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

Multi-modal learning for Speech Emotion Recognition: An Analysis and comparison of ASR outputs with ground truth transcription

Saurabh Sahu, Vikramjit Mitra, Nadee Seneviratne, Carol Espy-Wilson

Speech Communication Laboratory, University of Maryland, College Park, MD, USA

ssahu89@umd.edu, vikramjitmitra@gmail.com, nadee@terpmail.umd.edu, espy@umd.edu

Abstract

In this paper we plan to leverage multi-modal learning and automated speech recognition (ASR) systems toward building a speech-only emotion recognition model. Previous studies have shown that emotion recognition models using only acoustic features do not perform satisfactorily in detecting valence level. Text analysis has been shown to be helpful for sentiment classification. We compared classification accuracies obtained from an audio-only model, a text-only model and a multi-modal system leveraging both by performing a cross-validation analysis on IEMOCAP dataset. Confusion matrices show it's the valence level detection thats being improved by incorporating textual information. In the second stage of experiments, we used two ASR application programming interfaces (APIs) to get the transcriptions. We compare the performances of multi-modal systems using the ASR transcriptions with each other and with that of one using ground truth transcription. We analyze the confusion matrices to determine the effect of using ASR transcriptions instead of ground truth ones on class-wise accuracies. We investigate the generalisability of such a model by performing a cross-corpus study.

1. Introduction

Speech is the most common and efficient way of interaction that occurs on a daily basis and its non-invasive nature has also resulted in speech features being popular for various tasks one of them being emotion recognition. It has applications in several fields including building intelligent voice-assistants, psychiatry, analysis of human interaction and other behavioral studies [1]. Affect recognition or emotion recognition is a well-researched field and the results demonstrate that using speech features does a better job at predicting arousal levels (intensity) than valence (pleasantness) level of the utterance. In [2], the authors employed a support vector machine based regressor and found that the metric concordance correlation coefficient (CCC) is higher for predicting arousal levels than valence. They managed to improve the valence prediction task using information from other modalities such as video and physiological signals. The work in [3] shows similar results on a couple of databases after extracting features from raw waveform and spectrogram using a convolutional neural network and passing them through a neural network based regressor to get the predicted arousal and valence scores. In [4], the authors employed a fuzzy inference based system and their results show a lower mean absolute error and a higher CCC in predicting arousal than valence across three different languages. From the results shown in [5] it can be observed that the same is still true even after employing curriculum learning. The work in [6] compared different neural network based systems in classifying between angry, sad, neutral and happy and it was observed that all of them struggled in classifying the 'happy' samples correctly.

These results indicate that audio-based systems can be improved in predicting valence levels by leveraging information from other modalities. Since our aim was to build an emotion recognition model that only uses speech as input and modern state of the art ASR models can generate good transcriptions, we looked at previous works using audio and text features. In [7], the authors combined audio, video and phoneme level transcripts for multi-modal emotion classification and showed an improvement as compared to a uni-modal classifier. In [8], the authors use word level acoustic, vision and text features to implement an attention architecture that captures cross-modal dynamics. In [9], similar features were input to a deep neural framework that was implemented to capture the dynamics between speakers in a dyadic conversation. However, none of these works have provided an insight of why multi-modal learning helps for emotion classification and what is the contribution from each modality. Furthermore, all of them have used ground truth text transcription for their experiments which can be time-consuming and expensive to obtain. In [10, 11] the authors trained an ASR model on the dataset at hand and used the spoken words along with acoustic features for emotion recognition. However, due to unavailability of ground truth transcriptions they were unable to compare how much is the loss in performance when they use ASR transcriptions instead of ground truth. In this paper we analyze the performance of a multi-modal system employing audio and text features, with the hypothesis that while audio features help us with detecting arousal levels, the text features help us with valence prediction. We also developed a system that uses transcriptions obtained from different ASR models and compare its performance with that of a system that uses only audio features and a multi-modal system using ground truth transcriptions. In the next section, we provide a background of the datasets used for our cross-validation and cross-corpus study, followed by a detailed explanation of our experiment methodology in Section 2. We show our results and analyze them in Section 3. Finally we present our conclusions and future directions in Section 4.

