Segment-Based Speech Emotion Recognition Using Recurrent Neural Networks

Efthymios Tzinis
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou 15780, Greece
Email: etzinis@gmail.com

Alexandros Potamianos
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou 15780, Greece
Email: potam@central.ntua.gr

***Abstract*—Recently, Recurrent Neural Networks (RNNs) have produced state-of-the-art results for Speech Emotion Recognition (SER). The choice of the appropriate time-scale for Low Level Descriptors (LLDs) (local features) and statistical functionals (global features) is key for a high performing SER system. In this paper, we investigate both local and global features and evaluate the performance at various time-scales (frame, phoneme, word or utterance). We show that for RNN models, extracting statistical functionals over speech segments that roughly correspond to the duration of a couple of words produces optimal accuracy. We report state-of-the-art SER performance on the IEMOCAP corpus at a significantly lower model and computational complexity.**

1. Introduction

Building emotionally aware Human-Machine Interfaces (HMIs) ultimately relies on the automatic emotion recognition. Understanding the underlying dynamics of the affective cues and integrating cognition in an HMI, is a challenging task. Speech is one of the most common channels for expressing emotion, thus a machine which interacts with humans should adeptly interpret and utilize affective signals from speech.

One of the most demanding problems in Speaker-independent Speech Emotion Recognition (SER) is finding a representative set of emotion features and the optimal time-scale for emotional context extraction. Extensive experimentation has been made for Low Level Descriptors (LLDs) or local features like Mel Frequency Cepstral Coefficients (MFCCs), pitch, energy, quality of voice but also for global features that are calculated as statistics over local features [3]. The fusion of statistical functions over multiple frame-wise features, extracted from the whole utterance, also yields good performance for SER tasks [6]. Significant progress on SER has been made by employing different classifiers [4] by using both types of aforementioned features. Specifically, Hidden Markov Models (HMMs) [1] and Gaussian Mixture Models (GMMs) [2] have been used with local and global features, respectively. Statistical features were also extracted for multimodal utterance-level emotion recognition with Support Vector Machines (SVMs) [5] and gender dependent utterance-level SER with autoencoders [9].

Progressively, deeper architectures which independently learned abstract emotional context from simple local features, were applied. Namely, Deep Neural Networks (DNNs) [15], Extreme Learning Machines (ELMs) [16], [17] and Recurrent Neural Networks (RNNs) [23], [10] were trained on LLDs vectors corresponding to 1 or more consecutive frames. Consequently, the decision time-scale for each emotional state was based on frame level (30ms) or phoneme level (10-30 concatenated LLDs-vectors). However, every frame contains different amount of information, i.e. each frame has different affective saliency [11]. As a result, building a system which focuses only on highly emotion-condense frames, is crucial. Attention based models were build for this cause and firstly introduced for Speech Recognition (SR) in [7] and integrated in SER tasks [8], [18]. An affective saliency model is proposed in [11], where the contribution of local features is weighted via a regression model, trained on acoustic features.

However, the time-scale at which frame-based features are concatenated or statistical segment-based features are extracted has substantial implications for SER performance [20], [21]. Little progress has been made in analyzing the appropriate decision time-scale from which RNNs can explore a more abstract sequential representation either from local or global features.

In this paper, we utilize both feature sets (local and global) for training a Long Short Time Memory (LSTM) unit and investigate the appropriate decision time-scale (frame, phoneme, word or utterance) for each one. In essence, we examine the effect of the selected number of concatenated LLDs-vectors (for local features) and the duration of speech segments from which we obtain the statistics (global features). We propose a novel segment-wise learning approach with global features for resolving the decision of the emotional context. As a result, state-of-the-art performances on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [25] corpus are obtained with a simple LSTM and exploitation of segment-wise statistical features for speaker-independent SER.

The remainder of this paper is organized as follows. In Section 2 we outline previous related work on SER. In Section 3 feature extraction methods and applied models are described. The experimental procedure is demonstrated in Section 4 and its results are evaluated in Section 5. We conclude this study in Section 6.

