Understanding Speaking Styles of Internet Speech Data with LSTM and Low-resource Training

Xixin Wu, Zhiyong Wu, Yishuang Ning, Jia Jia,
Lianhong Cai
Tsinghua-CUHK Joint Research Center for Media Sciences,
Technologies and Systems, Shenzhen Key Laboratory of
Information Science and Technology
Graduate School at Shenzhen, Tsinghua University
Shenzhen, China
xixinwood@gmail.com, ningys13@mails.tsinghua.edu.cn,
zywu@se.cuhk.edu.hk, {jjia,clh-dcs}@tsinghua.edu.cn

Helen Meng
Department of Systems Engineering and Engineering
Management
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong SAR, China
hmmeng@se.cuhk.edu.hk

***Abstract*—Speech are widely used to express one's emotion, intention, desire, etc. in social network communication, deriving abundant of internet speech data with different speaking styles. Such data provides a good resource for social multimedia research. However, regarding different styles are mixed together in the internet speech data, how to classify such data remains a challenging problem. In previous work, utterance-level statistics of acoustic features are utilized as features in classifying speaking styles, ignoring the local context information. Long short-term memory (LSTM) recurrent neural network (RNN) has achieved exciting success in lots of research areas, such as speech recognition. It is able to retrieve context information for long time duration, which is important in characterizing speaking styles. To train LSTM, huge number of labeled training data is required. While for the scenario of internet speech data classification, it is quite difficult to get such large scale labeled data. On the other hand, we can get some publicly available data for other tasks (such as speech emotion recognition), which offers us a new possibility to exploit LSTM in the low-resource task. We adopt retraining strategy to train LSTM to recognize speaking styles in speech data by training the network on emotion and speaking style datasets sequentially without reset the weights of the network. Experimental results demonstrate that retraining improves the training speed and the accuracy of network in speaking style classification.**

Keywords—speaking style; long short-term memory; recurrent neural network; retraining

I. INTRODUCTION

As social network grows rapidly, large scale of multimedia data are generated, including text, speech and image data [1]. Compared to text and image, speech is a more natural way to express one's emotion, intention and desire in human social communication [2]. Understanding the speaking style of these speech data is an important problem in social multimedia research. Unlike conventional speech data, social speech data are uttered by different users with different speaking style categories. Traditional methods try to classify such data with training data in carefully predefined categories. It is time-consuming and requires much human labor to label data with predefined categories. The labeler should have background on psychology and speech processing. The speech data from the Internet are of rich speaking styles and are accessed easily. They provide a rich resource for training speaking style classification models. Nevertheless, they are raw data needed to be prepared carefully. In previous researches, unsupervised methods, including clustering [3], principal component analysis (PCA) [4], are applied. However, the classification accuracy is still very low. It is also a common method to use statistics of the acoustic feature value of the frames as the feature for the whole utterance. Nevertheless, this method may lose some local context information, which plays important role in recognizing different speaking styles. As shown in Figure 1, the two figures of spectrum (as well as F0 and energy contours) are with the same content while with different speaking styles. The one on the top is in question intonation and the bottom one is in neutral intonation. The F0 contour in the top one highlighted in the red rectangle shows the gradually increasing trend. When we use the global feature of the whole speech recording, like the statistics of acoustic feature value, we may lose such local information of speaking styles.

Recently, bidirectional long short-term memory (BLSTM) has achieved exciting successes in lots of research areas, such as handwriting recognition [5] and speech recognition [6]. BLSTM is an extended architecture of RNN. It replaces each hidden layer units of RNN with a memory block, which can store context information. The special architecture of BLSTM allows it to retrieve both past and future context information. To utilize local context information, we apply BLSTM to classify speech data with different speaking styles. However, training BLSTM needs huge number of labeled data. In most circumstances, it is difficult to obtain enough labeled data. In [7], the method of retraining is adopted to improve the adaptability of long short term memory (LSTM). It tries to train a LSTM to recognize digit in speech data. The LSTM is first trained with digit data spoken by girls and then trained with digit data spoken by men. Experimental results show that the retraining method can improve the training speed and accuracy of the network. [9] tries to train BLSTM to recognize speech. First trained with speech data without noise, the network is then trained with the data with noise. [7] and [9] are