2. Methodology

In this section we explain the databases, feature sets and classifiers used for our experiments. We then talk about the different ASR models employed to get the transcriptions to be used instead of ground truth transcriptions.

2.1. Datasets

2.1.1. IEMOCAP

We use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [12] as one of the datasets in our experiments. The dataset consists of five sessions. In each session, two actors act out scenarios which are either scripted or improvised. No two sessions have the same actor participating in

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-1149

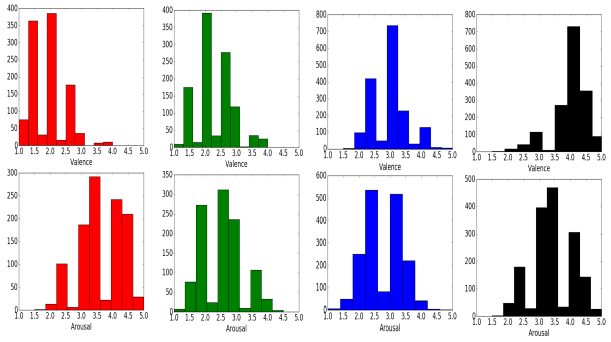

Figure 1: *Distribution of valence (top) and arousal (bottom) values for utterances in IEMOCAP belonging to classes angry (red), sad (green), neutral (blue) and happy (black) classes*

them. This enabled us to perform a five fold leave-one-session out cross-validation analysis on IEMOCAP. The conversations have been segmented into utterances which are then labeled by three annotators for emotions such as happy, sad, angry, excitement and, neutral. Manual transcriptions provided with the dataset are considered as ground truth transcriptions. For our experiments, we only use utterances for which we could obtain a majority vote and assign that as the ground truth label. We used approximately 7 hours of data from the dataset which amounts to 5530 utterances : neutral (1708), angry (1103), sad (1083), and happy (1636). Apart from annotating for categorical emotions, the utterances were also rated on a scale of 1-5 in terms of their arousal and valence; 1 being low arousal/valence and 5 being high arousal/valence. In Figure 1 we show a class-wise distribution of arousal and valence values. It can be seen that while 'anger' is low valence and high arousal, 'sad' is low both in terms of valence and arousal. 'Neutral' is more or less symmetrical along the mean for arousal and valence. The emotion 'happy' is high in valence and has more proportion of utterances with higher arousal (arousal value > 3) than 'neutral' and 'sad' but lesser proportion of high arousal utterances than that of 'angry'. Following the observations in [13] we set the length of utterances as 7.5 seconds. Shorter utterances were pre-padded with zeros while longer ones were clipped.

2.1.2. MSP-IMPROV

MSP-IMPROV [14] has actors participating in dyadic conversations across six sessions and like IEMOCAP they also have been segmented into utterances. But unlike IEMOCAP, it also includes a set of pre-defined 20 target sentences that are spoken with different emotions depending on the context of conversation. There are 7798 utterances belonging to the same four emotion classes. The class distribution is unbalanced with the number of utterances belonging to happy/neutral class more than three times that of angry/sad. We didn't have ground truth transcriptions available for this dataset. We used MSP-IMPROV to perform a cross-corpus study where we used it as a test set while IEMOCAP was used as training set.

2.2. Feature extraction

We extracted two sets of features for the speech based model and compared their performances. The first set was the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) extracted using the openSMILE toolkit [15]. It is a 23 dimensional feature set consisting of prosodic features like pitch, loudness, jitter, shimmer and spectral parameters. These features were computed for every 20 ms window with a 10 ms overlap. To reduce the computation time, we took the mean of every ten such consecutive frames so that we have a smoother feature summary vector every 100ms which was then fed to the classifier described in section 2.3. A similar approach was employed in [8] to get word level acoustic features from frame level features. The second feature set was computed using the toolkit pyAudioAnalysis [16]. This feature set was also used in [17]. The motivation behind using such a feature set is the expectation that it would be more helpful towards building a speaker agnostic emotion recognition model since they don't include prosodic features. Speaker-based normalization was applied to reduce speaker specific effects using only the neutral speech [18]. Real world emotion recognition systems usually have access to such samples, so its a fair assumption to make that they can normalize the utterances from test speakers [19].

