INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Exploration of Acoustic and Lexical Cues for the INTERSPEECH 2020 Computational Paralinguistic Challenge

Ziqing Yang, Zifan An, Zehao Fan, Chengye Jing, and Houwei Cao

Department of Computer Science, New York Institute of Technology
(zyang23, zan01, zfan06, cjing, and hcao02)@nyit.edu

Abstract

In this paper, we investigate various acoustic features and lexical features for the INTERSPEECH 2020 Computational Paralinguistic Challenge. For the acoustic analysis, we show that the proposed FV-MFCC feature is very promising, which has very strong prediction power on its own, and can also provide complementary information when fused with other acoustic features. For the lexical representation, we find that the corpus-dependent TF.IDF feature is by far the best representation. We also explore several model fusion techniques to combine different modalities together, and propose novel SVM models to aggregate the chunk-level predictions to the narrative-level predictions based on the chunk-level decision functionals. Finally we discuss the potential for improving prediction by combining the lexical and acoustic modalities together, and we find that fusion of lexical and acoustic modalities do not lead to consistent improvements over elderly Arousal, but substantially improve over the Valence. Our methods significantly outperform the official baselines on the test set in the participated Mask and Elderly Sub-challenges. We obtain an UAR of 75.1%, 54.3%, and 59.0% on the Mask, Elderly Arousal and Valence prediction tasks respectively.

Index Terms: lexical features, fisher vector, model fusion, Computational Paralinguistic Challenge

1. Introduction

In this study, we set out to compare a variety of representations of acoustics and lexical usage for the ComParE Challenge 2020 [1]. This year's challenge addresses three new problems within the field of Computational Paralinguistics in a challenge setting: the Elderly Emotion, the Mask, and the Breathing Sub-Challenges. Particularly, we participate in the Elderly Emotion and Mask Sub-challenges. The goal of our work is to establish which of acoustic and/or linguistic representations are the most suitable for predicting mask-wearing conditions, and Arousal (A) and Valence (V) in the speech of elderly individuals, and to what extent different features can be combined to further improve the prediction accuracy.

In this paper we describe in detail on our official submission system[1] to the ComParE Challenge 2020. We use the same training, development, and test corpus definitions as the challenge. The results are therefore comparable with the challenge benchmarks and with results published by other participants. In the Mask Sub-Challenge, the Mask Augsburg Speech Corpus (MASC) [1] is used. We explore novel Fisher Vector (FV) encoding over the traditional 39 MFCC features, and the 130 LLDs from the standard ComParE feature set [2], to predict whether a speaker wears a surgical mask or not. The proposed FV-MFCC feature outperforms all of the four baseline acoustic features. We also develop several fusion techniques to combine various acoustic indicators together, and demonstrate the unique contribution of the FV-MFCC feature in model fusion. Our best SVM fusion models are trained with the decision scores from six individual acoustic features, which achieve 75.1% UAR on the testing set, 3.3% higher than the official baseline.

In the Elderly Emotion Sub-Challenge, we use the Ulm State of Mind in Speech-elderly (USOMS-e) corpus [1], which contains both the elderly emotional speech and the manual transcription of the speech. The task is to predict the elderly emotional speech in both Arousal and Valence dimension in the three levels of (L)ow, (M)edian, and (H)igh. We explore various acoustic and lexical representations in this sub-challenge. In general, we found that the acoustic features work better on A and the linguistic features work better on V. As the original audio narrative files were segmented into 5 sec chunks at first for acoustic analysis and prediction, we need to aggregate the acoustic predictions from chunk-level to generate the final prediction on narratives. We further explore several model fusion techniques to combine different modalities together, and propose novel SVM models to aggregate the chunk-level predictions to the narrative-level predictions based on the chunk-level decision functionals. We demonstrate on the development set that the chunk to narrative aggregation seems playing a very important role on final system performance, and significant improvement can be achieved by using our proposed aggregation model. Finally we discuss the potential for improving prediction by combining the lexical and acoustic modalities together, and we find that fusion of lexical and acoustic modalities do not lead to consistent improvements over Elderly Arousal, but substantially improve over the Valence. Our proposed acoustic and linguistics models significantly outperform the baselines on predicting the Arousal and Valence in elderly speech, which achieve mean UAR of 56.6%, 6.9% higher than the baseline.

2. Audio Representation

Four different types of acoustic features, including the ComParE acoustic feature set [2], Bag-of-Audio-Words (BoAWs) [3], DEEP SPECTRUM [4][5] and AUDEEP [6][7], are given as the baseline features for the challenges. Besides those baseline features, we further investigate the following three types of feature sets: Fisher vector representations over 39 MFCC, over 130 LLDs from the standard ComParE feature set, and MFCC functional features.