2. Related Work

In [20], Provost introduced a variable window length approach for capturing emotional speech in sub-utterance scale. For every window, the presence of an affective cue (Emotion Profile (EP)) was estimated and the final emotion flow was modeled with a trajectory which served as input for a HMM. The fusion of EPs and LLDs in a hierarchical correlation model with multiple layers corresponding to each type of feature, was used in [21]. Emotion probabilities of every layer's unit were fed in an SVM classifier.

Recently, DNNs alongside with ELMs were employed [15]. LLDs were extracted from frames and were concatenated in chunks of 265ms. The highest energy segments were fed into a DNN and emotion class posterior probabilities were computed for each one of them. Statistics of these probabilities were fed into the ELM kernel for utterance-level classification, while the model was tested on the IEMOCAP database. Metallinou et al. [10] demonstrated the significance of using Bi-directional LSTMs (BLSTMs) and generally RNNs for integrating long-term temporal context in SER by experimenting in activation-valence space. Lee et al. [16] proposed an RNN-Connectionist Temporal Classification (CTC) schema in which every frame can be assigned to all emotional classes and an extra NULL label for silence frames. They applied ELM over BLSTM and showed significant improvement on the IEMOCAP by relatively boosting the accuracy around 12%.

Contemporary SER work also focuses on classifying emotion by utilizing spectrogram features with minimal speech processing. In [19] Ghosh et al. introduced innovative ways for training a BLSTM, like representation and transfer learning. The former is based on glottal flow signals while the latter learns emotion representation from activation-valence space, both methods were tested on improvised and scripted utterances. In a recent study [23], different deep implementations were compared. Models were trained on spectrogram features and state-of-the-art results on IEMOCAP were reported using Convolutional Neural Network (CNN).

Building upon the tremendous expressiveness of RNNs and the necessity for emphasizing emotionally-dense frames of speech, attention models were deployed for SER. In [8] promising scores were presented on IEMOCAP. An attention-based BLSTM was employed for a better conceptual of emotional sub-utterances and yielded absolute improvement of 1.46% in accuracy from no attention model. Finally, Misramadi et al. [18] extracted frame-wise raw spectral features and LLDs for training a BLSTM. They demonstrated that LLDs outperform raw spectral features and used a weighted pooling layer with attention on top, in order to deal with the voting of emotional-irrelevant frames. In addition, a variety of pooling methods was compared, over the last hidden layer, for the final utterance classification.

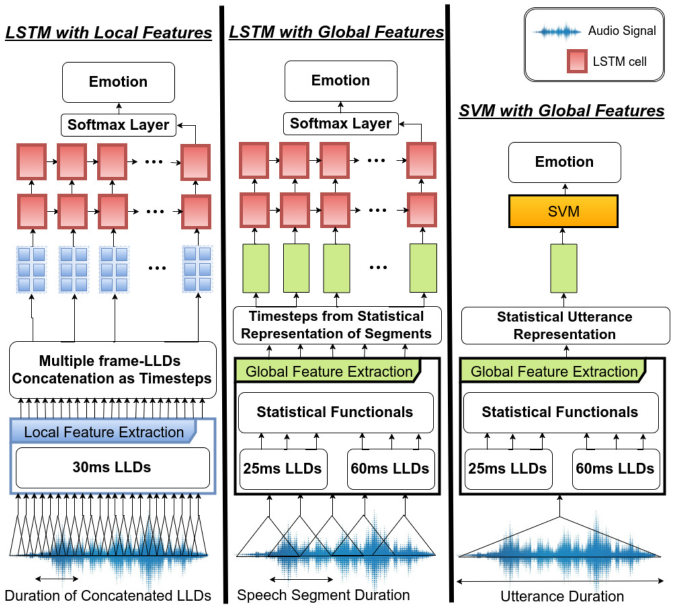

Figure 1. Feature Extraction and Models' Architectures: (Left) LSTM trained on concatenated LLDs, (Middle) LSTM trained on segment global statistical features, (Right) SVM trained on utterance statistical features.

