USING REGIONAL SALIENCY FOR SPEECH EMOTION RECOGNITION

Zakaria Aldeneh, Emily Mower Provost

University of Michigan, Ann Arbor, MI 48109, USA
Computer Science and Engineering
{aldeneh, emilykmp}@umich.edu

ABSTRACT

In this paper, we show that convolutional neural networks can be directly applied to temporal low-level acoustic features to identify emotionally salient regions without the need for defining or applying utterance-level statistics. We show how a convolutional neural network can be applied to minimally hand-engineered features to obtain competitive results on the IEMOCAP and MSP-IMPROV datasets. In addition, we demonstrate that, despite their common use across most categories of acoustic features, utterance-level statistics may obfuscate emotional information. Our results suggest that convolutional neural networks with Mel Filterbanks (MFBs) can be used as a replacement for classifiers that rely on features obtained from applying utterance-level statistics.

Index Terms— speech emotion recognition, convolutional neural network, machine learning

1. INTRODUCTION

Allowing machines to understand human emotion from speech has many important applications, including aiding in the diagnosis of depression [1, 2], and monitoring mood state for bipolar patients [3, 4]. Building accurate speech emotion recognition (SER) systems, however, is a challenging task and is still an open research problem.

Traditional SER systems follow one of three major approaches. In the first approach, utterance-level statistical functionals are applied to low-level descriptors (LLDs) extracted from utterances of variable lengths to obtain fixed-length features that describe the global characteristics of the given utterances. These fixed-length features can then be used to train machine learning classifiers (e.g. [5, 6]). While popular, we hypothesize that this approach dilutes important regional information by combining it with potentially irrelevant information from neighboring frames.

Two recent papers [7, 8] showed that one can train classifiers using only a portion of the information contained within utterances and still achieve competitive results. In particular, Le et al. [7] showed that state-of-the-art results can be obtained on the FAU Aibo 2-class problem using less than 50% of the data contained within an utterance. Kim et al. [8] showed that emotional information in an utterance is regionalized and follows specific patterns. Echoing the findings of Le et al. they showed that, in some cases, systems that use only 59% of the data within an utterance can achieve performance that is similar to that achieved by systems that use 100% of the data. This suggests that traditional SER approaches include irrelevant information when creating fixed-length features.

In the second approach, statistical functionals are applied to windowed segments of utterances to create statistical descriptions of the segments. These statistics are then classified to create sequences of emotion confidences. Given this sequence of emotion confidences, the problem becomes a time series classification problem (e.g. [9]). This approach assumes that all segments take the same emotional label as their parent utterance and thus assumes that all regions of utterances contain relevant emotional information.

Finally in the third approach, frameworks that are capable of directly modeling temporal LLDs are used to build SER systems. Many of these approaches were inspired by approaches proposed in the automatic speech recognition (ASR) community. Notable approaches include HMM-DNN hybrids [10] and deep end-to-end systems [11].

We hypothesize that focusing on emotionally salient regions of utterances can allow us to build robust SER systems that do not require defining statistical functionals or making any assumptions about frame-level emotional labels. In this work, we use convolutional neural networks (CNNs) to learn emotion classifiers from speech. CNNs have shown tremendous success in the fields of ASR [12], computer vision [13], and sentence classification [14]. CNNs allow multiple regions of the input to share the same weights; overcoming the scalability problem of regular neural networks. In addition, CNNs can be applied to inputs of variable sizes, thus easing one of the challenges of dealing with variable length speech data.

The contributions of this work are as follows: (1) we show how ideas presented in the sentence classification literature are applicable to the field of SER; (2) we show how a simple CNN that uses minimally hand-engineered features can yield competitive results when compared to results obtained from systems trained on popular emotion feature sets; (3) we show how applying statistical functionals to temporal LLDs can washout information causing loss of performance; (4) we

show how using speed augmentation can improve the performance of SER systems.