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

IEEE computer society

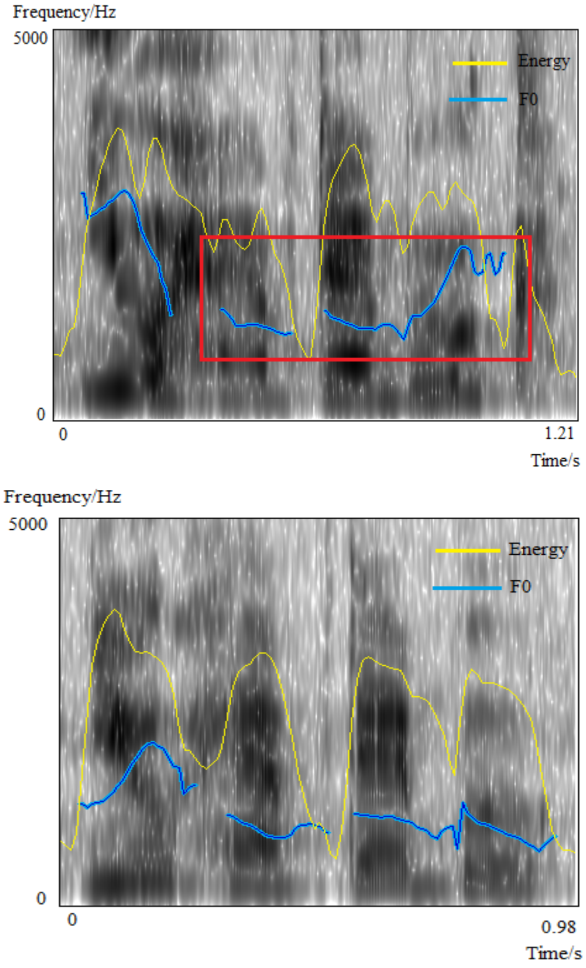

Fig. 1. The spectrum as well as the F0 and energy contours of the speech recordings with different intonations. Top figure comes from the speech with question intonation and bottom figure from the speech with neutral intonation. The blue curve is the F0 contour and the yellow curve is the energy contour. As can be seen, the F0 value in the red rectangle area increases gradually.

similar to our work, but our work is special in the three following perspectives. First, unlike [7], in our experiments, the task of the network is different in the two stages of retraining. In the initial training stage, we train the network to recognize emotion in speech. While in the retraining stage, the network is trained to classify speech into different speaking styles. Second, the corpora in the two stages are different. In the first stage, the corpus used is the interactive emotional dyadic motion capture (IEMOCAP) database [10] and in the second stage, we utilize the audiobook data released by Blizzard Challenge 2012 [11]. Last, the sizes of training data in the two stages are different. The training data in the second stage is much less than the data in the first stage. Our work proves that it's helpful to consider context information in speaking style classification. This work is valuable to present the application of BLSTM in low-resource scenario. The results of our experiments also demonstrate retraining method is effective even though the task objectives and the corpora of the two training stages are quite different.

The rest of this paper is organized as follows. Section 2 introduces the architecture of BLSTM. Section 3 gives the details of the corpora used and the preparation of the data. Retraining strategy is then described in Section 4. Experiments and results are presented in Section 5 and Section 6 draws the conclusions.

II. Bidirectional Long Short Term Memroy (BLSTM)

Bidirectional long short term memory (BLSTM) is an extended architecture of recurrent neural network (RNN) [12]. With memory blocks in the hidden layer, which is also referred to as the LSTM layer, BLSTM can store information for long time duration and learn relevant context information for classification task. This characteristic is helpful in the classification of speech with different speaking styles as context information can provide clues for distinguishing different styles. Given the input sequence with T time steps $x = (x_1,\ldots,x_T)$, the corresponding output sequence $y = (y_1,\ldots,y_T)$ and the hidden layer sequence $h = (h_1,\ldots,h_T)$, the hidden layer and the output sequence are computed by the network as follows:

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \qquad (1)$$

$$y_t = W_{hy}h_t + b_y \qquad (2)$$

where t iterates from 1 to T. W are the weight matrices between layers, e.g. W_{xh} denotes the weight matrix between input $\boldsymbol{x}$ and hidden layer $\boldsymbol{h}$. b are the bias vectors of different layers, e.g. b_h is the bias vector of hidden layer. H is the hidden layer function, in this work, we use sigmoid function. The recurrent characteristic of RNN is shown in Equation (1) that the hidden layer is connected to input and the activations of hidden layer on previous step.

