Interspeech 2018
2-6 September 2018, Hyderabad

# Self-Assessed Affect Recognition using Fusion of Attentional BLSTM and Static Acoustic Features

*Bo-Hao Su*[1,2], *Sung-Lin Yeh*[1,2], *Ming-Ya Ko*[1,2], *Huan-Yu Chen*[1,2], *Shun-Chang Zhong*[1,2], *Jeng-Lin Li*[1,2], *Chi-Chun Lee*[1,2]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
cclee@ee.nthu.edu.tw

## Abstract

In this study, we present a computational framework to participate in the Self-Assessed Affect Sub-Challenge in the INTERSPEECH 2018 Computation Paralinguistics Challenge. The goal of this sub-challenge is to classify the valence scores given by the speaker themselves into three different levels, i.e., low, medium, and high. We explore fusion of Bi-directional LSTM with baseline SVM models to improve the recognition accuracy. In specifics, we extract frame-level acoustic LLDs as input to the BLSTM with a modified attention mechanism, and separate SVMs are trained using the standard ComParE_16 baseline feature sets with minority class upsampling. These diverse prediction results are then further fused using a decision-level score fusion scheme to integrate all of the developed models. Our proposed approach achieves a 62.94% and 67.04% unweighted average recall (UAR), which is an 6.24% and 1.04% absolute improvement over the best baseline provided by the challenge organizer. We further provide a detailed comparison analysis between different models.

**Index Terms**: computational paralinguistics, BLSTM, affect recognition, attention mechanism

## 1. Introduction

Computing paralinguistic attributes from speech is becoming more prevalent across a variety of tasks. Aside from focusing solely on automatic speech recognition, modeling speech signals to extract a variety of other relevant attributes of human states and traits (e.g., cold and snoring [1], Alzheimer disease [2, 3], and Autism diagnoses [4], etc.) has sparked many technical research effort - many of these works show that these higher-level human attributes could indeed be estimated from speech signals. The potential application scenario is vast; in fact, a series challenges have been proposed to tackle the issues of robust recognition for different human states and traits. The ComParE 2018 Challenge consists of four sub-challenges as following: Atypical Affect Sub-Challenge, Self-Assessed Affect Sub-Challenge, Crying Sub-Challenge and Heart Beats Sub-Challenge. In this work, we present our algorithm in the participation of Self-Assessed Affect Sub-Challenge.

Many of these real-life tasks suffer naturally from limited data samples, and also the exact mechanism in the manifestation of these attributes in the speech signal is often complex and intertwined with other unwanted factors, e.g., individual idiosyncratic factors, environmental noise, other human attributes and traits, etc. In this work, we focus on affect recognition. Particularly, these are self-assessed affect states (instead of conventionally perceptual-based affect states recognition). In scenarios of mental illness such as depression, the patient's emotion would influence the outcome throughout the therapy process or the morbidity of the illness. If the degradation of the patient's emotion well-being continue to worsen, the patients may even lose the ability to do anything in their daily life. Research has indicated that if a patient's self-assessed affect states improves with therapy, it indeed could create a substantial impact in improving his/her quality of life [5, 6] .

While being an important health indicator, in practice, most of these people tend not to self assess and disclose their own affective states. The ability to automatically sense and detect these individuals' self-assessed valence states using unobtrusive and easily-obtainable behavior signals, such as speech and facial expressions, is becoming more and more important, especially in the health-related applications.. In this work, we present a technical framework in fusing various approaches to achieve robust self-assessed valence attributes recognition from speech. In specifics, we utilize two different types of model: time-series model and static model. The static model is obtained by training SVM classifier using the ComParE_16 feature set with functional encoding (6373 dimensional features). In order to further improve the recall accuracy on the minority class (low), we also train static SVM using ComParE_16 feature set with minority class upsampling.