3. Our Approach

The core of our approach lies in investigating emotion decision scales for different feature sets. Consecutive frames with local features can be concatenated and consequently change the emotion decision-scale. For example, consider that a concatenation of 5 frames may represent a *sad* emotion but each frame independently might not. Accordingly, emotion decision-scale is sensitive to the speech segment's duration on which global features are extracted. A speech segment can be considered as a timespan longer than a frame (e.g. the whole speech utterance could serve as a segment). We focus on the impact that number of concatenated frames and segments' duration have on LSTM SER. Thus, we find the appropriate decision time-scale for local and global features, respectively.

First, various Low Level Descriptors (LLDs) are extracted from speech frames (local features) and statistical functionals are computed over them (global features), as described in Sections 3.1 and 3.2, respectively. For the extraction of both local and global features, openSMILE toolkit [12] is used. For all the extracted features, pre-emphasis and smoothing is performed with a low pass filtering when we need to produce the delta values of some attributes. The employed models are described in 3.3. All models' architectures and feature extraction methods are displayed in Figure 1.

3.1. Frame-Wise Local Features

We create 30ms frames with 15ms step size from the audio signal and window them with a Hamming window.

TABLE 1. Local & Global selected features

(RMS) Root Mean Square, (ZCR) Zero Crossing Rate, (HNR) Harmonics to Noise Ratio, (DDP) Difference of Difference of Periods, (LSP) Line Spectral Pairs, (SHS) Sub-Harmonic Sum, (ACF) Autocorrelation Function, (MFB) Mel Frequency Band.

LLDs	1st Delta	Local Features	Global-Features Applied Functional Sets*
RMS Energy	✓	✓	✗
Quality of Voice	✓	✓	✗
ZCR	✓	✓	✗
Jitter Local	✗	✓	A
Jitter DDP	✗	✓	A
Shimmer Local	✗	✓	A
F0 by SHS	✓	✓	A,C
Loudness	✓	✓	A,B
Probability of Voicing	✓	✓	A,B
HNR by ACF	✓	✓	A,B
MFCCs[0-14]	✓	✓	A,B
LSP Frequency [0-7]	✓	✗	A,B
log MFB [0-7]	✓	✗	A,B
F0 Envelope	✓	✗	A,B

**Statistical Functional Sets (A,B,C) are defined in Table 2.*

We extract 24 LLDs for every frame and we also extract the derivatives for all of them (except jitter and shimmer). Selected features per frame are presented in Table 1 under column "Local Features". Now every frame is endowed with 47 attributes. Because of the large number of local features, the most prominent LLDs are selected [16], [18].

3.2. Segment-Wise Global Statistical Features

We follow a similar procedure with Schuller et al. [6] who extracted a robust set of emotional features. The attributes for the corresponding speech segments are based on LLDs which are extracted from frames, (see Section 3.1), enriched with other LLDs. Next, we apply statistical functions on them in order to obtain the global features per segment (see Table 2). Different LLDs require different window sizes for proper extraction. In short, we extracted pitch by using a Gaussian window of 60ms and 10ms step size, while for all the other LLDs we use a Hamming window of 25ms and 10ms step size. The list of the selected LLDs alongside with their statistical functional sets are presented in Table 1 under column "Global-Features Applied Functional Sets". As a result, every speech segment is represented with a fixed-length vector of 1582 features.

3.3. Models

We employ an LSTM RNN as described in [24] and train it with multiple timesteps which correspond either in concatenated LLDs (local features, Figure 1: Left) or statistical representation of speech segments (global features, Figure 1: Middle). These fixed-length vectors of attributes correspond to different timesteps and vary between utterances. For every sequence of timesteps, the expected output is an emotion label (this is often called many-to-one training). RNNs can adequately encode the information enclosed in a sequence of timesteps and produce the expected output on the last timestep. When LSTM layers are stacked, the output of every layer is fed as an input in the subsequent layer. Our LSTM has 2 hidden layers and on top of the final timestep representation there is a dense output layer which leads to an corresponding emotional category. An SVM with a Radial Base Function (RBF) kernel is built, using LibSVM [14] and trained on statistical features over the whole utterance (Figure 1: Right).