2. RELATED WORK

CNNs have been used for SER. Most notably, Mao et al. [15] used CNNs to learn salient features to be used by an SVM for classification. The authors followed three steps to build their SER system. First, they used sparse auto-encoders to learn filters from spectrogram segments. The authors convolved the learned filters with spectrogram fragments to produce feature vectors. Second, the authors mapped the feature vectors into two smaller feature vectors using a semi-supervised objective function. The objective function disentangled affect-salient features from other non-salient features. Third, the authors used the affect-salient features to train SVMs. The authors finally compared the discriminative performance of features obtained from different stages of the CNN.

Other works used neural networks and recurrent neural networks for SER. Le et al. [10] followed an approach that is similar to those followed in ASR literature and used a HMM-DNN hybrid approach [16] to train an SER system. The authors investigated different ways to model emotion as an HMM and finally drew a contrast between the fields of emotion and speech recognition.

Han et al. [17] and Lee et al. [18] both took a multi-step approach to the problem of SER. In the first step, Han et al. [17] trained a neural network using frame-level features (along with contextual information) while Lee et al. [18] trained a 2-layer bidirectional long short-term memory (BLSTM) network. The trained models were used to produce frame-level emotional predictions (four channel time-series). Both authors applied statistical functionals to the time-series data before feeding the results into another simple neural network for utterance-level classification.

Xia et al. [19] used denoising autoencoders to build SER models that take gender into account. The authors trained gender-specific models using neutral speech obtained from a large ASR dataset. The results suggested that modeling gender variability can be useful for emotion recognition. In other work, Xia et al. [6] used a multi-task learning approach to leverage additional data with continuous labels (as opposed to categorical labels) to train a network for SER. The authors showed that using regression as a secondary task can improve the overall performance of the system when compared to a single-task system that only relies on examples with categorical labels.

Finally, motivated by a recent trend in deep learning where raw data is used with minimal feature pre-processing, Trigeorgis et al. [11] devised an end-to-end deep network that worked on raw time-domain signals. The authors first applied convolutions to extract features before they fed the extracted features into a LSTM structure for prediction in the valence-activation space.

The approaches followed in the cited related work do at least one of the following: (1) make assumptions about the length of utterances and the temporal resolution of labels [11]; (2) rely on manual feature engineering [6, 17–19]; (3) apply statistical functionals on top of temporal LLDs [6, 19]; (4) follow a multi-step process for building the emotion recognition system [15, 17, 18]; (5) make assumptions about frame-level emotional labels and/or dynamics of emotion [10, 15, 17, 18]. In contrast, the approach that we take in this paper does not do any of the aforementioned points.

3. MODEL

Motivated by architectures used in the field of sentence classification (e.g. [14]), where the goal is to predict the class of a given variable length sentence (e.g. positive/negative review), we build a simple four-layer CNN for SER (Figure 1). Our model has four major components: (1) convolutional layer; (2) max-pooling over time layer; (3) dense layer; and (4) softmax layer. The convolutional layer identifies emotionally salient regions within variable length utterances and creates a sequence of feature maps. The max-pooling over time layer propagates features with the highest value to the dense layer. The max-pooling over time layer induces time invariance and creates a fixed-size feature vector from a variable length input. Finally, the dense and softmax layers provide further modeling and prediction. We describe each component in more detail in this section.

Let $\mathbf{x}_i^u \in \mathbb{R}^d$ be a d dimensional feature vector available at frame i of an utterance u. Then, we represent an utterance u with T frames as:

$$\mathbf{X}^u = [\mathbf{x}_1^u, \mathbf{x}_2^u, \ldots, \mathbf{x}_T^u]$$

note that d is fixed while T varies across utterances. A temporal convolution operation applies a filter $\mathbf{w} \in \mathbb{R}^{d \times s}$, where s is the width of the filter, to produce a new feature set of length $T - s + 1$. So convolving filter $\mathbf{w}$ with $\mathbf{X}^u$ yields:

$$\mathbf{c}^u = \left[c_1^u, c_2^u, \ldots, c_{T-s+1}^u\right]$$

where each $c_i^u \in \mathbb{R}$ is obtained using the following operation:

$$c_i^u = \sum_{m=1}^{s} \sum_{n=1}^{d} \left(\left[\mathbf{x}_i^u, \ldots, \mathbf{x}_{i+s-1}^u\right] \odot \mathbf{w}\right)_{m,n}$$

where $\odot$ denotes the element-wise multiplication operation. We leave out the bias term in the above equation for simplicity.

