SPEECH EMOTION RECOGNITION WITH DUAL-SEQUENCE LSTM ARCHITECTURE

Jianyou Wang[1], *Michael Xue*[1], *Ryan Culhane*[1], *Enmao Diao*[1], *Jie Ding*[2], *Vahid Tarokh*[1]

[1] Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA
[2] School of Statistics, University of Minnesota Twin Cities, Minneapolis, MN, USA

ABSTRACT

Speech Emotion Recognition (SER) has emerged as a critical component of the next generation of human-machine interfacing technologies. In this work, we propose a new dual-level model that predicts emotions based on both MFCC features and mel-spectrograms produced from raw audio signals. Each utterance is preprocessed into MFCC features and two mel-spectrograms at different time-frequency resolutions. A standard LSTM processes the MFCC features, while a novel LSTM architecture, denoted as Dual-Sequence LSTM (DS-LSTM), processes the two mel-spectrograms simultaneously. The outputs are later averaged to produce a final classification of the utterance. Our proposed model achieves, on average, a weighted accuracy of 72.7% and an unweighted accuracy of 73.3%—a 6% improvement over current state-of-the-art unimodal models—and is comparable with multimodal models that leverage textual information as well as audio signals.

Index Terms— Speech Emotion Recognition, Mel-Spectrogram, LSTM, Dual-Sequence LSTM, Dual-Level Model

1. INTRODUCTION

As the field of Automatic Speech Recognition (ASR) rapidly matures, people are beginning to realize that the information conveyed in speech goes beyond its textual content. Recently, by employing deep learning, researchers have found promising directions within the topic of Speech Emotion Recognition (SER). As one of the most fundamental characteristics that distinguishes intelligent life forms from the rest, emotion is an integral part of our daily conversations. From the broad perspective of general-purposed artificial intelligence, the ability to detect the emotional contents of human speech has far-reaching applications and benefits. Furthermore, the notion that machines can understand and perhaps some day produce emotions can profoundly change the way humans and machines interact.

Previous work in SER models on the benchmark IEMOCAP dataset [1] can be generally divided into two categories: unimodal and multimodal. Research that focuses on unimodal data uses only raw audio signals, whereas research in multimodal data leverages both audio signals and lexical information, and in some cases, visual information. Not surprisingly, since they take advantage of more information, multimodal models generally outperform unimodal models by 6-7%. Traditionally, unimodal models extract high level information from raw audio signals, such as MFCC features, and then pass the output through a recurrent neural network [2]. Recently, researchers have begun transforming raw audio signals into spectrograms or mel-spectrograms [3, 4], which contain low level information and can be converted back to raw audio. These spectrograms are then mapped into a latent time series through several convolutional layers before going through a recurrent layer.

Some researchers think that audio data alone is not enough to make an accurate prediction [5], and thus many have turned to using textual information as well. However, it is possible that two utterances with the same textual content can have entirely different meanings when fueled with different emotions. Therefore, using textual information too liberally may lead to misleading predictions. It is our opinion that the full potential of audio signals has not been fully explored, and we propose several changes to the existing state-of-the-art framework for unimodal SER [6, 7].

In this paper, we make three major contributions to the existing unimodal SER framework. First, we propose a new dual-level model that contains two independent neural networks that process the MFCC features and mel-spectrograms separately, but are trained jointly. Similar to other dual-level architectures [8], we found that our proposed dual-level model provides a significant increase in accuracy. Second, inspired by the time-frequency trade-off [9], from each utterance we calculate two mel-spectrograms of different time-frequency resolutions instead of just one. Since these two spectrograms contain complementary information—namely, one has a better resolution along the time axis and the other has a better resolution along the frequency axis—we propose a novel variant of LSTM [10], denoted as Dual-Sequence LSTM (DS-LSTM), that can process these two sequences of data simultaneously and harness their complementary information effectively. It should be noted that previous research in multi-dimensional LSTM (MD-LSTM) [11, 12, 13], especially in ASR [14, 15], focused on adapting the LSTM to

This work was supported in part by Office of Naval Research Grant No. N00014-18-1-2244.