In the hidden layer of BLSTM, each memory block contains one or more memory cells, as well as three multiplicative gates: input, output and forget gate, which respectively simulate the write, read and reset operations of a memory cell. The activation function of the input, output and forget gates are denoted as i, o and f. The cell activation vectors are denoted as c. The hidden layer function can be implemented as (6):

$$i_t = \theta(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \qquad (3)$$

$$f_t = \theta(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \qquad (4)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \qquad (5)$$

$$o_t = \theta(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \qquad (6)$$

$$h_t = o_t \tanh(c_t) \qquad (7)$$

where θ is the logistic sigmoid function.

BLSTM contains a forward and a backward layer and thus can utilize the past and future information. $\vec{h}$ and $\overleftarrow{h}$ denote the forward and backward hidden sequences respectively.

$$\vec{h} = H(W_{x\vec{h}}x_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1} + b_{\vec{h}}) \qquad (8)$$

$$\overleftarrow{h} = H(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}) \qquad (9)$$

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

The output layer is connected to both forward and backward layers, thus the output sequence can be written as:

$$y_t = W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \tag{10}$$

III. CORPUS AND DATA PREPARATION

A. IEMOCAP Dataset

One of the training data used in our experiments is from IEMOCAP [10]. IEMOCAP database is recorded by SAIL lab at USC [13]. It contains approximately 12 hours of audio and visual data acted by 10 actors and annotated by 3 human annotators into 9 different emotion states (anger, happiness, excitement, sadness, frustration, fear, surprise, other, neutral). We use the audio data on which at least 2 annotators give the same emotion state label. Since the data of three emotion states (surprise, other and fear) are much less than that of the other states, in our experiments, we use data of the other 6 annotated states. IEMOCAP consists of 5 sessions. We use the first 4 sessions as the training set and the last session as the test set. The data was recorded with sampling frequency of 16 KHz. The total numbers of training and test sequences are 4869 and 1377 respectively. The running time of training data and test data are 362 minutes and 100 minutes.

B. Audiobook Dataset

The retraining data is from the audiobook data "A Tramp Abroad" released by Blizzard challenge 2012 [11]. The book is written by Mark Twain and the audio data is uttered by John Greenman. The book consists of 56 chapters and the running time of the speech recordings is about 16 hours. The data was sampled at 44.1 KHz and the recording conditions were acceptable. The speaker tries to express speech in different speaking styles. When the speaker utters the text in quote, he tries to speak in the role's tone and also expresses various emotions to match the intention of the text. To ensure each utterance contains only one kind of speaking style, we follow previous works [3][14] to segment the audio data into three categories, i.e. data corresponding to text in quote, text out of quote and sentence without quote. We totally obtain 5574 segmentations from all of the 56 chapters.

From the 5574 segmentations, we manually annotate those utterances with obvious speaking styles into 5 classes. The annotation method is the same as [14], which is proved effective by listening test experiment. The brief procedure is as follows. Firstly set the neutral set as null. Then listen to each segmented unit, classify the unit into neutral set if it is neutral without obvious expressivity. If the unit is perceived similar to an existing subset, add it into that subset, otherwise create a new subset containing the unit. Finally combine the similar subsets with few units. The final subsets, except the neutral set, are the annotated dataset, which contains 5 classes (471 units) as shown in Table 1. Please note that the units in Class 1 do not belong to one role. They are classified into Class 1 for they are perceived similarly in speaking styles. We randomly select 20 percent of the annotated dataset as the test set (96 units) and the rest as the training set (375 units). The running time of the training set and that of the test set is about 28 minutes and 5 minutes. All the results reported in our experiments are recorded on the test set.

C. Feature Extraction

In our experiments, for each utterance we extract acoustic features with 25 ms frame length and 10 ms window shift. For each frame, we extract 12 Mel-frequency cepstral coefficients (MFCC), F0, voice probability, logarithmic harmonics-to-noise ratio (HNR) and their delta, forming a 30 dimensional feature vector [14]. We normalize the vectors in the IEMOCAP dataset and audiobook dataset to have zero mean and unit variance respectively.