The convolution operation allows the network to extract local features from an utterance. The width of the convolutional filters dictates the size of the region from which we create the feature maps. Wider filters capture long-term interactions while narrower filters capture short-term interactions. We can apply multiple filters, each with different weights, to

extract different information from the same region. It is customary to apply a non-linearity activation function to the outputs of the convolution operation. We use the rectified linear unit (ReLU) in this work [20].

We follow the convolutional layer by a max-pooling over time operation. Given a sequence of features, the max-pooling over time operation returns the maximum feature within that sequence. This ensures that only emotionally salient information is propagated. We follow the max-pooling layer by dense layers and then by a softmax layer for prediction. The softmax layer takes a C-dimensional feature vector and outputs a C-dimensional probability distribution.

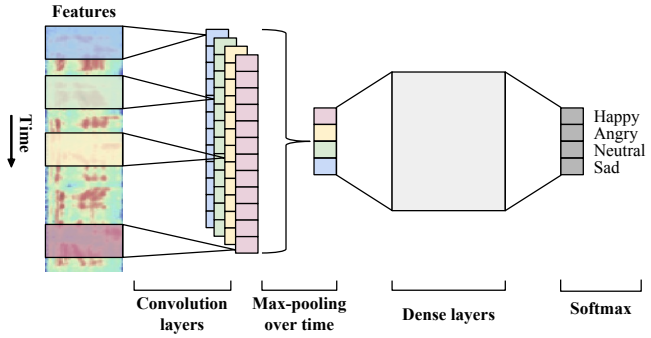

Fig. 1. Network Overview (four filters shown).

4. DATASETS AND RECIPE

4.1. Datasets

We evaluate our system on two emotion datasets: IEMOCAP [21] and MSP-IMPROV [22]. We only use the audio portion of the datasets. Both datasets were collected following theatre theory in order to simulate natural dyadic interactions between actors. We use categorical evaluations with majority agreement for both datasets. We only use four emotional categories in our work: *Happy*, *Sad*, *Angry*, and *Neutral*.

IEMOCAP. The IEMOCAP dataset is comprised of five sessions, where each session contains utterances from two speakers (one male and one female). This results in 10 unique speakers. To be consistent with previous work [6], we include *excitement* utterances with *happiness* ones. The final dataset contains a total of 5531 utterances (1103 *Angry*, 1708 *Neutral*, 1084 *Sad*, 1636 *Happy*).

MSP-IMPROV. The MSP-IMPROV dataset is comprised of six sessions, where each session contains utterances from two speakers (one male and one female). This results in 12 unique speakers. The final dataset contains a total of 7798 utterances (792 *Angry*, 3477 *Neutral*, 885 *Sad*, 2644 *Happy*).

4.2. Feature Extraction and Data Augmentation

We use the openSmile toolkit [23] to extract 40-dimensional log Mel filterbank features (MFBs) from each utterance. We create our initial segments by sliding a Hamming window of width 25ms with an overlap of 10ms. We perform speaker-specific z-normalization on all features.

We increase the size of our training data by creating two different copies of each utterance following the approach described in [24]. In particular, for a given training utterance, we apply the *speed* effect found in the *Sox*[1] audio manipulation tool at factors of 0.9 and 1.1 to create two versions of the original utterance. We report the performance with and without augmentation in the results section.

4.3. Experimental Recipe

We follow a leave-one-speaker-out evaluation scheme for both datasets. In each session, we use utterances from one speaker for testing and utterances from the other speaker for validation and early stopping. We use utterances from all other speakers for training. This scheme allows using a validation speaker who has similar acoustic and recording conditions to those of the test speaker. We report the mean and standard deviation of the unweighted average recall (UAR) from all speakers. UAR is a popular metric used in SER because of imbalanced datasets.

We implement our network using the Keras deep learning library. In our experiments, we fix the dense network to have three layers with shape 1024:1024:4. We regularize our network using early stopping. We randomly initialize the weights of our network following recommendation by He et al. [25].