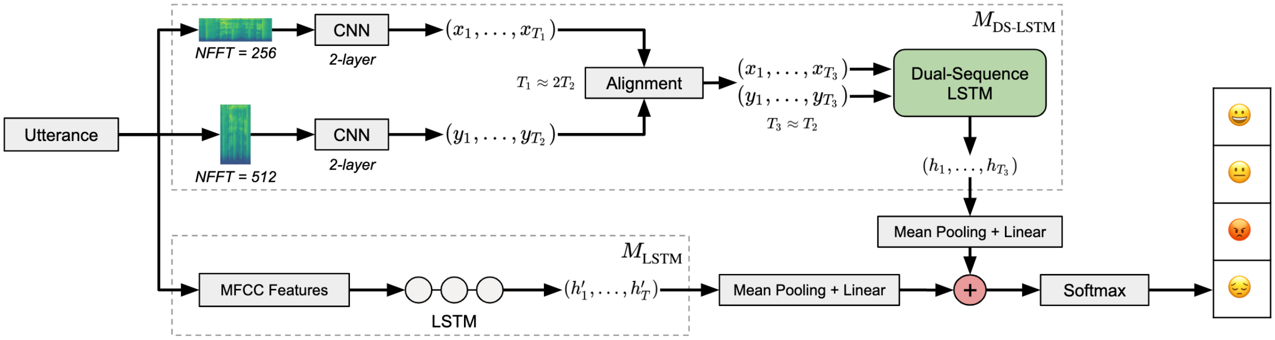

Fig. 1: Dual-level model with DS-LSTM cell

a multi-dimensional data format. Although similar in concept, our proposed DS-LSTM has a distinct architecture, and is designed to process two sequences of one-dimensional data instead of multi-dimensional data. Third, we propose a novel mechanism for data preprocessing that uses nearest-neighbor interpolation to address the problem of variable lengths between different audio signals. We have found that interpolation works better than more typical methods such as truncating and padding data, which lose information and also increase the computational cost.

2. RESEARCH METHODOLOGY

2.1. Dataset Description

We used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [1] in this work, a benchmark dataset containing about 12 hours of audio and video data, as well as text transcriptions. The dataset contains five sessions, each of which involves two distinct professional actors conversing with one another in both scripted and improvised manners. In this work, we utilize data from both scripted and improvised conversations, as well as only audio data to stay consistent with the vast majority of prior work. We also train and evaluate our model on four emotions: *happy*, *neutral*, *angry*, and *sad*, resulting in a total of 5531 utterances (*happy*: 29.5%, *neutral*: 30.8%, *angry*: 19.9%, *sad*: 19.5%). We denote these 5531 utterances in the set $\{u_1, \ldots, u_{5531}\}$.

2.2. Preprocessing

For extracting MFCC features, we used the openSMILE toolkit [12], a software that automatically extracts features from an audio signal. Using the MFCC12_E_D_A configuration file, we extracted 13 Mel-Frequency Cepstral Coefficients (MFCCs), as well as 13 delta and 13 acceleration coefficients, for a total of 39 acoustic features. These features are extracted from 25 ms frames, resulting in a sequence of 39-dimensional MFCC features per utterance $u_i \in \{u_1, \ldots, u_{5531}\}$.

For each utterance, we also propose to derive two mel-spectrograms of different time-frequency resolutions instead of just one, as done in previous research. One (denoted by S_{1i}) is a mel-scaled spectrogram with a narrower window and thus a better time resolution, while the other (denoted by S_{2i}) is a mel-scaled spectrogram with a wider window and thus a better frequency resolution. In our work, S_{1i} and S_{2i} are calculated from a short-time Fourier transform with 256 and 512 FFT points, respectively. The hop length and the number of mel channels are 50% and 25% of the number of FFT points, respectively.

The standard method to deal with variable length in utterances is padding or truncation. Since there are rises and cadences in human conversation, we cannot assume the emotional contents are uniformly distributed within each utterance. Therefore, by truncating data, critical information is inevitably lost. On the other hand, padding is computationally expensive. We propose a different approach to deal with variable length between utterances: nearest-neighbor interpolation, in which we interpolate along the time axis for each mel-spectrogram to the median number of time steps for all the spectrograms, followed by a logarithmic transformation.

2.3. Proposed Model

2.3.1. *Dual-Level Architecture*

Our proposed dual-level architecture is illustrated in Figure 1. It contains two separate models, M_{LSTM} and $M_{\text{DS-LSTM}}$, the first for the MFCC features and the second for the two mel-spectrograms. Each of these two models has a classification layer, the outputs of which are averaged to make the final prediction. The loss function is also the average of two different cross entropy losses from the two models.

2.3.2. *LSTM for MFCC Features*

The MFCC features for each utterance are represented by $Z = \{z_1, \ldots, z_T\}$, with each $z_i \in \mathbb{R}^{39}$. Each Z is fed into a standard two-layer single-directional LSTM, whose outputs,

$H = \{h'_1, \ldots, h'_T\}$, as specified by Figure 1, are mean pooled before being fed into the final classification layer [2].