IV. RETRAINING PROCESS

Retraining strategy is simply to train a network with new training data after the network is already trained with some data. It starts from the point with lowest classification error in previous training stage. The key point of retraining is between the training stages, the weights of the network are not reset. Instead, the weights of the network are initialized with that of the previous trained network.

In [7], multiple retraining is used. More than one training set are used to train the network in sequence. The data are all from the same corpus TIDIGITS, and the tasks of different training stages are the same, i.e. digit recognition. In the output layer, the output labels of the networks in two stages are the same. In this work, we aim at exploiting whether retraining benefits when there are only limited training data for the target task (i.e. low resource retraining). What's more, the initial training data and the retraining data are from different corpora, and the tasks of the two datasets are also different. The output labels are different in two stages, as shown in Figure 2. In the

TABLE I. STATISTICS OF THE ANNOTATED AUDIOBOOK DATASET.

Class	Number of units
Class1 (Role1)	179
Class2 (Role2)	109
Class3 (Women)	78
Class5 (Sadness)	81
Class4 (Happiness)	24
Total	471

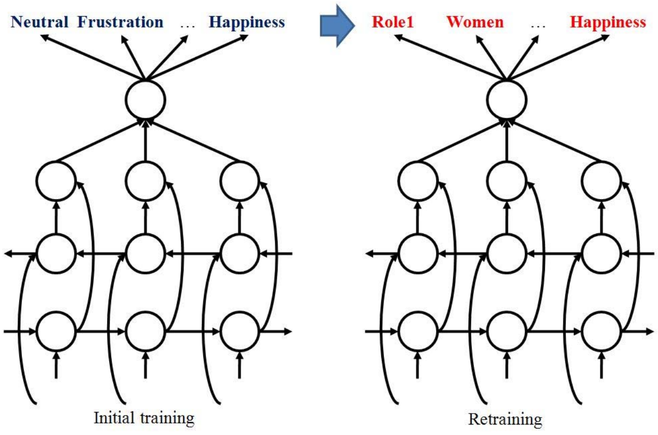

Fig. 2. Output labels in initial training and retraining stages are different.

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

first stage, the output labels are emotion categories, including neutral frustration, etc.. In the second stage, the output labels are speaking styles, like role1, role2, women, etc.

We train a BLSTM with the audiobook annotated dataset as the baseline system, which is denoted as ANNOT in the following. As both IEMOCAP and audiobook dataset have the categories of sadness and happiness, to exploit whether similar category of output in training data will affect the final performance of the network, we conduct the following 4 retraining schemes. In all the schemes, we retrain the network with audiobook dataset with all 5 classes.

- NFESH_ANNOT: The network is initially trained on the IEMOCAP training set and then retrained with audiobook dataset. In the IEMOCAP training set, the emotion classes are neutral, frustration, excitement, sadness and happiness. In the audiobook dataset, the speaking style classes are role1, role2, women, sadness and happiness. The emotion classes and the speaking style classes are in sequence in the same position of the output layer of the network. Note that in this scheme, the initial training data and the retraining data both have the labels of sadness and happiness and they are in the same output position.
- NFEAH_ANNOT: The network is initially trained on the IEMOCAP training data with neutral, frustration, excitement, angry and happiness; and then retrained with the audiobook data. In this scheme, the happiness label of the two datasets is still in the same position.
- NFESA_ANNOT: The network is firstly trained on the IEMOCAP data with neutral, frustration, excitement, sadness and angry; and then retrained with audiobook data. The sadness label is in the same position.

We also conduct an experiment with randomly selected emotion class set in random order from the IEMOCAP data as initial training data.

- HNSEF_ANNOT: From the 6 categories of IEMOCAP, we randomly select 5 labels, including happiness, neutral, sadness, excitement, frustration in order. In this scheme, each output label of the output layer are different in two stages.

V. Experiment and Results

A. BLSTM Setup

We build the BLSTM based on RNNLIB [15]. The input layer consists of 30 units. The extracted acoustic feature vector from each frame is fed into the input layer. The input layer fully connects to the hidden layer. In the hidden layer, we use two LSTM layers. Both forward and backward layer contains 50 memory blocks. Each block contains one cell with forget gate. The hidden layer is fully connected to the output layer. The output layer contains 5 units corresponding to the 5 target classes as in the previous section. There are totally 135 units and 33205 weights in the network. The weights of the network are randomized uniformly in range [-0.1, 0.1] in the initial training stage. The network is trained using stochastic gradient descent method with momentum 0.9 and learning rate 1e-4. Gradient descent updates at the end of a sequence. The final classification is based on the most active output at the end of an input sequence.

