all other training methods by focusing on emotional parts of utterances. Compared with traditional SVM solution, the proposed algorithm achieves +5.7% and +3.1% absolute improvements in WA and UA, respectively. Also presented in Table 3 are the results of jointly learning both LLDs and temporal aggregation from raw spectral data by a deep network of two hidden relu layers followed by a BLSTM layer. Although the joint learning provides slightly lower performance here, we attribute it to the lack of sufficient training examples to learn the parameters for both tasks. Given sufficient training examples, the parameters of short-term characterization, long-term aggregation, and the attention model can be jointly optimized for best performance.

Table 2. Accuracy comparison between hand-crafted LLDs and learned LLDs from raw spectral features.

Features	Classifier	HSFs	WA	UA
raw spectral	DNN[2]	Mean	56.4%	53.4%
		Mean, Min, Max	59.3%	54.9%
		Full	58.3%	54.4%
emotion LLDs	SVM	Mean	53.3%	49.3%
		Mean, Min, Max	55.4%	52.9%
		Full[1]	57.8%	55.7%

[1] Mean, std, min, max, range, skewness, kurtosis.
[2] Two relu hidden layers of 512 nodes (Fig. 1(a)).

Table 3. Accuracy comparison between RNN architectures

Features	Temporal aggregation	WA	UA
raw spectral	RNN–frame-wise (Fig.1(b))	57.7%	53.8%
	RNN–final frame (Fig.1(c))	54.4%	49.7%
	RNN–mean pool (Fig.1(d))	56.9%	55.3%
	RNN–weighted pool with attention (Fig.1(e))	61.8%	56.3%
emotion LLDs	RNN–frame-wise (Fig.1(b))	57.2%	51.6%
	RNN–final frame (Fig.1(c))	53.0%	54.9%
	RNN–mean pool (Fig.1(d))	62.7%	57.2%
	RNN–weighted pool with attention (Fig.1(e))	63.5%	58.8%

4. CONCLUSIONS

We presented different RNN architectures for feature learning in speech emotion recognition. It was shown that using deep RNNs, we can learn both frame-level characterization as well as temporal aggregation into longer time spans. Moreover, using a simple attention mechanism, we proposed a novel weighted time-pooling strategy which enables the network to focus on emotionally salient parts of an utterance. Experiments on IEMOCAP data suggests that the learned features provide better classification accuracy compared to traditional SVM-based SER using fixed designed features.

5. REFERENCES

[1] Marie Tahon and Laurence Devillers, "Towards a small set of robust acoustic features for emotion recognition: challenges," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 16–28, 2016.

[2] Björn Schuller, Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, et al., "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals.," in *INTERSPEECH*, 2007, pp. 2253–2256.

[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] Björn Schuller, Dejan Arsic, Frank Wallhoff, Gerhard Rigoll, et al., "Emotion recognition in the noise applying large acoustic feature sets," *Speech Prosody, Dresden*, pp. 276–289, 2006.

[5] Aitor Álvarez, Idoia Cearreta, Juan Miguel López, Andoni Arruti, Elena Lazkano, Basilio Sierra, and Nestor Garay, "Feature subset selection based on evolutionary algorithms for automatic emotion recognition in spoken spanish and standard basque language," in *International Conference on Text, Speech and Dialogue*. Springer, 2006, pp. 565–572.

[6] André Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Günter Meier, and Björn Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5688–5691.

[7] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine.," in *Interspeech*, 2014, pp. 223–227.

[8] Jinkyu Lee and Ivan Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech*, 2015.

[9] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Stefanos Zafeiriou, et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[10] Carlos Busso, Murtaza Bulut, and SS Narayanan, "Toward effective automatic recognition systems of emotion in speech," *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds*, pp. 110–127, 2013.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.