INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Learning Temporal Clusters Using Capsule Routing for Speech Emotion Recognition

Md Asif Jalal[1], *Erfan Loweimi*[2], *Roger K Moore*[1], *Thomas Hain*[1]

[1]Speech and Hearing Research Group (SPandH), University of Sheffield, Sheffield, UK
[2]Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, UK
{majalal1, r.k.moore, t.hain}@sheffield.ac.uk, e.loweimi@ed.ac.uk

Abstract

Emotion recognition from speech plays a significant role in adding emotional intelligence to machines and making human-machine interaction more natural. One of the key challenges from machine learning standpoint is to extract patterns which bear maximum correlation with the emotion information encoded in this signal while being as insensitive as possible to other types of information carried by speech. In this paper, we propose a novel temporal modelling framework for robust emotion classification using bidirectional long short-term memory network (BLSTM), CNN and Capsule networks. The BLSTM deals with the temporal dynamics of the speech signal by effectively representing forward/backward contextual information while the CNN along with the dynamic routing of the Capsule net learn temporal clusters which altogether provide a state-of-the-art technique for classifying the extracted patterns. The proposed approach was compared with a wide range of architectures on the FAU-Aibo and RAVDESS corpora and remarkable gain over state-of-the-art systems were obtained. For FAO-Aibo and RAVDESS 77.6% and 56.2% accuracy was achieved, respectively, which is 3% and 14% (absolute) higher than the best-reported result for the respective tasks.

Index Terms: Speech emotion recognition, representation learning, BLSTM, CNN, capsule network

1. Introduction

Speech is the most natural way of human communication and reflects many aspects of us. This turns it into a complicated signal, encoding a large amount of information that can be categorised into lingual content, speaker-dependent attributes and environmental clues. Emotion is among the speaker-related information and plays an important role in human-human communication. As speech-driven user interfaces become more common in everyday life, lack of emotional intelligence is becoming more evident and adding this dimension to human-machine interaction is highly desirable.

From pattern recognition viewpoint, speech emotion recognition (SER) requires a front-end which extracts a set of features that ideally bear maximum correlation with emotion attribute while having the least sensitivity to other speech aspects. However, such signal parameterisation through feature engineering is challenging. In practice, general features such as MFCC, filterbank used along with classifiers such as HMM [1], SVM [2] and GMM [3] for emotion recognition from speech. In [4], Schuller et al. introduced a set of features handcrafted for emotion recognition problems. Hybrid models using HMMs and DBNs (deep belief networks) were also used for SER [5].

Deep neural networks (DNNs) can solve the data representation problem through learning a series of task-specific transformations. The network layers extract abstract representations and also filter out the irrelevant information which leads to a more accurate classification [6, 7] and better generalisation [8, 9]. Temporal models were also proposed for modelling sequential data with mid to long-term dependencies [10, 11].

In this paper, we present a novel architecture consisting of bidirectional long short term-memory (BLSTM), CNN and capsule layers. The BLSTM-CNN and the Capsule net in the proposed network play complementary roles: the former deals with the sequential nature and the temporal dynamics of the speech and the later further distils the information and classifies the extracted representations. The goal is to build a deep temporal model of utterances through leveraging the information encoded in the speech dynamics and its sequential nature. The experimental results prove the effectiveness of the proposed hybrid topology, leading to the state-of-the-art performance on FAU-Aibo [12–14] and RAVDESS databases [15] in both binary and 8-class emotion classification tasks.

The rest of this paper is organised as follows. In Section 2, representation learning through RNNs, 1D-CNN and capsule routing networks are reviewed and discussed. Section 3 explains the proposed architecture and its advantages. In Section 4 the experimental results are presented along with discussion and Section 5 concludes the paper.

2. Approaches to Representation Learning

Speech is a sequential data with a high temporal dynamics [16]. The speaker-related properties like emotion are distributed in the utterance and vary at a slower pace than the lingual content. To adequately capture such attributes, employed algorithm should be capable of handling sequence properties and go beyond mere short-term processing techniques. There are two main approaches in neural networks (NN) to deal with such sequential dynamics: augmenting the input by stacking the previous/next frames or using a network with some memory, representing the temporal evolution of the system's internal state.

2.1. Recurrent Neural Networks

In a regular feedforward NN, the temporal information is provided through the input by stacking neighbouring contextual frames. Setting the context length is done empirically and is a task and data-dependent practice [17]. On the other hand, Recurrent Neural Networks (RNN) by utilising the internal state, keep track of what has happened in the past and consider such temporal evolution while making a decision at each time step.