B. Experimental Results

As stated in Section 4, we conducted 4 retraining experiments to compare their performances with the baseline system. The results are presented in Table 2. The initial training epoch is the training iteration number of the previous training stage, i.e. training with IEMOCAP data, and the retraining epoch is the iteration number of the retraining stage on the audiobook data. Initial classification error is the result of testing the initially trained network on the audiobook data and the retraining classification error is the result of retrained network. As can be seen, the result of HNSEF_ANNOT is rather bad compared to the results of the other schemes. While on the other hand, the result of NFESH_ANNOT is better than others. One possible explanation is, in the scheme of NFESH_ANNOT, the initial training output and the retrain output are similar. The optimizing objectives of the two training stages are similar. The successful recognition of sadness and happiness state in IEMOCAP data probably offers correct recognition of these two classes on the audiobook test set. For the HNSEF_ANNOT scheme, the objectives of the two training stages are totally different, thus it takes some iterations to adjust the weights of the network to generate correct recognition result. As show in Table 2, the retraining classification error of HNSEF_ANNOT finally outperforms that of the baseline.

We can also find that the classification errors of the 4 retraining schemes are obviously less than that of the ANNOT scheme, as illustrated in Figure 3. The red line shows the classification error of the network initially trained with IEMOCAP data. The green line shows the classification error of the retrained network. The blue line presents the classification error of the baseline system, i.e. trained with the audiobook data only (ANNOT). This result demonstrates that the retraining strategy is effective in improving the performance of the network even though the training data of the target task is limited (i.e. low-resource). NFESH_ANNOT and NFESA_ANNOT both outperform the baseline on training speed. For the other two schemes, though they take more epochs, as shown in Figure 3, their classification error curves are steeper and their classification errors decrease faster. Therefore, the retraining method contributes to improve the training speed of the network.

TABLE II. Result of Different Training Schemes

Training Scheme	Initial Classification Error(%)	Retraining Classification Error (%)	Retraining Epochs	Initial Training Epochs
ANNOT	-	39.58	112	-
NFESH_ANNOT	67.71	**32.29**	66	52
NFESA_ANNOT	68.75	**34.38**	55	111
NFEAH_ANNOT	76.04	**31.25**	170	123
HNSEF_ANNOT	91.67	**32.29**	177	46