2.3.3. *CNN for Mel-Spectrograms*

As mentioned earlier, for each utterance u_i, we produce two mel-spectrograms with different time-frequency resolutions. We pass these two spectrograms into two independent 2D CNN blocks, each of which consist of two convolution and max-pooling layers. After both spectrograms go through the two convolution and max-pooling layers, they have a different number of time steps, one with T_1 and the other with T_2, where $T_1 \approx 2T_2$. Before passing both sequences into the DS-LSTM, we use an alignment procedure to ensure they have the same number of time steps, taking the average of adjacent time steps in the sequence of length T_1. After alignment, both sequences have the same number of time steps T_3, where $T_3 \approx T_2$.

2.3.4. *Dual-Sequence LSTM*

Following the alignment operation, we obtain two sequences of data, $X = \{x_1, \ldots, x_{T_3}\}$ and $Y = \{y_1, \ldots, y_{T_3}\}$, with the same number of time steps. Here, X comes from mel-spectrogram S_{1i}, which records more information along the time axis, and Y comes from mel-spectrogram S_{2i}, which records more information along the frequency axis. It is entirely conceivable that sequences X and Y will complement each other, and therefore it will be beneficial to process them through a recurrent network simultaneously.

As Figure 2 indicates, we propose a Dual-Sequence LSTM (DS-LSTM) that can process two sequences of data simultaneously. Let $\odot$ denote the Hadamard product, $[a, b]$ the concatenation of vectors, σ the sigmoid activation function, $\tanh$ the hyperbolic tangent activation function, and rbn the recurrent batch normalization layer, which keeps a separate running mean and variance for each time step [16].

$$f_t = \text{rbn}(\sigma(W_f[x_t, y_t, h_{t-1}] + b_f)) \tag{1}$$

$$i_{T_t} = \text{rbn}(\sigma(W_{i_T}[x_t, y_t, h_{t-1}] + b_{i_T})) \tag{2}$$

$$i_{F_t} = \text{rbn}(\sigma(W_{i_F}[x_t, y_t, h_{t-1}] + b_{i_F})) \tag{3}$$

$$o_t = \text{rbn}(\sigma(W_o[x_t, y_t, h_{t-1}] + b_o)) \tag{4}$$

$$\tilde{C_{T_t}} = \tanh(W_T[x_t, h_{t-1}] + b_T) \tag{5}$$

$$\tilde{C_{F_t}} = \tanh(W_F[y_t, h_{t-1}] + b_F) \tag{6}$$

$$C_t = f_t \odot C_{t-1} + i_{T_t} \odot \tilde{C_{T_t}} + i_{F_t} \odot \tilde{C_{F_t}} \tag{7}$$

$$h_t = o_t \odot \tanh(C_t) \tag{8}$$

After the execution of (8), h_t is the hidden state for the next time step, but h_t also goes through a batch normalization layer to be the input for the next layer of the DS-LSTM at time t.

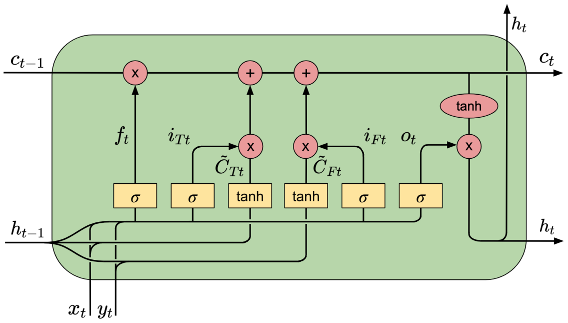

Fig. 2: The graphical representation of one DS-LSTM cell

While an LSTM is a four-gated RNN, the DS-LSTM is a six-gated RNN, with one extra input gate i_{F_t} at (3) and one extra intermediate memory cell $\tilde{C_{F_t}}$ at (6). The two intermediate memory cells $\tilde{C_{T_t}}$ and $\tilde{C_{F_t}}$ are derived from X and Y, respectively, with the intuition that $\tilde{C_{T_t}}$ will capture more information along the time axis, while $\tilde{C_{F_t}}$ will capture more information along the frequency axis. Empirical experiments suggest that the forget gate, two input gates, and output gate should incorporate the maximum amount of information, which is the concatenation of x_t, y_t, and h_{t-1}.