TABLE 2. Sets of Statistical Functionals

Statistical Functions	Set
position max/min arithmetic mean, standard deviation skewness, kurtosis linear regression coefficient 1/2 Quadratic & Absolute linear regression error quartile 1/2/3 quartile range 2-1/3-2/3-1 percentile 99 up-level time 75/90	A
percentile 1, percentile range 1-99	B
OnSets Number, Duration	C

4. Experimental Setup

For the experiments we use the IEMOCAP database. This database contains audio and visual data from 5 sessions. In each session 2 people perform an acted or an improvised dialogue. For each utterance, 3 human annotators labeled it with a categorical emotion label. We choose only the utterances for which at least 2 out of 3 annotators had similar opinion. Specifically, our dataset comprises of audio signals from 4 emotional categories: *angry*, *sad*, *happy* and *neutral*. We employ a Leave One Session Out (LOSO) schema for testing our models, same as [8], [16], [19]. In each fold (5 in total) one session is used for test and the remaining 4 for train. For every session we test, one speaker is used for validation set and the other for testing. We repeat the experiment by reversing the validation and test sets. The average accuracy from the two speakers is included for the final assessment. All features are normalized by the global mean and the standard deviation of the features in the training set. Evaluation is performed by using Weighted Accuracy (WA) which is the percentage of correct classification decisions over the test set and Unweighted Accuracy (UA) which is the average of accuracies over each emotional class.

The LSTM has 2 hidden layers with cell sizes of 512 and 256, respectively [1]. A softmax layer for classifying the 4 emotional categories is placed on top of the final timestep. LSTM is regularized by applying a dropout rate of *Dr=0.5* only on the non-feedback connections as described in [27]. LSTM performance is not further increased when we try different dropout rates. The model is optimized by

[1] The performance does not increase when we add extra hidden layers.

minimizing the categorical cross-entropy metric, by using Nadam optimizer [28], which is basically Adam optimizer with Nestorov momentum. Base value of learning rate is set to α_{base}=*0.002*. The number of training epochs is set to 100 but early stopping is applied when the loss function of the validation set does not improve for 10 consecutive epochs. The configuration which yields the best sum of WA and UA on the validation set is selected. Moreover, batch normalization layers are discarded because of their undesirable effects on faster convergence. Batch size is set equal to *400*, in order to fit in the memory of our Graphical Processor Unit (GPU). We implement the LSTM by using Keras [13] over a Theano [26] back-end.

When we concatenate multiple LLDs or create a set of global level features over longer speech-segments, the number of timesteps of our training LSTM becomes significantly small. Interestingly, with a reduced number of timesteps the LSTM converges immensely fast [23]. However, we proportionally adjust the initial learning rate exclusively in order to enforce a finer-tuning approach and avoid plateau convergence.

4.1. LSTM Training with Frame Level LLDs

For every frame we extract the LLDs as described in Section 3.1. Consecutive frame-vectors are concatenated in chunks of different lengths and are fed in the LSTM. The length of each chunk is the number of consecutive frame-vectors that it contains. Evaluation is performed for chunk-lengths that correspond to frame (30ms), phoneme (90ms-300ms) and longer segments (400ms-8sec) time-scales for deducing the emotional label. The learning rate for local extracted LLDs is given by the following equation, where α_{base}=*0.002*:

$$\alpha_{local} = \frac{\alpha_{base}}{Number\ of\ Frames\ in\ chunk} \qquad (1)$$