One issue with RNNs training is gradient vanishing or explosion [18] and to overcome that Long short-term memory (LSTM) [19] architecture was proposed. LSTM is an RNN with a special memory cell enclosed by three gates (input, forget and output) which can keep contextual information efficiently as far as it is required. This makes LSTMs outperform the normal

http://dx.doi.org/10.21437/Interspeech.2019-3068

RNNs in learning mid to long-term dependencies in the sequential data [20]. Further studies show that as well as the backward sequential context, the future context can also contribute to a more effective sequential processing. In this regard, bidirectional RNNs (BRNN) were proposed which consider both forward and backward contextual information [21]. BLSTM [17] is the combination of both LSTM and BRNNs and has the advantages of both architectures.

Here, for each time step t, we concatenate the activations of the unfolded network over the last $T-1$ time steps, providing a matrix of $T \times N$, where N is number of nodes of the BLSTM layer and T is the context length (including current frame t). In our work these BLSTM activations are the first temporal representation in the hierarchy.

2.2. 1D-Convolutional Capsules

CNNs have been successfully applied to various speech-related tasks such as ASR [22, 23] and emotion classification [9, 24]. CNN can model patterns with high robustness to variations and distortions [25]. A key component of our work is that we have applied 1D convolution to each of the temporal activation states of BLSTM for learning different abstract temporal representations. The output matrices of these convolution units are concatenated to form a capsule and a squashing function is applied to get vector representation for each capsule. A capsule is a group of neurons. The output of the capsule j, $\mathbf{o}_j^t$, is

$$\mathbf{o}_j^t = \mathbf{g}(\mathbf{s}_j), \tag{1}$$

where $\mathbf{s}_j$ is the input (concatenated 1D-Conv output) to the capsule j and $\mathbf{g}(.)$ is a squashing function. The intuition behind squashing is to shrink short vectors (less likely ones) to zero and long vectors (more likely ones) to nearly (below) 1. In Sabour et. al [26], $\mathbf{g}(\cdot)$ is defined as follows

$$\mathbf{g}(\mathbf{x}) = \frac{\|\mathbf{x}\|^{\mathbf{2}}}{\mathbf{1} + \|\mathbf{x}\|^{\mathbf{2}}} \frac{\mathbf{x}}{\|\mathbf{x}\|}. \tag{2}$$

In this paper, each 1D-Conv capsule consists of a group of neurons which collectively learn specific temporal entities presented by the BLSTM layer (shown in Fig 2). Contrary to the normal units which return a scalar, a capsule outputs a vector whose length is proportional with the likelihood of the entity presence, and direction represents the instantiation parameters.

2.3. Capsule Routing Network

After max-pooling the feature maps (outputs of the filters) in a CNN, an approximately translation invariant representation is achieved at the expense of losing orientational and relative spatial information about the parts or entities in an image [26]. For classification tasks where the input should be mapped into a label, this information loss may not pose a serious issue. However, when some segmentation is required, the (approximately) translation invariance rendered by max-pool becomes problematic due to the information loss it brings about. Sabour et al. [26] proposed a novel technique called *Routing by agreement* which yields a translation *equivariance* instead of the translation *invariance* in the CNNs. It deals with the problem mentioned above and better preserves the hierarchical relationships between lower and higher level features.

In the routing layer, the previous layer capsules try to estimate the output of the next layer. The capsules in lower layer predicts the output of the capsule n in the next layer. The input of the capsule $\mathbf{n}$, $\mathbf{s_n}$, is a weighted sum of such *predictions*

$$\mathbf{s_n} = \sum_{\mathbf{m}} \mathbf{c_{mn}} \hat{\mathbf{p}}_{\mathbf{n|m}}, \tag{3}$$

where $\hat{p}_{n|m}$ is the prediction of capsule m (in the lower level) about the output of capsule n and coefficients c_{mn} are weights. The weights are computed by a softmax function operating on b_{mn} coefficients which are learned through *routing-by-agreement* algorithm [26].