In terms of the time-series model, we first compute the low-level descriptors part of the ComParE_16 feature set (130 dimensions per frame). We utilize the Bi-directional LSTM as our model, which captures the forward and backward time-dependent acoustic information, to perform affect recognition. We also include the use of attention mechanism together with BLSTM, and we modify the conventional structure of attention weights by inserting a dense layer (fully-connected layer) in the computation of the attention weights for each time step of BLSTM. This additional non-linear transformation of dense layer helps in improving the recognition rates. Finally, the probability outputted obtained from each of these models are averaged to perform the final fused recognition. Overall, our proposed approach achieves a 62.94% and 67.04% unweighted average recall (UAR) in this three class recognition task, which is an 6.24% and 1.04% absolute improvement over the best baseline provided by the challenge organizer in the development and the blind test, respectively.

The rest of this paper is organized as follows. In section 2, we will elaborate the methods used in this work. In section 3, we will present the experimental results and discussions. In the last section, we conclude with future works.

## 2. Methodology

There are multiple components in our proposed approach. We will describe each in the following section.

10.21437/Interspeech.2018-2261

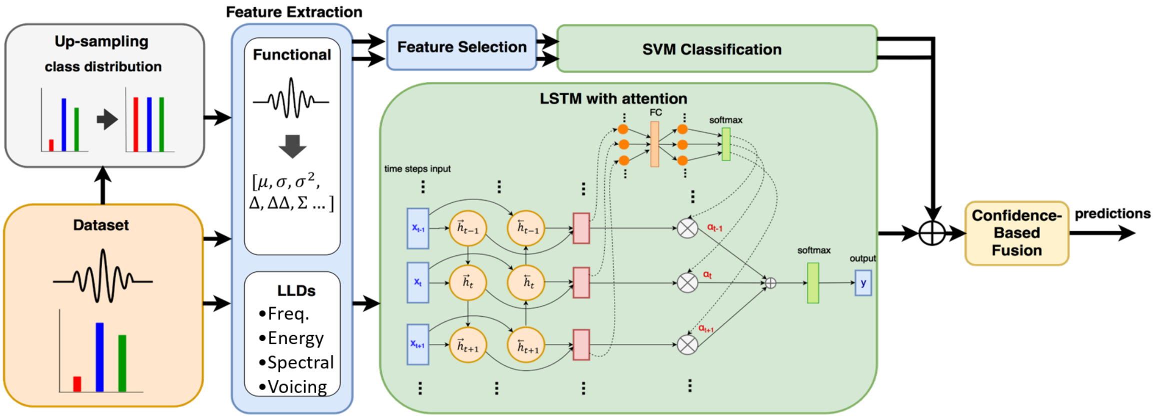

Figure 1: *The complete schematic of our framework: upsampling minority class in our database, training both time-series model (BLSTM with modified attention mechanism) and a static model (SVM with ComParE_16 features), and finally integrating diverse models in a decision-level fusion scheme*

## 2.1. BLSTM with Modified Attention Mechanism

### *2.1.1. Bi-directional LSTM*

Long Short-Term Memory (LSTM) Neural Network is first proposed by Hochreiter et al. [7]. LSTM preserves long term contextual information from data inputs in its hidden state. LSTM is an improvement over recurrent neural network (RNN) by introducing three control gates: input gate, output gate, and forget gate controlling write, read and reset operations for the hidden cells. This helps eliminate the gradient explosion and vanishing gradient problems for RNN. Conventional forward LSTM is uni-directional, i.e., the information can only flow from the past to the future due to the forward propagation of the network structure. Bidirectional LSTM (BLSTM) networks is an improvement over standard forward LSTM model that is capable of operating a sequence of features in both forward and backward directions.

The original LSTM state:

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$

$$h_t = o_t \tanh(c_t)$$

where $\sigma$ is the logistic sigmoid function, and $i$, $f$, $o$ and $c$ are input gate, forget gate, output gate and cell state.

The Bidirectional LSTM state:

$$h_i = [\overrightarrow{h_i} \oplus \overleftarrow{h_i}]$$

Using the combined hidden states allows us to preserve information from both past and future information at any given time step. This particular methodology has been shown to be useful for modeling tasks involving sequence modeling [8]. Another modification to LSTM is Gated Recurrent Unit (GRU) [9]. Similar to LSTM, GRU aims at tracking long-term dependencies effectively to prevent the vanishing/exploding gradient problems. The key difference is that GRU uses only two gates (reset and update gates). The relatively simpler structure of GRU help achieve faster training; however, the trade-off is that GRU remembers only shorter sequences in tasks requiring modeling long-distance relations.