4.2. LSTM Training with Statistical Features

For this experiment, global-statistical features are extracted as described in Section 3.2. We are using segments of lengths (0.5s-8s) with an overlap ratio of *OL=0.5* between them. The learning rate for global-statistical features is given by the following equation, where α_{base}=*0.002*:

$$\alpha_{global} = \frac{\alpha_{base} \cdot OL}{100 \cdot Segment\ Duration(sec)} \qquad (2)$$

4.3. SVM Utterance Level Statistical Features

As described in Section 3.2 statistical features are extracted from the whole utterance which now serves as the only segment. We use an RBF kernel and regularize gamma coefficient with the number of features for every utterance γ = *1/1582*. Values for cost coefficient C lie in the *[0.001,60]* interval. For every test speaker, the C value which yields the best sum of WA and UA for the corresponding validation speaker is chosen. All other hyperparameters are set to their default values [14].

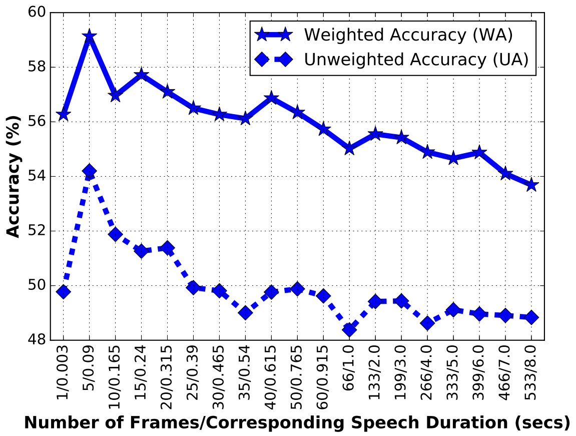

Figure 2. Weighted Accuracy (WA) and Unweighted Accuracy (UA) of LSTM trained on various lengths of concatenated frames with local features.

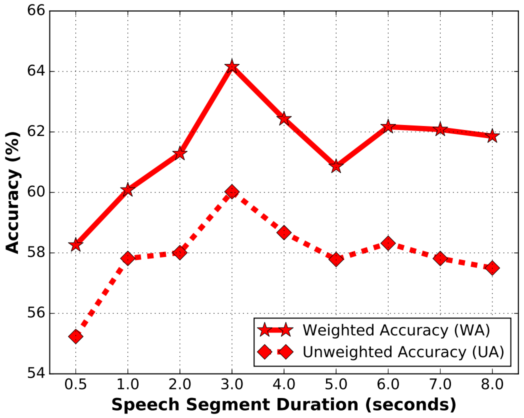

Figure 3. Weighted Accuracy (WA) and Unweighted Accuracy (UA) of LSTM trained on global features over various segments durations.

5. Evaluation Results

5.1. Comparison between Local & Global Features

Results for the local feature training experiment are displayed on Figure 2. It is evident that a concatenation of 5 frames' LLDs yields the best performance in both WA and UA metrics. This corresponds to phoneme time-scale (100ms) decision for the emotional context. As we are increasing the number of frames in every chunk, we change the decision scale to larger scales (1-8s) and notice a gradual decline in both metrics. LLDs misrepresent silent frames of speech and when they are concatenated for long speech durations, they drive the RNN to misleading abstractions for the emotional manifestation of utterances. Moreover, when the decision is made on frame level, silent frames are mistakenly characterized by the same emotion label as higher energy frames.

TABLE 3. ACCURACY OF MODELS IN THE LITERATURE

Model	Type of Features	WA (%)	UA(%)
Best LSTM [23]	Spectrogram	61.71	58.05
BLSTM-SUA [8]	LLDs	59.33	49.96
BLSTM-WPA [18]	LLDs	**63.5**	58.8
BLSTM-ELM [16]	LLDs chunks of 250ms	62.85	**63.89**

Weighted Pooling Attention (WPA), Sub-Utterance Attention (SUA)

TABLE 4. ACCURACY OF PROPOSED MODELS

Model	Type of Features	WA (%)	UA(%)
SVM	Statistical over the whole utterance	53.54	49.23
LSTM	LLDs chunks of 90ms	59.14	54.2
LSTM	Statistical over 3 seconds segments	**64.16**	**60.02**