The *agreement* of the predicted output and actual output indicates the correctness of the prediction of the capsule m in the lower level about the capsule n's output in the higher level in the hierarchy, namely $\mathbf{a_{mn}}$. It determines how the lower and higher level features should be linked together which is in contrast to the conventional networks where the higher level feature are merely a weighted sum of the lower level features. The prediction about capsule n is computed as a product of the transformation matrix W_{mn} and the output of the preceding layer p_m. Then, the agreement, a_{mn}, is computed using an inner product

$$\hat{\mathbf{p}}_{\mathbf{n|m}} = \mathbf{W_{mn}} \mathbf{p_m}, \qquad \mathbf{a_{mn}} = \mathbf{p_n} \cdot \hat{\mathbf{p}}_{\mathbf{n|m}}. \tag{4}$$

This inner product is added to b_{mn} (prior probablities initialised by zero) such as $b_{mn} = b_{mn} + a_{mn}$. The transformation matrices, W_{mn}, are learned by backpropagation.

If capsule m (in the lower level) contains an instantiation of an entity represented by capsule n (in the higher level), the routing process makes the link between m and n capsules stronger and vice verse. Hence, the impact of the features from the m^{th} capsule on the n^{th} capsule is dynamically adjusted. Max-pooling is a *static* form of routing where only the most active unit in the pool is routed to the higher level, without considering the *dynamic* of the agreement between the low and high-level features in the hierarchy.

2.4. Supervector Extraction Using Generative Models

For supervector extraction in building baseline systems, we have used the method proposed in [3]. In this technique, first a GMM with M components is estimated for each class. Then, the posterior probabilities of all the components of GMMs are computed for each frame, averaged over all the utterance frames and finally stacked into a supervector.

The length of the supervector is $M \times C$ where C is the number of classes. For UBMGMM [27] and iVector [28] the supervector length is $M \times D$ where D indicates the length of the raw feature vectors. For tasks where the number of classes is inherently small such as emotion recognition, C is notably less than D. As a result; this approach leads to a more compact representation which facilitates faster and more efficient learning. For more details about the advantages of this approach, please refer to [3]. eGeMAPS [29] have also been used to extract supervector for comparison purposes.

3. Proposed DNN Architecture

The general architecture of the proposed framework is illustrated in Fig. 1. In this architecture, BLSTM, Conv-Capsule and the Capsule routing layer are playing complementary roles. BLSTM is used to deal with the sequential nature of the speech and its temporal dynamic. The Conv-Capsule model learns more abstract and richer representations of those temporal features. Finally, the capsule routing layer further distils the extracted patterns and maps them to a categorical distribution.

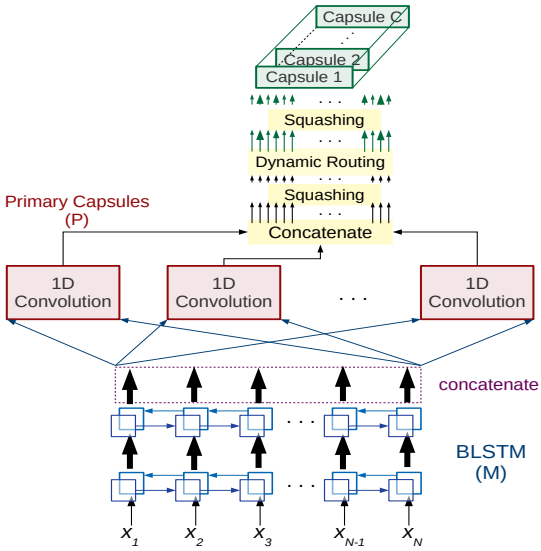

Figure 1: *The proposed architecture consisting bi-directional long-short term memory, 1D convolution and routing network.*

The Capsule net, in comparison with BLSTM, has a lower capability in handling and processing the forward/backward contextual information encoded in a sequential data like speech. On the other hand, BLSTM is not as powerful as the Capsule net in dealing with the static patterns. For example, it is not translation invariant. The order, i.e. using 1D convolution capsules on top of the BLSTM is justifiable as follows: first, temporal features enriched by contextual information, is extracted through BLSTM and then more abstract information distillation is carried out through the capsules and routing process. We preserve all the temporal cell state sequences using 1D convolution. It was noticed that using 2D convolution distorts the temporal alignment. As such, each part is used in a task which it best fits and the other component compensates for its shortcoming. This makes the structure *super-additive* and improves the overall performance as verified by the experiments.

Getting into more details, the overall network is comprised of two BLSTM layers, 1D conv-capsule layer consisting of capsules and a capsule routing layer. Input layer consisted of 70 nodes (length of the feature vector), and each BLSTM hidden layer (M in Fig. 1) contained 256 units.