A recurrent batch normalization layer (rbn) is used to normalize the output of the forget gate, input gates, and output gate in order to speed up training and provide the model with a more robust regularization effect.

In summary, Section 2.3.2 describes the vanilla model M_{LSTM}. Sections 2.3.3 and 2.3.4 describe the architecture for our proposed DS-LSTM model, denoted as $M_{\text{DS-LSTM}}$. Together, M_{LSTM}+$M_{\text{DS-LSTM}}$ describes our proposed Dual-Level model as illustrated in Figure 1.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

For the CNN block used to process the mel-spectrograms, a 4×4 kernel is used without padding, and the max pooling kernel is 2×2 with a 2×2 stride. For each layer of the CNN, the output channels are 64 and 16, respectively. All gate neural networks within LSTM and DS-LSTM have 200 hidden nodes. Each LSTM is single-directional with two layers. The weight and bias for the recurrent batch normalization parameters are initialized as 0.1 and 0, respectively, as suggested by the original paper [16]. An Adam optimizer is used with the learning rate set at 0.0001.

3.2. Baseline Methods

Since several modifications are proposed, we create six baseline models that consist of various parts of the whole model in order to better evaluate the value of each modification.

Base 1: M_{LSTM}, which is the LSTM-based model with the MFCC features.
Base 2: CNN+LSTM, whose inputs, $\{S_{11}, \ldots, S_{1n}\}$, are spectrograms with 256 FFT points. Inputs are passed through a CNN followed by an LSTM. Models such as these are developed in [6] and [17].
Base 3: CNN+LSTM, whose inputs, $\{S_{21}, \ldots, S_{2n}\}$, are spectrograms with 512 FFT points. Inputs are passed through a CNN followed by an LSTM. Note that the architecture is the same as **Base 2**.
Base 4: A combination of models of **Base 2** and **Base 3**: 2 × (CNN+LSTM), whose inputs are $\{S_{11}, \ldots, S_{1n}\}$ and $\{S_{21}, \ldots, S_{2n}\}$. In this model, two LSTMs process two sequences of mel-spectrograms separately, and their respective outputs are averaged to make final classifications. Note this is different from our proposed DS-LSTM, which processes these two sequences within a single DS-LSTM cell.
Base 5: A combination of models of **Base 1** and **Base 4**.
Base 6: A combination of models of **Base 1** and **Base 2**.

In addition to the above six baseline models, we propose two models, $M_{\text{DS-LSTM}}$ and the dual-level model, M_{LSTM}+$M_{\text{DS-LSTM}}$. We compare these models with the baseline models, as well as four state-of-the-art models that use standard 5-fold cross-validation for evaluation.

3.3. Results and Analysis

	Mean WA	Mean UA
Base 1 = M_{LSTM}	64.7±1.4	65.5±1.7
Base 2 = CNN+LSTM	63.5±1.6	64.5±1.5
Base 3 = CNN+LSTM	62.9±1.0	64.3±0.9
Base 4 = Base 2 + Base 3	64.4±1.8	65.2±1.8
Base 5 = Base 1 + Base 4	68.3±1.3	69.3±1.2
Base 6 = Base 1 + Base 2	68.5±0.8	68.9±1.2
D. Dai et. al (2019) [18]	65.4	66.9
S. Mao et. al (2019) [19]	65.9	66.9
R. Li et. al (2019) [6]	—	67.4
S. Yoon et. al (2018) [20] *	71.8±1.9	—
Proposed $M_{\text{DS-LSTM}}$	69.4±0.6	69.5±1.1
Proposed M_{LSTM}+$M_{\text{DS-LSTM}}$	72.7±0.7	73.3±0.8

Table 1: Mean WA and Mean UA are the average of weighted accuracy and unweighted accuracy, respectively, for 5-fold cross validation. Most results are reported with one standard deviation.
* indicates the model uses textual information.

Table 1 indicates our proposed model M_{LSTM}+$M_{\text{DS-LSTM}}$ outperforms all baseline models by at least 4.2% in mean weighted accuracy, and by at least 4.0% in mean unweighted accuracy. It also outperforms state-of-the-art unimodal SER models [18, 19, 6] by at least 6.8% in mean weighted accuracy and 5.9% in mean unweighted accuracy. Although multimodal SER models typically have a higher accuracy due to its access to both audio data and textual data, we see that our proposed model achieves comparable performance with [20] in mean weighted accuracy.