We used 100 dimensional Glove embeddings [20] to initialize the embedding layer of the text based neural network model. The embeddings are computed by essentially factorizing the logarithm of a word-word co-occurrence count matrix obtained from a 2014 dump of Wikipedia (glove.6B). The embedding layer was then fine-tuned for the task at hand by back-propagating the error values obtained from the output layer.

2.3. Classification models

We used recurrent cells to compute a sequence of high-level representations from the time-series of feature vectors capturing their contextual information as has been done in [21,22]. For the audio modality, we had two long short-term memory (LSTM) layers with 256 and 128 hidden units, respectively, followed by a dense layer of 64 neurons with rectified linear unit (ReLU) activation which was connected to the output layer consisting of four neurons with softmax activation. Our text based model had a similar architecture except there was an embedding layer that matched the words with their corresponding Glove vector which was input to the first LSTM layer. For our multi-modal system, both outputs from the second LSTM layer of the audio modality and the text based model were concatenated to form a 256 dimensional vector. This was followed by a ReLU activated dense layer with 64 neurons and finally the output layer. A recurrent dropout probability of 0.3 was applied to all the recurrent layers in all the models. The hyper-parameters such as the number of recurrent/dense layers, number of recurrent units, batch size, dropout probability etc. were decided based on the cross-validation study done on IEMOCAP.

2.4. ASR models employed

Our next experiments involved running two free ASR applications to generate the transcriptions and using them in our experiments instead of ground truth transcriptions. This automatically generated transcription enables us to have an emotion recognition model that only requires speech as its input so that we can do away with the manual transcription of the utterances. We used the codes from [23] to get the transcriptions by implementing the models from Wit.ai (a Facebook company) and Google. We note that the ASR engines from Google and Wit.ai were not able to generate transcriptions for all the utterances mainly due to troubles with communicating with the API's server. For IEMOCAP, Google and Wit.ai could transcribe 89.9% and 78.3% of the samples, respectively. For MSP-IMPROV, the percentage of utterances for which we could obtain transcriptions using APIs from Google and Wit.ai are

90.55% and 60.24%, respectively. Below we show a few ground truth (GT) annotations and their ASR transcriptions.

1. GT: *You're going to fill out a form on your desk*
   Google: *fill out a form on your desk*
   WIT.ai: *out a form on your desk*
2. GT: *you have to tell me*
   Google: *you have to tell me*
   WIT.ai: *you have to tell me*
3. GT: *Really you don't work for anybody it's just you*
   Google: *really you don't work for anybody it's just you*
   WIT.ai: *really don't work for anybody is up*

3. Results and discussion

Here we show the results and our analysis for the experiments performed. Our metric would be un-weighted accuracy (UWA) which is the average of class-wise accuracies. Since our datasets are not perfectly balanced, we believe it would be a better metric to use than the overall accuracy or weighted accuracy. The results shown have been averaged across four runs with different random seeds.