Deriving higher level statistical functions from multiple LLDs over speech segments, leads to a more salient representation of underlying emotional dynamics. However, when we extract global features from the whole utterance and apply a simple SVM kernel (Table 4), poor results are obtained. Most probably because of the statistical misrepresentation of the whole emotion utterance. Results in Figure 3 demonstrate the effect of training an LSTM with global features over different time-scales. Statistical features on phoneme (0.5s) and utterance (6-8s) time-scales do not perform as well as words (3-4 seconds) time-scales. Speech segments which downscale to phoneme level durations do not enclose sufficient emotional information when we represent them with global statistical features.

Overall, concatenation of multiple LLDs at SER tasks necessitates the existence of a mechanism that confines the influence of non-emotional frames [15], [8]. The need of such a mechanism can be avoided by employing statistical representation on a word time-scale. Intuitively, this resembles the human system of emotional deduction by taking into consideration information from a couple of words.

Our proposed LSTM model, trained on statistical features over a 3 seconds speech segment, obtains a relative improvement of 5.02% in WA (59.14% → 64.16%) and 5.82% (54.2% → 60.02%) in UA from the best performing LSTM trained directly on LLDs and 10.62% in WA (53.54% → 64.16%) and 10.79% (49.23% → 60.02%) in UA from an utterance level decision SVM. See Table 4.

5.2. Comparison with the Literature on IEMOCAP

Comparable literature models are presented in Table 3. Our model surpasses all attention based RNNs: weighted pooling attention BLSTM [18] in both WA and UA by 0.66% and 1.22%, sub-utterance attention based BLSTM [8] in both WA and UA by 4.83% and 10.06%, respectively. Authors in [16] test their model only on improvised scripts which is a subset of the IEMOCAP database and as it is demonstrated in [19] they tend to give higher scores in WA and UA. Despite that, our simple LSTM scores 1.31% higher in WA from their proposed BLSTM-ELM model. Finally, the current state-of-the-art approach in [23] was obtained by exclusively using the final session as a validation set and testing only on the other 4. Nevertheless, their proposed CNN architecture yields very similar results (64.78% WA and 60.89% UA) to ours (64.16% WA and 60.02% UA). Irregardless of the exclusion of sessions, our model clearly outperforms all simple RNN architectures (LSTM, BLSTM) reported in the literature. Comparing with the best RNN-LSTM architecture reported in [23], we obtain a relative improvement of 2.45% in WA and 1.97% in UA.

6. Conclusions

We have shown that the time-scale on which we extract features for SER, has profound effect on RNNs performance. LLDs work well on a time-scale that roughly corresponds to phoneme level. Conversely, statistical features remarkably encode emotional context on a time-scale that corresponds to word level. In our proposed LSTM model, we extract statistical features over segments of speech which correspond to a couple of words and we report state-of-the-art results on the IEMOCAP database, with a much lower computational complexity. In future work, we would like to integrate a mechanism which allows multi-scale decision levels over the emotional sequence and would serve as an input for our RNN. We also want to evaluate our model on other databases.

Acknowledgments

This work has been partially supported by the BabyRobot project supported by EU H2020 (grant #687831).

References

[1] Nogueiras, A., Moreno, A., Bonafonte, A. and Marino, J., B.,“Speech emotion recognition using hidden Markov models.” *INTERSPEECH*, pp. 2679–2682, 2001.

[2] Busso, C., Lee, S., Narayanan, S., “Analysis of emotionally salient aspects of fundamental frequency for emotion detection,” *IEEE Transactions on Audio, Speech and Language Processing,* vol.17, no. 4, pp. 582–596, 2009.

[3] El Ayadi, M., Kamel, M., S., Karray, F., “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[4] Shashidhar, G., K., and Sreenivasa, K., R., “Emotion recognition from speech: a review,” *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[5] Mower, E., Mataric, M., J. and Narayanan, S., “A framework for automatic human emotion classification using emotion profiles,” *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 5, pp. 1057–1070, 2011.