The Fisher Vector (FV) [8], combining the advantages of generative statistical models (e.g., Gaussian Mixture Model) and those of discriminative methods (e.g., support vector machines), has been widely used in computer vision [9] and information retrieval tasks [10]. FV encoder encapsulates the first and second order differences between the pooled local features

[1] https://github.com/MannyCooper/INTERSPEECH-2020-ComParE

http://dx.doi.org/10.21437/Interspeech.2020-2999

and the dictionary which is built using Gaussian Mixture Models (GMM). Recently, the FV has been successfully applied in affective computing applications [11].

To apply FV to extract acoustic representations, we first need to generate low-level acoustic descriptors, and then use the Gaussian Mixture Models (GMM) to build the dictionary to model the distribution of the low-level descriptors. This GMM-based dictionary can be considered as probabilistic acoustic dictionary. Using this dictionary, weighted measures of the descriptors are assigned to multiple clusters in the GMM to generate the final FV representations.

In this paper, we first use the frame-level MFCC features as the low-level features. MFCC features are the state-of-the-art acoustic features for various speech applications, including speech recognition, speaker verification, language identification, etc. The frame-level feature vector has 39 components: 13 MFCC and their first and second-order time derivatives. Then, we use all the MFCCs extracted from the training set to fit GMMs and to extract the **FV-MFCC** representation. We fit GMMs with N = 2, 4, 8, 16, 32, 64 and 128 components. In addition, we consider the 130 LLDs from the baseline ComParE feature set, and use them to fit GMMs to extract the **FV-ComParE** representation.

We also extract the **MFCC functional** features by using the openSMILE toolkit [12]. It is a comprehensive utterance-level feature set containing 1,400 static features resulting from the computation of functionals (e.g., mean, standard deviation, percentiles and peak statistics) over frame-level MFCC attributes. It is similar to the baseline ComParE features, while we only use 28 LLDs (13 MFCCs along with energy, and their first time derivatives), instead of 130 LLDs used in the ComParE features.

3. Lexical Representation

Many words have strong positive or negative connotations. And often what people say (the words they use) carry rich information about their affective states. Much work in text processing has shown that subjectivity, opinion and emotion can be successfully estimated simply on the basis of lexical features [13][14][15]. Inspired by the conventional bag of words representations in which texts are represented as sparse vectors of occurrence counts of words from a predefined vocabulary, we investigate several lexical representations with various sparse feature spaces with different vocabulary as follows.

- **BoW:** Bag-of-word features. The feature space consists of all the words in the training data. The representation of utterance has value zero for words not in the utterance, and value one for words that do appear in it.
- **TF.IDF**: Term Frequency–Inverse Document Frequency features. The feature space consists of all the words in the training data. The values of components are determined by the term frequency–normalized word counts. The inverse document frequency for a word is determined by the number of all conversations and the number of conversations that contain the word.
- **Sparse PMI**: Sparse lexical representation with point-wise mutual information (PMI) selected words. The value of the component corresponding to a particular word is equal to the PMI between the word and the affect dimension that is being predicted.
- **PMI–TF.IDF**: Sparse lexical representation with point-wise mutual information (PMI) selected words. The value of the component corresponding to a particular word is equal to the PMI between the word and the affect dimension that is being predicted, weighted by the TF.IDF scores.
- **Sparse NGD**: Sparse lexical representation with normalized Google distance (NGD) selected words. This representation is similar to the BoW one, however the value of the component corresponding to a particular word is equal to the NGD between the word and the affect dimension that is being predicted.
- **NGD–TF.IDF**: Sparse lexical representation with normalized Google distance (NGD) selected words. This representation is similar to the BoW one, however the value of the component corresponding to a particular word is equal to the NGD between the word and the affect dimension that is being predicted, multiplied by the TF.IDF scores.

BoW and **TF.IDF** are two of the commonly used feature sets in information retrieval and text classification tasks. In this study, we build the feature space by considering all the $2,836$ words that appeared in the training dataset. Each story is represented by a vector of length $2,836$.