Before further investigating the effectiveness of each integrated part of the proposed dual-level model M_{LSTM}+$M_{\text{DS-LSTM}}$, we note that Base 1∼3 and 6 have less parameters than our proposed models. However, we have verified that simply adding more nodes or layers to these models does not make any empirical difference in its predictive power, which suggests that these aforementioned baseline models have already reached their full potential. Therefore, we can objectively compare these models.

Both Base 2 and Base 3 take a single sequence of mel-spectrograms, and both perform slightly worse than Base 1, which only uses MFCC features. This supports the claim that mel-spectrograms are harder to learn than MFCC features. Base 4 is a naive combination of Base 2 and Base 3, and because the two LSTMs in Base 4 do not interact with each other, the complementary information between these two sequences of mel-spectrograms is not fully explored; therefore, Base 4 is also slightly worse than Base 1. Base 5 and Base 6 are both dual-level models that consider both MFCC features and mel-spectrograms, and they both outperform Base 1∼4, demonstrating the effectiveness of the dual-level model.

More importantly, we observe that the proposed $M_{\text{DS-LSTM}}$ significantly outperforms Base 1∼4. Comparing $M_{\text{DS-LSTM}}$ with Base 4, we see that when two separate LSTMs are replaced by the DS-LSTM, which has only six neural networks in its cell instead of eight neural networks in two LSTMs together, the weighted accuracy increases by 5% and the parameters are reduced by 25%. This shows that the DS-LSTM is a successful upgrade from two separate LSTMs. When we consider the dual-level model M_{LSTM}+$M_{\text{DS-LSTM}}$, it outperforms all baseline methods significantly.

4. CONCLUSION

In this paper, we have demonstrated the effectiveness of combining MFCC features and mel-spectrograms produced from audio signals for emotion recognition. Furthermore, we introduced a novel LSTM architecture, denoted as DS-LSTM, which can process two mel-spectrograms simultaneously. We also outlined several modifications to the data preprocessing step. Our proposed model significantly outperforms baseline models and current state-of-the-art unimodal models on the IEMOCAP dataset, and is comparable with multimodal models, showing that unimodal models, which only rely on audio signals, have not reached their full potential.

5. ACKNOWLEDGEMENTS

We are grateful for the assistance of Mr. Reza Soleimani and the support of Professor Robert Calderbank from the Rhodes Information Initiative at Duke University.

6. REFERENCES

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

[2] 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)*, March 2017, pp. 2227–2231.

[3] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *INTERSPEECH*, 2017.

[4] Jianfeng Zhao, Xia Mao, and Lijiang Chen, "Speech emotion recognition using deep 1d and 2d cnn lstm networks," *Biomed. Signal Proc. and Control*, vol. 47, pp. 312–323, 2019.

[5] E. Kim and J. W. Shin, "Dnn-based emotion recognition based on bottleneck acoustic features and lexical features," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6720–6724.

[6] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6675–6679.

[7] S. Yeh, Y. Lin, and C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6685–6689.

[8] I. Choi S. H. Bae and N. S. Kim, "acoustic scene classification using parallel combination of lstm and cnn," *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, pp. 11–15, 2016.

[9] D. Donoho and P. Stark, "Uncertainty principles and signal recovery," *SIAM Journal on Applied Mathematics*, vol. 49, no. 3, pp. 906–931, 1989.

[10] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[11] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber, "Multi-dimensional recurrent neural networks," *ICANN*, 2007.

[12] Alex Graves and Jürgen Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," pp. 545–552, 2009.

[13] Marijn F. Stollenga, Wonmin Byeon, Marcus Liwicki, and Jürgen Schmidhuber, "Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation," *NIPS*, 2015.

[14] Jinyu Li, Abdul Raheem Mohammad, Geoffrey Zweig, and Yifan Gong, "Exploring multidimensional lstms for large vocabulary asr," in *ICASSP*, March 2016.

[15] Bo Li and Tara N. Sainath, "Reducing the computational complexity of two-dimensional lstms," in *INTERSPEECH*, 2017.

[16] Tim Cooijmans, Nicolas Ballas, César Laurent, and Aaron C. Courville, "Recurrent batch normalization," *CoRR*, vol. abs/1603.09025, 2016.

[17] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch, "Cnn+lstm architecture for speech emotion recognition with data augmentation," 2018.

[18] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 7405–7409.

[19] S. Mao, D. Tao, G. Zhang, P. C. Ching, and T. Lee, "Revisiting hidden markov models for speech emotion recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6715–6719.

[20] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," *CoRR*, vol. abs/1810.04635, 2018.