3.1. Comparing audio and text modalities

Our initial set of experiments were carried out to show the worth of multi-modal systems. We compared the two different audio feature sets eGeMAPS and the ones obtained using pyAudioAnalysis but didn't notice a big difference in the accuracies. We believe that since the feature sets have undergone speaker based normalization prior to being fed to the neural network model, we are getting rid of speaker specific characteristics and hence the speaker-specific prosodic features used in eGeMAPS don't deteriorate the performance of the audio-only model.We chose to use the pyAudio feature set for further experiments. Next we investigated the performance of a text-based system and a multi-modal system. It can be seen from Table 1 that both of those models perform better than an audio-only model. To verify our assumption that the audio modality is better for detecting arousal while the text modality is better at detecting valence, we provide the confusion matrices in Figure 2. It can be seen that the audio-based model (left) performs better than the text-based model (center) in detecting 'anger' which is a high arousal emotion. From the first row of the matrices, we can also see that the text based model is more likely to confuse the 'angry' and 'sad' classes than is the audio based model. This is because both anger and sadness are low valence emotions but they differ in their arousal level, thereby making it easier for the audio modality to distinguish between the two. Both the models perform similarly when it comes to identifying the 'neutral' speech samples. This is probably because the 'neutral' class lies somewhere in the middle of the arousal and valence axes and not at one of the extremes. Hence neither of the modalities end up having any advantage over the other. However, text based models do a much better job in identifying the 'happy' samples than the audio based model. While the audio based model classifies 26% of 'happy' samples as angry, our text based model does a better job at distinguishing between the two classes. It further strengthens our hypotheses that text based models are better than audio based models in distinguishing between high and low valence utterances. While anger is a low valence emotion, happiness is high valence. Combining the two modalities we see that class-wise accuracies either improve or remain almost the same for all of the classes. Accuracies for 'sad' and 'neutral' obtained using multi-modal system are better than that of uni-modal systems indicating that speech and text features supplement each other while identifying samples from these two classes.

Table 1: *UWA obtained from 5-fold cross-validation on IEMOCAP. Ground truth text transcriptions are used here.*

Model	pyAudio	eGeMAPS	Glove	pyAudio + Glove
UWA	56.94	56.85	61.89	**68.18**

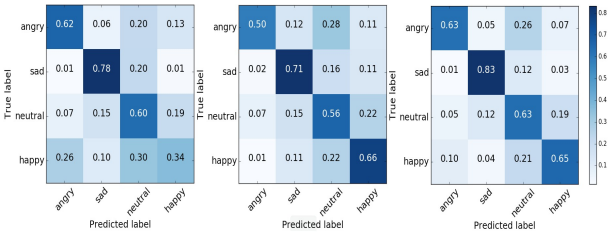

Figure 2: *Confusion matrices for one of the validation splits showing the class-wise accuracies of audio-only (left), text-only (center) and multi-modal (right) systems. Numbers shown are in percentages*

3.2. ASR model output vs ground truth transcriptions used for multi-modal classification

Having performed the multi-modal experiments on ground truth transcriptions, we now created a pipeline where we only used audio data as input. We used the ASR transcriptions generated from audio in the multi-modal system instead of ground truth transcriptions. Since, the different APIs were able to transcribe different numbers of utterances, we ran the experiments comparing the models with a different train/test file-list for each API. This resulted in different accuracies even when only audio features were used or when they were used along with text features obtained from ground-truth transcriptions. Figure 3(a) shows the comparison between the cross-validation UWAs obtained from the audio only model, the multi-modal system using ground truth text and the multi-modal system using the API's transcription. It indicates that the model trained on ground truth transcriptions perform better than the ASR transcriptions as expected. We get a relative loss of 4% and 5.3% in accuracy compared to ground truth transcriptions when using Google's and Wit.ai's ASR engine, respectively. To compare the quality of the transcriptions generated, we computed the word error rate (WER) by measuring the Levensthein distance [24] (LD) between the generated transcriptions and the ground truth ones for each IEMOCAP utterance and then averaging it over the entire dataset. Levensthein distance between two sentences measures the minimum number of insertions/deletions/substitutions of words required to convert one sentence to another. In general, longer utterances are more likely to have a higher LD when compared to shorter utterances because there are more words where the ASR model can make an error in transcribing. Since the different API's transcribed different numbers of utterances, this measure could provide us with a skewed idea about the performance of APIs. Hence, we also computed a normalized Levensthein distance (NLD) where we divide LD by the number of words in the ground truth transcription. Figure 3 compares the performance of the two ASR APIs in terms of those two metrics. We see that the difference is less stark in case of NLD, however both the metrics show similar trends. The lower drop in UWA compared to ground truth transcriptions was obtained