We minimize the cross-entropy loss function using RMSprop [26] with an initial learning rate of 1e-4. We use a maximum batch size of 50. To create batches, we first edge-pad utterances so that they have lengths that are integer multiples of 32. Then, we group the resulting same-length utterances for batch training. To deal with class-imbalance, we scale the loss function using weights that are inversely proportional to class frequencies. For a given sample i, assume that $\mathbf{y}_i$ is the true label vector (all zeros but with a one at the correct class) and $\hat{\mathbf{y}}_i$ is the predicted probability distribution from the softmax layer, then the loss function takes the following form:

$$L_i = -w_i \sum_{j=0}^{C-1} y_{i,j} \log(\hat{y}_{i,j})$$

where C is the total number of classes and w_i is the scaling factor associated with sample i.

We compute the UAR on the validation set at the end of each epoch. If the UAR does not improve, then we restore the learned weights to their initial values at the beginning of the

[1] http://sox.sourceforge.net

epoch and reduce the learning rate by 1.4. The process stops if the UAR does not improve for 10 consecutive epochs. For each setup, we train 10 models and average their predictions.

5. EXPERIMENTS

We try to answer the following questions in our experiments: (1) does capturing regional information using CNNs provide an advantage over computing utterance-level statistics? (2) how does the performance of a system that focuses on emotionally salient regions compare to those of systems trained with popular large feature sets?

To answer the first question, we capture utterance-level features by applying the 12 IS09 statistical functionals [27] to 40 MFBs to get fixed-length feature vector of size 480. We remove the convolutional component of the CNN and train the dense layers directly using the captured statistical features. The first row of Table 1 shows the results we obtain from training a dense network on utterance-level statistical functionals.

Next, we train a CNN directly on temporal MFBs without applying any statistical functionals. We vary the width of the filters from 8 to 128. To ensure a fair comparison, we adjust the number of filters in each setup such that the total number of learnable parameters are equal to those used in the dense network trained on utterance-level statistical features. Table 1 shows the results we obtain for different filter widths.

To answer the second question, we train a set of SVMs using popular feature sets. We extract IS09 [27], IS13 [28], GeMAPS and eGeMAPS [29] features. We apply the same 12 statistical functionals to IS09 and IS13 LLDs. We use an RBF kernel and do a grid search using validation data to pick the optimal hyper-parameters in $C \in \{2^0, 2^2, \ldots, 2^{12}\}$, and $\gamma \in \{2^{-15}, 2^{-13}, \ldots, 2^{-3}\}$. We scale the SVM cost parameter to take class-imbalance into account. We use augmented data for all SVM experiments to ensure a fair comparison. Table 2 shows the results we obtain using different sets of features.

Next, we train a CNN that uses multi-width filters (8, 16, 32, 64) directly on temporal MFBs. Combining multiple widths allows the network to consider multiple contextual dependencies simultaneously. This approach showed promise in some sentence classification applications [30]. We use 384 filters for each width to set the total number of inputs to the dense layers to be equal to the total number of features we obtain from IS13 features. The first two rows of Table 2 shows the results we obtain from this setup.

6. RESULTS

Table 1 shows that focusing on regional information when training a network is better than training a network using features obtained from statistical functionals. When focusing on regional content, we see a significant improvement ($p < 0.05$) of 2.2% on IEMOCAP and a minor improvement of 0.7% on MSP-IMPROV over results of networks that rely on utterance-level statistics.

Table 1. Regions vs. utterance-level statistics (40 MFBs) ("*" indicates $p < 0.05$ under paired t-test with first row)

Filter Width	UAR (%)	
	IEMOCAP	MSP-IMPROV
statistics	58.5 ± 3.0	49.8 ± 4.7
8	58.1 ± 3.0	50.2 ± 3.7
16	60.0 ± 2.8*	50.5 ± 3.5
32	60.2 ± 3.1*	50.4 ± 2.9
64	60.7 ± 2.6*	50.2 ± 3.9
128	57.9 ± 3.2	48.0 ± 3.7

Table 2. System performance comparison ("*" indicates $p < 0.05$ under paired t-test with first row)