The next layer consists of four 1D-CNN for two class categorisation task and 10 1D-CNN for eight class categorisation task. Each of these CNN has 90 filters and operates on the same receptive field of BLSTM temporal activation. The output feature maps are concatenated to form a capsule. The output of a capsule is squashed (Eqs. 1, 2) for getting the vector representation of the feature learned by that capsule.

The routing capsule layer is connected to the previous layer through the transformation matrices which is similar to a fully-connected layer in the conventional NNs, except for replacing the scalar-to-scalar with a vector-to-vector transform. The number of capsules in this layer equals the number of classes, and each output layer capsule is connected to all the capsules in the previous layer. The previous layer capsules compute the prediction of the output of capsules in the next layer (Eq. 4). The agreement coefficient is achieved by measuring the distance between the predicted output and the actual output (Eq. 4). Finally, the output of these capsules is computed using Eq. 3. The output layer capsules compute the posterior probabilities. The transformation matrices are learned by back-propagation.

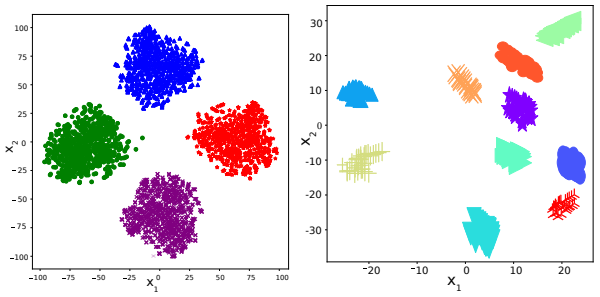

Figure 2: *Scatter plot of the four conv-capsule (left) and ten conv-capsule (right) outputs for RAVDESS after dimensionality reduction via t-SNE. Each color represents one cluster.*

Table 1: *UA(%) on FAU speaker independent scenario (Mont/Ohm as train/test) and vice versa*

Method	Train	Test	2-class UA(%)	8-class UA(%)
Supervector+SVM	Mont	Ohm	62.8	29.8
	Ohm	Mont	56.5	36.3
Supervector+CNN	Mont	Ohm	68.0	53.6
	Ohm	Mont	70.8	58.7
eGeMAPS+CNN	Mont	Ohm	61.7	42.3
	Ohm	Mont	68.2	55.7
(*Feature**) + Capsule	Mont	Ohm	70.5	53.3
	Ohm	Mont	71.3	59.0
(*Feature**)+BLSTM	Mont	Ohm	71.7	53.4
	Ohm	Mont	72.2	58.7
(*Feature**) + Proposed Framework	**Mont**	**Ohm**	**74.5**	**55.3**
	Ohm	**Mont**	**75.3**	**61.8**
(*Feature**)+ Capsule + BLSTM	Mont	Ohm	70.4	53.8
	Ohm	Mont	71.2	58.6
(*Feature**)+ BLSTM + CNN	Mont	Ohm	72.1	54.4
	Ohm	Mont	72.9	58.8

4. Experimental Results

4.1. Features

The eGeMAPS [29] and supervector [3] features were used in the baseline systems, the default parameters reported in their respective publications were applied. We have also used the log-spectrogram feature with 128 filterbanks (*FB*128). The feature vector consists of the fundamental frequency (F_0), 23-dimensional MFCC and log-energy augmented by delta and delta-delta, denoted by *Feature**. To further enrich the input of the DNNs with contextual information, each frame's feature vector was appended with the feature vectors of the preceding/following 45 frames. This paves the way for better capturing the mid to long-term properties of the speech through processing a context of about 900 ms. Networks were trained by PyTorch [30] and optimisation was done by Adam [31].