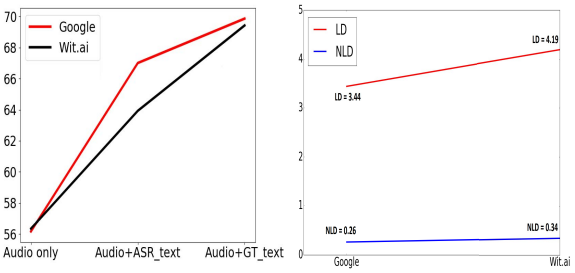

Figure 3: *(a) Left figure compares the performance of an audio-only model with multi-modal systems using ground truth text or the ASR transcriptions for the different ASR systems (b) In the right we show the performance of the different ASR modules used in our experiment on IEMOCAP. Lower is better*

using Google's system. This can be explained by its lower word error rate as obtained for the IEMOCAP dataset. Google's and Wit.ai's APIs have probably been trained on a large amount of data so that the deep learning models used for ASR in both the APIs were more generalisable giving us satisfactory performance on an unseen dataset. Wit.ai's API seems to perform worse than Google's API in terms of UWA, but we should also keep in mind that we are using different subsets of the dataset to evaluate the models. Also we are using less data to train the pipeline using Wit.ai's transcriptions (as explained in section 2.4) which could also be one of the reasons for its worse performance. Having looked at the WER of the two APIs, we now compare the average confusion matrix obtained over five cross-validation sets for multi-modal systems using ground truth transcription vs ASR outputs in Figure 4. It can be observed that class-wise accuracies are higher when ground truth transcriptions are used as expected. Comparing models using Google API's output with that of using ground truth transcriptions, the absolute increase in percentage of 'happy' samples being classified as 'angry' and 'neutral' is more than that of the 'sad' class. This could be because the arousal value distribution of 'happy' utterances is more similar to 'angry' and 'neutral' than that of 'sad' utterances (from Figure 1). When using Wit.ai API's output instead of the ground truth transcriptions we see more 'angry' samples being miss-classified as 'happy' probably for a similar reason. The same is true when there are more 'sad' samples being miss-classified as 'neutral' and 'happy' when Wit.ai is used to transcribe. These observations show that using worse quality transcriptions leads to more confusion between classes with similar arousal values which points to the fact that audio features contribute to the classification to a greater extent in such cases.

3.3. Cross-corpus analysis

To verify the generisability of our model, we did a cross-corpus analysis where we trained our model using IEMOCAP and tested it on MSP-IMPROV. We preferred IEMOCAP for training because it is more balanced. We have compared the performance between an audio-only system and a multi-modal system using the generated transcriptions. The tokenizer used in these experiments were generated from the IEMOCAP dataset. Doing so would allow us to capture the cross-domain difference in their vocabulary. Utterances in MSP-IMPROV for which we could not find any of the words in the tokenizer were not used in the experiment. We see a similar trend where using ASR transcriptions along with audio results in a better emotion recog-

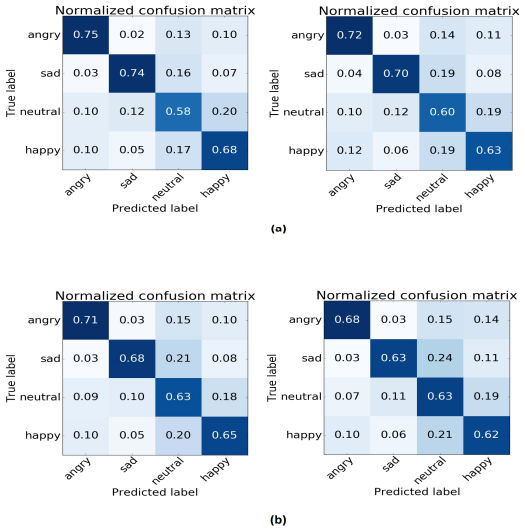

Figure 4: *Confusion matrix obtained from multi-modal systems using ground truth transcriptions (left) and ASR transcriptions (right) using (a) Google's and (b) Wit.ai's API*