Method	UAR (%)	
	IEMOCAP	MSP-IMPROV
CNN + 40 MFBs	61.8 ± 3.0	52.6 ± 3.8
CNN + 40 MFBs (no aug)	59.5 ± 3.1*	49.8 ± 2.9*
SVM + IS09	60.5 ± 2.9	53.3 ± 5.0
SVM + IS13	61.7 ± 2.9	53.8 ± 6.0
SVM + GeMAPS	57.9 ± 3.2*	52.1 ± 4.7
SVM + eGeMAPS	58.7 ± 2.7*	52.4 ± 5.0

Table 2 shows that a network that combines multi-width filters that is trained using temporal MFBs yields UARs that are statistically comparable ($p \geqslant 0.05$) to those obtained from SVMs trained using IS09 and IS13 feature sets. Our results suggest that CNNs with MFBs can be used as replacement for traditional SVMs with hand-engineered features for SER. Table 2 also shows that augmenting the dataset using speed perturbation gives a significant improvement ($p < 0.05$) of 2.3% and 2.8% on IEMOCAP and MSP-IMPROV datasets respectively.

The SVM + IS13 setup yields the highest UAR for the MSP-IMPROV dataset (though not significantly higher than the UAR obtained from the CNN + 40 MFBs setup). IS13 contains a total of 1560 (130×12) features. These features include spectral, energy, and voicing features. In contrast, our system only uses 40 MFBs as features.

Xia et al. [6] obtained a UAR of 62.4% on IEMOCAP after training a deep neural network using 1582 hand-engineered features and utilizing a multi-task learning approach to incorporate more data. In contrast, our system is simpler, requires minimal feature engineering, and is trained in an end-to-end fashion.

For future work, we plan to study the effect of combining filters with varying widths and study the effect of appending additional LLDs (e.g. energy, pitch, etc.).

7. REFERENCES

[1] S. Scherer, G. Lucas, J. Gratch, and L. Morency A. Rizzo, "Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews," *IEEE Transactions on Affective Computing*, 2016.

[2] N. Cummins, J. Epps, V. Sethu, and J Krajewski, "Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech," in *ICASSP*, 2014.

[3] J. Gideon, E. Mower Provost, and M. McInnis, "Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder," in *ICASSP*, 2016.

[4] Z. Karam, E. Mower Provost, S. Singh, J. Montgomery, et al., "Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech," in *ICASSP*, 2014.

[5] E. Mower Provost, M. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[6] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, 2016.

[7] D. Le and E. Mower Provost, "Data selection for acoustic emotion recognition: Analyzing and comparing utterance and sub-utterance selection strategies," in *Affective Computing and Intelligent Interaction (ACII)*, 2015.

[8] Y. Kim and E. Mower Provost, "Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions.," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

[9] Y. Kim and E. Mower Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *ICASSP*, 2013.

[10] D. Le and E. Mower Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *Automatic Speech Recognition and Understanding (ASRU)*, 2013.

[11] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, S. Zafeiriou, et al., "ADIEU features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*, 2016.

[12] T. Sainath, M. Abdel-rahman, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *ICASSP*, 2013.

[13] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems (NIPS)*, 2012.

[14] Y. Kim, "Convolutional neural networks for sentence classification," *Conference on Empirical Methods in Natural Language Processing*, 2014.

[15] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, 2014.

[16] D. Yu and L. Deng, *Automatic Speech Recognition*, Springer, 2012.

[17] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proceedings of Interspeech*, 2014.

[18] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proceedings of Interspeech*, 2015.

[19] R. Xia, J. Deng, B. Schuller, and Y. Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *ICASSP*, 2014.

[20] M. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, et al., "On rectified linear units for speech processing," in *ICASSP*, 2013.

[21] C. Busso, M. Bulut, et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[22] C. Busso, S. Parthasarathy, A. Burmania, M. Abdel-Wahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, 2015.

[23] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *ACM international conference on Multimedia*, 2013.

[24] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of Interspeech*, 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[26] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, 2012.

[27] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge.," in *Proceedings of Interspeech*, 2009.

[28] B. Schuller, S. Steidl, A. Batliner, et al., "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of Interspeech*, 2013.

[29] F. Eyben, K. Scherer, B. Schuller, et al., "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.

[30] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," *arXiv preprint arXiv:1510.03820*, 2015.