4.2. Setup

FAU-Aibo [12–14] and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) databases [32] have been used. The RAVDESS is an audio-visual database and only its speech part is utilised here which covers eight acted emotional expressions: neutral, calm, happy, sad, angry, fearful, surprise and disgust while FAU consists of five emotional classes: anger, emphatic, neutral, positive and rest (other categories). FAU consists of children speech recordings who were communicating with Sony's pet robot Aibo, so the emotions are natural and spontaneous. FAU consists of two sets, namely Ohm and

Table 2: *UA(%) on RAVDESS speaker independent scenario*

Method	2-class UA(%)	8-class UA(%)
FB128 **+ Proposed Framework**	**66.3**	**50.1**
Feature* **+ Proposed Framework**	**70.4**	**56.2**
COVAREP + LSTM [33]		41.2

Table 3: *Performance on RAVDESS (75%/25% for train/test)*

Method	2-class UA(%)	8-class UA(%)
Supervector+SVM	65.8	36.3
Supervector+CNN	65.9	34.6
eGeMAPS+CNN	71.4	33.0
(*Feature**) + Capsule	60.3	25.5
(*Feature**)+BLSTM	74.2	63.9
(*Feature) + Proposed Framework**	**79.5**	**69.4**
(*FB*128) + Proposed Framework	**73.9**	**68.1**
(*Feature**)+ Capsule + BLSTM	66.8	35.4
($F_0 + MFCC^\star$)+ BLSTM + CNN	74.8	51.3

Mont which cover 55% and 45% of the whole data, respectively, with totally disjoint speakers.

For training the system, 75% of data (randomly chosen) were employed, and the remaining 25% were used for testing. The RAVDESS speaker independent scenario is performed by using 19 speakers for training and four different speakers for testing. For FAU corpus, we also followed another approach: since it consists of two subsets, i.e. Ohm and Mont, one was used for train and the other one to test. These two sets are disjoint in terms of speakers which makes the test condition more challenging than 75/25% case and provides a better platform for evaluating the robustness of the system. The downside, however, is that a lower amount of data becomes available for training. In the second approach, namely choosing 75/25% for the train/test, we have run the experiments ten times and reported the mean. The baseline systems trained with different hyperparameters, and the results have been reported for comparison purposes. No transfer learning mechanism was used and the classifiers were trained from scratch.

4.3. Results and Discussion

We hypothesized at the beginning of the paper that each 1D conv-capsule would learn different temporal properties for the same BLSTM temporal activation. The output of those capsules are extracted and after dimensionality reduction they are plotted in Fig 2. Each of these capsules learn totally different and unique temporal features. As can be observed, they form non-overlapping clusters for both cases of four and ten capsules.

Tables 1-4 show the unweighted accuracy (UA) for binary (positive vs negative) as well as 5-class (FAU) and 8-class (RAVDESS) emotion classification tasks. The proposed approach in comparison with different systems and baselines leads to a notably better performance. To visualise how well the proposed network separates classes, we performed a dimensionality reduction using t-SNE [34] on the output layer for RAVDESS test and train data. Fig. 3 illustrates the network successfully clusters the representations in the output layer.

The combination of the supervector (as input) with SVM and CNN is shown for FAU corpus in Table 1. SVM (with RBF basis) is outperformed by most of the DNN-based back-ends. Comparing the Capsule net with CNN in combination with BLSTM shows the superiority of the Capsule network which can be explained considering the advantages of the routing-by-agreement process over the max-pooling, as explained in Section 2.3. It should be mentioned that the training time for the capsules on our system was noticeably higher than CNN.

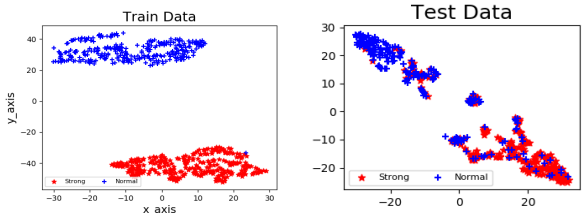

Figure 3: *Scatter plot of the proposed system's output for RAVDESS in 2-class task (strong/normal emotion) after dimensionality reduction via t-SNE.*

Table 4: *FAU (75%/25% for train/test) in binary emotion classification task.*

Method	**Proposed Framework**	DBN [35]	Sparse AE+ SVM [35]
UA(%)	**77.6** $\pm 0.2\%$	74.1	71.7

As seen in Table 1, the order of the BLSTM and the Capsule nets is also important and using the BLSTM on top of the Capsule net obviously degrades the performance and leads to a sub-additive combination. This can be explained based on the argument put forward in Section 3. Table 2 shows the results for RAVDESS database in which similar trends can be observed in terms of the ranking of the different systems.