Table 2: *Cross corpus results with IEMOCAP as training set and MSP-IMPROV as test set.*

Model	**Google**	**Wit.ai**
pyAudio	35.93	38.06
pyAudio + Glove from ASR output	**39.45**	**40.08**

nition model. The Google based system gives a relative improvement of 9.8% and using Wit.ai's ASR API results in a relative improvement of 5.2% compared to an audio-only model. However, the improvements weren't as much as observed in cross-validation experiments, possibly due to cross-domain differences in the vocabulary of IEMOCAP and MSP-IMPROV.

4. Conclusion

Our experiments demonstrate that the acoustic features help in detecting the level of arousal whereas the text based model helps in detecting the valence level. Combining information from both to build a multi-modal system seems to increase the class-wise accuracies. When using ASR transcriptions instead of ground truth ones, audio features seem to contribute more towards deciding which class an utterance should belong to. Deep learning based ASR models trained on thousands of hours of data [25] improves their generalisability thereby giving us meaningful transcriptions for unseen datasets which we can leverage to get higher cross-corpus accuracies. Hence, we can take advantage of the genralisability of ASR models to improve the generalisability of emotion classification models. In the future we plan to investigate the utility of articulatory features by incorporating them in our multi-modal system. We also aim to explore various word embeddings other than Glove or sub-word embeddings which are better at handling out of domain vocabulary words. We also plan to look at ways we can get word embeddings specific for an emotion recognition/sentiment classification task [26]. It would also be interesting to explore text features obtained using dictionaries used specifically for an emotion recognition/sentiment classification tasks. Additionally, we plan to explore novel ways to combine the information from the audio and text modes in the multi-modal learning framework.

5. References

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

[2] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[3] Z. Yang and J. Hirschberg, "Predicting arousal and valence from waveforms and spectrograms using deep neural networks," *Proc. Interspeech 2018*, pp. 3092–3096, 2018.

[4] X. Li and M. Akagi, "A three-layer emotion perception model for valence and arousal-based detection from multilingual speech," *Proc. Interspeech 2018*, pp. 3643–3647, 2018.

[5] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *arXiv preprint arXiv:1805.10339*, 2018.

[6] J. Kim, K. P. Truong, G. Englebienne, and V. Evers, "Learning spectro-temporal features with 3d cnns for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 383–388.

[7] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2462–2465.

[8] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," *arXiv preprint arXiv:1802.00923*, 2018.

[9] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, vol. 1, pp. 2122–2132.

[10] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[11] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion recognition from speech: putting asr in the loop," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4585–4588.

[12] C. Busso, M. Bulut, C.-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, 2008.

[13] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[14] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, , no. 1, pp. 67–80, 2017.

[15] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[16] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, pp. e0144610, 2015.

[17] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," *arXiv preprint arXiv:1701.08071*, 2017.

[18] C. Busso, A. Metallinou, and S. S. Narayanan, "Iterative feature normalization for emotional speech detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5692–5695.

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

[20] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[21] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," 2015.

[22] C.-W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition.," in *INTERSPEECH*, 2016, pp. 1387–1391.

[23] A. Zhang, "Speech recognition (version 3.8)," https://github.com/Uberi/speech_recognition#readme, 2017.

[24] W. J. Heeringa, *Measuring dialect pronunciation differences using Levenshtein distance*, Ph.D. thesis, Citeseer, 2004.

[25] R. Prabhavalkar, T. N. Sainath, B. Li, K. Rao, and N. Jaitly, "An analysis of attention in sequence-to-sequence models,," in *Proc. of Interspeech*, 2017.

[26] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, vol. 1, pp. 1555–1565.