### *2.1.2. Modified Attention Mechanism*

Attention mechanism is a widely used in sequence based encoder-decoder model. Due to the fixed length input vector to the encoder, the encoder-decoder architecture has superior performance on short sequences but not the long ones. As the sequence grows longer, the information contained inside often becomes more complex where a fixed length input vector can no longer support. A simple encoder model results in learning an unreliable representation for such long sequence, leading to poor decoder output. Attention mechanism helps mitigate such an issue by applying weights on the intermediate outputs from each step [10]; in other words, the outputs are generated under a selection mechanism from inputs.

In this work, we also apply an attention mechanism in the building of our time-series BLSTM model. Specifically, the time pooling technique applied to our BLSTM model is performed by computing weighted sum over time [11]. The standard method to use attention mechanism for BLSTM is to choose a simple logistic-regression-like weighted sum as the pooling layer. This weighted sum is the inner product computed between the frame-wise outputs of the BLSTM, $y_t$, and weights $u$ being a vector of parameters as in an attention model. To keep the weight summation as unity, we apply softmax function to the inner product.

$$\alpha_t = \frac{\exp(u^T y_t)}{\sum \exp(u^T y_\tau)}$$

After obtaining the weights, we can calculate the weighted sum over time to get the hidden representation to integrate attention mechanism in our BLSTM.

$$z = \sum \alpha_t y_t$$

In our approach, we modify this attention mechanism by adding a fully-connected layer in the computation of attention, i.e., instead of directly computing dot product between feature output and the label, we enhance the modeling power of attention weights by introducing the use of a more sophisticated nonlinear transformation (see Figure 1 for its network structure). Finally, the newly weighted hidden representation (with modified attention weights, $\alpha_t'$), $z'$, is later fed into another softmax dense layer to compute the final probability of each class. The entire network is jointly optimized over these modules.

$$\alpha_t' = G(w^T \alpha_t + b)$$

$$z' = \sum \alpha_t' y_t$$

Table 1: *A summary of the experiment results for the various model structure, Up-Samp means up-sampling the minority class samples, Aug. means general Data augmentation. The accuracy presented is evaluated on the development set with metric of UAR*

| | Baseline | Model 1 | Up-Samp | Data Aug. | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
|---|---|---|---|---|---|---|---|---|---|
| Low Recall | 37.97 | 24.05 | 54.43 | 18.98 | 29.11 | 18.98 | 53.16 | 19.98 | 29.11 |
| Medium Recall | 60.32 | 74.19 | 51.29 | 67.09 | 59.67 | 65.80 | 72.58 | 51.93 | 45.16 |
| High Recall | 71.10 | 89.51 | 69.97 | 77.05 | 64.87 | 62.88 | 58.38 | 53.54 | 73.37 |
| Average Recall | 56.50 | 64.24 | 57.48 | 54.77 | 51.22 | 49.22 | 61.50 | 48.27 | 49.21 |

where $w$, $b$, $G$ means the weight, bias and activation function of softmax respectively.

### 2.2. Up-Sampling

In the Self-Assessed Affect database, the imbalance of class distribution negatively impacts the recognition accuracy. Resampling is a method to alleviate this problem by balancing class distribution [12]. There are usually two different methods in resampling: up-sampling or down-sampling. Since the database only includes a limited number of utterances, down-sampling while efficient woud result in a loss of modeling power in our models. In our approach, we choose to directly up-sampling (duplicating data samples) the minority class in the database.

### 2.3. Decision Score Fusion

In order to combine various models to obtain a better prediction, we use confidence-based decision-level method, which is similar to decision score fusion to generate our final results [13]. The confidence score from the time-series model is obtained from softmax layer, and the estimated probabilities from the SVM classifications of the static model is used as the confidence score. These confidence scores, i.e., one for each class, predicted from multiple models are then further summed up together. The class with the highest confidence sum is our final prediction for each instance.