To the best of our knowledge, the state-of-the-art accuracy for FAU (2-class) is 74.1% [35]. Latif et al. [35] used deep belief network (DBN) and randomly selected 75% of the data for training purpose and 25% of the data for testing purpose. To do a fair comparison, similar to their approach, the data was divided into splits (75/25%) and ran ten times and our result is the average of ten runs. As seen in Table 4, the proposed hybrid architecture leads to 77.6% $\pm 0.2\%$ accuracy (standard deviation $\pm 0.2\%$) for FAU, which is 3.5% (absolute) higher than the state-of-the-art performance.

The state-of-the-art accuracy for 8-class RAVDESS audio speech emotion classification is 41.2% [33] while the proposed system leads to remarkably higher accuracy of 56.2%.

5. Conclusion

In this paper, a hybrid architecture consisting of BLSTM, 1D conv-capsule and capsule routing layers was proposed for speech emotion recognition. The BLSTM is tasked with handling the temporal dynamic of the speech as a sequential data and extracting contextually-rich representations through forward/backward processing of the short-term features. The Capsule layers provide a state-of-the-art system for further distilling and processing the patterns extracted by the BLSTM. This structure results in a hierarchical temporal modelling that facilitates information clustering and categorisation. The proposed architecture was compared with a wide range of alternative networks and the state-of-the-art performance was achieved. Applying this architecture to language and speaker recognition tasks is recommended for future research.

6. References

[1] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 2, pp. 401–404, 2003.

[2] H. Cao, R. Verma, and A. Nenkova, *Computer Speech and Language*.

[3] E. Loweimi, M. Doulaty, J. Barker, and T. Hain, "Long-Term statistical feature extraction from speech signal and its application in emotion recognition," in *Lecture Notes in Computer Science*, vol. 9449, 2015.

[4] B. Schuller, S. Steidl, and A. Batliner, "A.: The interspeech 2009 emotion challenge," in *In ISCA, ed.: Proceedings of Interspeech*, 2009, pp. 312–315.

[5] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec 2013, pp. 216–221.

[6] S. Mirsamadi, E. Barsoum, and C. 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.

[7] S. Zhang, T. Huang, and W. Gao, "Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval - ICMR '16*, 2016, pp. 281–284.

[8] K. Kawaguchi, L. Pack Kaelbling, and Y. Bengio, "Generalization in Deep Learning," *ArXiv e-prints*, Oct. 2017.

[9] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 801–804.

[10] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Dec 2016, pp. 1–4.

[11] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Deep temporal models using identity skip-connections for speech emotion recognition," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: ACM, 2017, pp. 1006–1013.

[12] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009, pp. 312–315.

[13] S. Steidl, "Automatic classification of emotion related user states in spontaneous children's speech," 2009.

[14] A. Batliner, S. Steidl, C. Hacker, and E. Nöth, "Private emotions versus social interaction: A data-driven approach towards analysing emotion in speech," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1-2, pp. 175–206, Feb. 2008.

[15] S. Livingstone and F. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018.

[16] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition." ISCA - International Speech Communication Association, September 2015.

[17] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM networks," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 4, 2005, pp. 2047–2052.

[18] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 06, no. 02, pp. 107–116, 1998.

[19] S. Hochreiter and J. U. Schmidhuber, "LONG SHORT-TERM MEMORY," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] M. Sundermeyer, R. Schl, and H. Ney, "LSTM Neural Networks for Language Modeling," *Proc. Interspeech*, pp. 194–197, 2012.

[21] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[22] 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 *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5200–5204.

[23] D. Palaz, M. Magimai-Doss, and R. Collobert, "Analysis of cnn-based speech recognition system using raw speech as input," in *INTERSPEECH*, 2015.

[24] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec 2014.

[25] Y. LeCun and Y. Bengio, "The handbook of brain theory and neural networks," M. A. Arbib, Ed. Cambridge, MA, USA: MIT Press, 1998, ch. Convolutional Networks for Images, Speech, and Time Series, pp. 255–258.

[26] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, 2017, pp. 3859–3869.

[27] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, p. 2000.

[28] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 4, pp. 788–798, 2011.

[29] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, 2016.

[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[32] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018.

[33] R. Beard, R. Das, R. W. M. Ng, P. G. K. Gopalakrishnan, L. Eerens, P. Swietojanski, and O. Miksik, "Multi-modal sequence fusion via recursive attention for emotion recognition," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 251–259.

[34] L. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-SNE," *Journal of Machine Learning Research*, 2008.

[35] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Cross corpus speech emotion classification- an effective transfer learning technique," *CoRR*, vol. 1801.06353, 2018.