[6] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., “The INTERSPEECH 2010 Paralinguistic Challenge,” *INTERSPEECH*, pp. 2794–2797, 2010

[7] Chorowski, J., K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y., ”Attention-based models for speech recognition,” *Advances in Neural Information Processing Systems*, pp. 577–585, 2015.

[8] Huang, C., W., Narayanan, S., “Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition,” *INTERSPEECH*, pp. 1387–1391, 2016.

[9] Xia, R., Deng, J., Schuller, B. and Liu, Y., “Modeling gender information for emotion recognition using denoising autoencoder,” *ICASSP*, pp. 990–994, 2014.

[10] Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B., and Narayanan, S., “Context-sensitive learning for enhanced audiovisual emotion classification,” *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[11] Chorianopoulou, A., Koutsakis, P., Potamianos, A., “Speech Emotion Recognition Using Affective Saliency,” *INTERSPEECH*, pp. 500–504, 2016.

[12] Eyben, F., Wollmer, M. and Schuller, B., “Opensmile: The munich versatile and fast open-source audio feature extractor,” *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459–1462, 2010.

[13] Chollet F., “Keras,” in *hhtp://keras.io*, 2015.

[14] Chang, C.-C. and Lin, C.-J., “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology,* vol. 2, no. 3, pp. 1–27, 2011.

[15] Han, K., Yu, D. and Tashev, I., “Speech emotion recognition using deep neural network and extreme learning machine,” *INTERSPEECH*, pp. 223–227, 2014.

[16] Lee, J. and Tashev, I., “High-level feature representation using recurrent neural network for speech emotion recognition,” *INTERSPEECH*, pp. 1537–1540, 2015.

[17] Wang, Z.-Q. and Tashev, I., (in press), “Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks,” *ICASSP*, 2017.

[18] Mirsamadi, S., Barsoum, E. and Zhang, C., (in press), “Automatic Speech Emotion Recognition Using Recurrent Neural Networks With Local Attention,” *ICASSP*, 2017.

[19] Ghosh, S., Laksana, E., Morency, L.-P. and Scherer, S., “Representation Learning for Speech Emotion Recognition,” *INTERSPEECH*, pp. 3603–3607, 2016.

[20] Provost, E., M., “Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow,” *ICASSP*, pp. 3682–3686, 2013.

[21] Wu, C., H., Liang, W., B., Cheng, K., C. and Lin, J., C., “Hierarchical modeling of temporal course in emotional expression for speech emotion recognition,” *ACII*, pp. 810–814, 2015.

[22] Huang, C.-W. and Narayanan, S., “Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition,” *INTERSPEECH*, pp. 1387–1391, 2016.

[23] Fayek, H., M., Lech, M. and Cavedon, L., (in press), “Evaluating deep learning architectures for Speech Emotion Recognition,” *Neural Networks*, 2017.

[24] Hochreiter, S. and Schmidhuber, J., “Long short-term memory,” *Neural computation,* vol. 9, no. 8, pp. 1735–1780, 1997.

[25] Busso, C., Bulut, M., Lee, C. and Kazemzadeh, A., “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[26] Bergstra, J., Bastien, F., Breuleux, O., Lamblin, P., Pascanu, R., Delalleau, O., Desjardins, G., Warde-Farley, D., Goodfellow, I., Bergeron, A., et al., “Theano: Deep learning on gpus with python,” *NIPS*, BigLearning Workshop, Granada, Spain, 2011.

[27] Bayer, J., Osendorfer, C., Korhammer, D., Chen, N., Urban, S. and van der Smagt, P., “On fast dropout and its applicability to recurrent networks,” *arXiv preprint arXiv:1311.0701*, 2013.

[28] Dozat, T., “Incorporating Nesterov momentum into Adam,” *Stanford University, Tech. Rep.*, Available: http://cs229.stanford.edu/proj2015/054report.pdf, 2015.