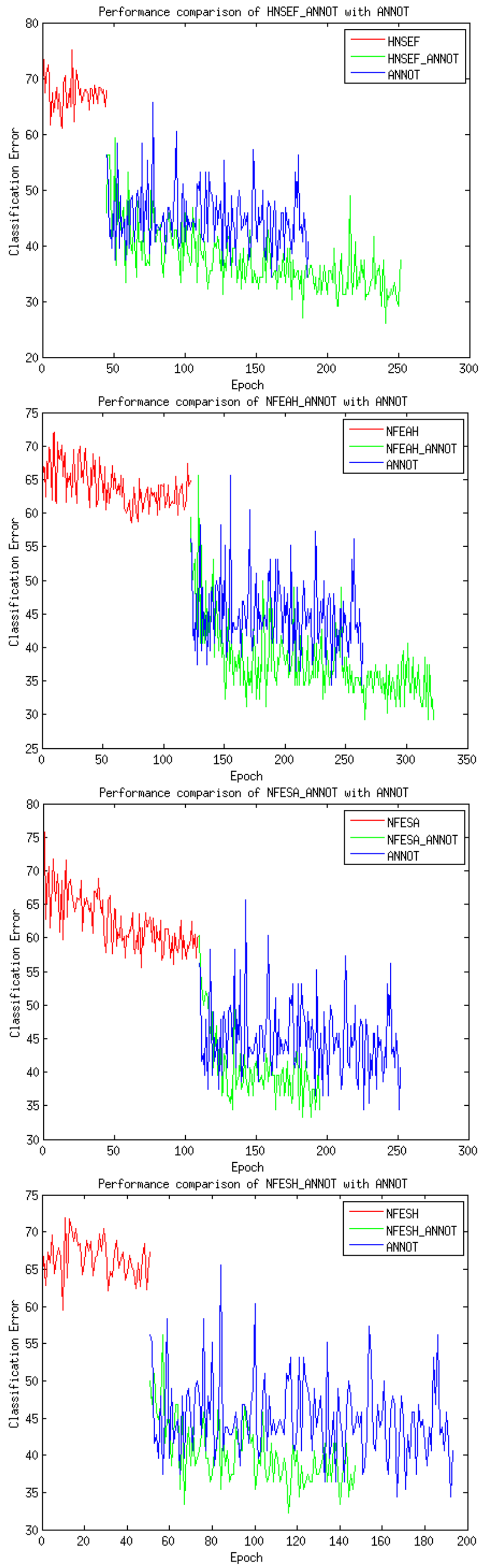

Fig. 3. Performance comparison of the retraining method and the baseline system.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we exploit the retraining strategy on BLSTM for the task of speaking style classification. Experimental results demonstrate that considering local context information is helpful for speaking style classification. The strategy is effective to improve the accuracy and the training speed of the network. What's more, it also shows that even though the task goals are different in the two stages (during network initial training and retraining) and the data are from different corpora, this method still works. In the future, we would like to train an ETTS system with the classified data.

ACKNOWLEDGEMENT

This work is supported by the National Basic Research Program of China (2012CB316401) and National High Technology Research and Development Program of China (2015AA016305). This work is also partially supported by the Hong Kong SAR Government General Research Fund (14205814), the National Natural Science Foundation of China (61375027, 61433018 and 61370023) and the Major Program for the National Social Science Fund of China (13&ZD189).

REFERENCES

[1] Z. Ren, J. Jia, Q. Guo, K. Zhang, and L. Cai, "Acoustics, content and geo-information based sentiment prediction from large-scale networked voice data," in ICME 2014, IEEE Int. Conf. on Multimedia and Expo, July 14-18, Chengdu, China, 2014, pp. 1-4.

[2] Z. Ren, J. Jia, L. Cai, K. Zhang and J. Tang, "Learning to infer public emotions from large-scale networked voice data," in MMM 2014, 20th Anniversary Int. Conf. on Multimedia Modeling, January 6-10, Dublin, Ireland, Proc., 2014, pp. 327-339.

[3] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. F. Gales, and K. Knill, "Unsupervised clustering of emotion and voice styles for expressive TTS," in ICASSP 2012, 14th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, March 25–30, Kyoto, Japan, Proc., 2012, pp. 4009–4012.

[4] M. Charfuelan and M. Schroder, "Correlation analysis of sentiment analysis scores and acoustic features in autiobook narratives," in Int. Workshop on Corpora for Research on Emotion Sentiment & Social Signals, May 26, Istanbul, Turkey, Proc., 2012, pp. 99–103.

[5] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural network," in NIPS 2009, Advances in Neural Information Process Systems, 2009, pp. 545–552.

[6] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with deep recurrent neural networks," in ICASSP 2013 – 15th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, May 26–31, Vancouver, Canada, Proc., 2013, pp. 6645–6649.

[7] A. Graves, N. Beringer, and J. Schmidhuber, "Rapid retraining on speech data with LSTM recurrent networks," Technical Report IDSIA-09-05, IDSIA, www.idsia.ch/techrep.html, 2005.

[8] R. G. Leonard, "A database for speaker-independent digit recognition," in ICASSP 1984 – IEEE Int. Conf. on Acoustics, Speech and Signal Processing, March 19–21 San Diego, USA, Proc., 1984, pp. 328–331.

[9] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in ICASSP 2013 – 15th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, May 26–31, Vancouver, Canada, Proc., 2013, pp. 6645–6649.

[10] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Language resources and evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[11] S. King, and V. Karaiskos, “The Blizzard challenge 2012,” in Blizzard Challenge Workshop, September 14, Portland, USA, 2012.

[12] 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.

[13] http://sail.usc.edu/iemocap/, [2015-06-02].

[14] X. Wu, Z. Wu, J. Jia, H. Meng, L. Cai, and W. Li, “Automatic speech data clustering with human perception based weighted distance,” in ISCSLP 2014 – 9th Int. Symp. on Chinese Spoken Language Processing, September 12–14, Singapore, Proc., 2014, pp. 216–220.

[15] A. Graves, “http://sourceforge.net/projects/rnnl/”, [2015-06-02].