## 3. Experimental Results and Discussions

### 3.1. Experimental Setup

We extract standard ComParE features set as our low level descriptors every 10 msec. These low-level descriptors are used in the BLSTM model, which consisting 130 dimensions. This feature set includes voicing, energy, spectral related features and their derivatives [14]. The functionals of these LLDs are regarded as the static acoustic representation for SVM model, and the learned output from attention BLSTM with the LLDs as inputs are the time-series model.

The architectures of our BLSTM models are: a bidirectional LSTM layer with 64 cells (32 for each direction) followed by a fully connected layer with 64 nodes. The activation function is ReLU, and 50% of dropout [15] is utilized to prevent over-fitting, which is applied to the fully-connected layer. The parameters of BLSTM models are optimized using learning rate of 0.0005, batch size as 256 and gradient clipping as 1 to limit the magnitude of the gradient during training process. We conduct and compare our recognition results with the following list of models, and all of the evaluation results are computed on the development set using the metric of unweighted average recall (UAR):

- Model 1 : SVM
- Mdoel 2 : BLSTM method with Attention
- Model 3 : B-GRU method with Attention
- Model 4 : BLSTM + Modified Attention
- Model 5 : Input_Fc + BLSTM + Attention
- Model 6 : Input_Fc + BLSTM + Modified Attention

The Input_Fc means that the inputted low-level descriptors are passed though a fully-connected layer before feeding it into the BLSTM training.

### 3.2. Experimental Results

Table 1 summarizes the performances of each model. In short, two classification models are used in our work, which is SVM and BLSTM. We observe that SVM is better at the medium and high class recall but performs poorly on the low class. Up-sampling data when classified using SVM helps improve the recall rate on low. The BLSTM method, on the other hand, performs well on low and medium class but not on high class.

Due to the difference in the these modeling characteristics, we propose the fusion models of static and time-series model. The final fusion model used, determined empirically as:

- **Fusion : Model 1 + Up-sampled Model 1 + Model 4**

After fusing these three models (SVM, SVM-with-Upsample, BLSTM-Modified-Attention), we obtain the best recognition rates. The confusion matrix of this model on the development set is shown in Figure 2. In summary, our best fused model obtains a 62.94% and 67.04% unweighted average recall (UAR) in the three-class recognition tasks of Self-Assessed Affect task(as shown in Table 2). We obtain an 6.24% and 1.04% absolute improvement over the best baseline provided by the organizer.

### 3.3. Model Comparison and Analysis

In this section, we provides various comparison between different models used in our work.

#### *3.3.1. Model 1 v.s. Baseline*

The Baseline model uses ComParE_2016 functional features to train a linear SVM model. The imbalance class distribution in this sub-challenge leads to worse classification on minority class (low). From Table 1, we observe an increased

Table 2: *Comparison between baseline model and our best fused model (Model 1 + Up-sampled Model 1 + Model 4)*

| | Baseline | Our best fused model |
|---|---|---|
| Dev UAR | 56.7% | 62.94% |
| Test UAR | 66.0% | 67.04% |

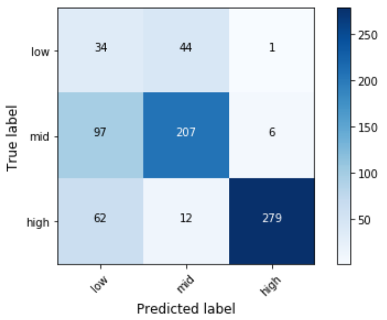

Figure 2: *Confusion Matrix of the Best Fused Model on the Development Set*

improvement for UAR of class Low (24.05% to 54.43%) by up-sampling method. However, the UAR scores of class high and medium drop slightly compared with the original method without the up-sampling method. Note that there is a trade-off between low and medium/high performance. Finally, data-augmentation means to generate data samples (not specific to a particular class) by corrupting original data samples with Gaussian noise. This methodology introduces more noises into our dataset and effectively decrease the recognition accuracy.

### *3.3.2. Model 2 v.s. Model 3*

We further compare the performance between bi-directional GRU and bi-directional LSTM with a standard attention mechanism in each model. While the GRU cells show faster convergence rate during training process, the model with BLSTM cells obtains 2% to 3% higher UAR in average compared to bi-directional LSTM. The bidirectional LSTM with an attention layer achieves not only a high UAR of 61.5% but also shows better performance in both low and medium class recall rates.

### *3.3.3. Extension of Fully-Connected Layer*

The effect of using additional fully-connected layer in our recognition architecture is also analyzed.

- Model 2 v.s. Model 4

The use of dense layer in the computation of attention weights brings about 5% to 8% improvements in the UAR when comparing BLSTM using modified attention versus BLSTM using standard attention mechanism.

- Model 2 v.s. Model 5

In this comparison, we examine the difference of recognition rates obtained by placing the fully-connected layer in the attention weight computation or right after the input LLDs before feeding them into BLSTM. Model 5 shows an decrease in the recall rate in the low class around 10%, which indicates that the fully-connected layer should be placed in the attention mechanism not directly at the input space.

- Model 5 v.s. Model 6

By comparing between Model 5 and 6, we see that by adding additional dense layer is indeed beneficial in obtaining the higher recognition rates. Although, in general, these two models do not perform well due to the initial dense layer applied to the inputs before feeding into the BLSTM.

## 4. Conclusions and Future Works

In this work, we present our recognition framework in the participation of the Self-Assessed Affect Challenge. Our framework is composed of two parts: a standard utterance level baseline ComParE_16 features with SVM trained on original database and up-sampled database, and a BLSTM model with a novel modified attention mechanism. In order to alleviate the issue of data imbalance, we employ a straightforward up-sampling technique. This framework achieves an improved recognition rates for both the development set and the blind testing set. The introduction of modified attention mechanism, i.e., adding a fully-connected layer in the computation of attention weights, is beneficial in improving utilizing sequence based model in affect recognition from speech.

In our future work, we will continue to investigate advanced methods in integrating static-dynamic acoustic representation and learning model for complex human states and trait recognition. Since many of these higher-level internal states and traits are often complexly manifested in the recorded behavior signals, additional technical endeavor is required develop automatic system in consistently and reliably tracking these attributes. The continuous advancement in computational paralinguistics (e.g., complex affective phenomenon recognition) will further help create a tangible impact, especially relevant on applications domains of affective disorders and other related mental health.

## 5. References

[1] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017, pp. 3442–3446.

[2] J. Drapeau, N. Gosselin, L. Gagnon, I. Peretz, and D. Lorrain, "Emotional recognition from face, voice, and music in dementia of the alzheimer type," *Annals of the New York Academy of Sciences*, vol. 1169, no. 1, pp. 342–345, 2009.

[3] G. A. Gates, A. Beiser, T. S. Rees, R. B. D'agostino, and P. A. Wolf, "Central auditory dysfunction may precede the onset of clinical dementia in people with probable alzheimer's disease," *Journal of the American Geriatrics Society*, vol. 50, no. 3, pp. 482–488, 2002.

[4] J. I. Alcántara, E. J. Weisblatt, B. C. Moore, and P. F. Bolton, "Speech-in-noise perception in high-functioning individuals with autism or asperger's syndrome," *Journal of Child Psychology and Psychiatry*, vol. 45, no. 6, pp. 1107–1114, 2004.

[5] S. K. Mittal, L. Ahern, E. Flaster, J. K. Maesaka, and S. Fishbane, "Self-assessed physical and mental function of haemodialysis patients," *Nephrology Dialysis Transplantation*, vol. 16, no. 7, pp. 1387–1394, 2001.

[6] Y. Benyamini, E. L. Idler, H. Leventhal, and E. A. Leventhal, "Positive affect and function as influences on self-assessments of health: Expanding our view beyond illness and disability," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 55, no. 2, pp. P107–P116, 2000.

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

[8] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[11] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2227–2231.

[12] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[13] A. Sinha, H. Chen, D. Danu, T. Kirubarajan, and M. Farooq, "Estimation and decision fusion: A survey," *Neurocomputing*, vol. 71, no. 13-15, pp. 2650–2656, 2008.

[14] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.