To construct **Sparse PMI**, we first calculate the point-wise mutual information (PMI) between a word and a given affect dimension. PMI has been successfully applied for categorical emotion estimation and is widely used as a measure of association in a range of semantic processing applications [16, 17, 18]. The PMI between a word w and a dimension of affect ε can be computed as:

$$PMI(\varepsilon, w) = \log \frac{P(\varepsilon, w)}{P(\varepsilon)P(w)} = \log \frac{P(\varepsilon|w)}{P(\varepsilon)}, \qquad (1)$$

where $P(\varepsilon)$ is the prior probability of an affect dimension and $P(\varepsilon|w)$ is the conditional probability of the affect dimension given the word w. Both probabilities are computed directly from the data. For each word w in the training set and for each affect dimension A and V, we compute three PMI values, associated between the word and class `Low`, `Median`, and `High`.

Afterwards, the feature space of **Sparse PMI** is defined by the selected words with high PMI values. For each affect dimension, we build a set of $1,500$ words, consisting of 500 words with the highest PMI for class `Low`, `Median`, and `High` respectively.

The **PMI–TF.IDF** feature is similar to the **Sparse PMI** feature, and the TF.IDF weights of those words are also considered here to better assess the importance of a word to a story in the corpus. As a result, each story is represented by $1,500$ features with the corresponding PMI values multiplied by their TF.IDF weights.

The normalized Google distance (NGD) is originally used to measure the semantic similarity of keywords based on the number of hits returned by the Google search engine for a given set of keywords [19]. Here, we use NGD to measure the semantic similarity between a word w and a given affect dimension ε:

$$NGD(\varepsilon, w) = \frac{\max(\log f(\varepsilon), \log f(w)) - \log f(\varepsilon, w)}{\log N - \min(\log f(\varepsilon), \log f(w))}, \qquad (2)$$

where N is the total number of stories in the training set multiplied by the average number of words in each story; $f(\varepsilon)$ and $f(w)$ are the numbers of occurrences for an affect dimension ε

and a given word w, respectively; and $f(\varepsilon, w)$ is the number of stories which both ε and w occur.

Similar to **Sparse PMI**, the **Sparse NGD** feature select 500 words with the highest NGD for class `Low`, `Median`, and `High` respectively, and each story is represented by $1,500$ features with the corresponding NGD values. The **NGD–TF.IDF** feature is the **Sparse NGD** feature multiplied by their TF.IDF weights.

4. Fusion Techniques

4.1. Combination of Different Modalities

In this paper we investigate several fusion techniques, including early-stage feature fusion and late-stage decision fusion, to combine different modalities together. In the early fusion, we directly combine various types of features together to train the prediction models. We examine four approaches for late-stage fusion based on the decision scores from different modalities: 1) Majority Vote (MV), 2) Harmonic Mean, 3) SVM fusion, and 4) DNN fusion.

In MV-based fusion, the final predictions will be given based on the majority prediction from different feature sets. In fusion based on harmonic mean, we combine predictions from different modalities by taking the harmonic mean of their output decision scores, and the utterance is classified as the class for which it achieved the highest average scores. In SVM fusion, we use the decision scores obtained from different modalities as features to train a second layer SVM to generate the final prediction. In DNN fusion, we use the decision scores obtained from different modalities as input to train a deep neural network to predict the final results.

4.2. Chunks to Narrative Prediction

For the elderly sub challenge (ESC), the original audio narrative files were segmented into 5 sec chunks at first for acoustic analysis and prediction. In the baseline system, the final prediction on narrative is based on the majority voting of chunk-level predictions on all chunks in the narrative. In this paper, we propose two approaches to better aggregate the predictions from chunk-level to narrative-level, in addition to MV.

The first approach is to train an SVM to predict the narrative-level prediction with the chunk-level decision scores. For each narrative, we first compute the 12 statistics and functionals (e.g., max, min, mean, std, range, argmax, argmin, and linear regression coefficients like intercept, kurtosis, meanstd_err, skew and slope.) over the chunk-level prediction scores on `Low`, `Median`, and `High` classes. The proposed functional analysis can help us extract a fixed-length feature vector from a varying-length narratives, while still capturing the dynamic information within narratives. As a result, each narrative will be represented by 36 static features, and those features will be used to train an SVM classifier to predict the affective states on narratives.

The second approach is to use RNN (Recurrent Neural Network) [20] and LSTM (Long short-term memory) [21] to explore the potential sequential patterns between audio chunks in the same narrative. The chunk-level prediction scores on `Low`, `Median`, and `High` classes are used as input features to train RNN and LSTM models. Our RNN network consists of one single layer with 16 hidden units, and our LSTM network consists of two stacked hidden layers with 16 and 8 units respectively. The output layer is one fully connected layer with *softmax* activation function.

5. Experiments and Results

5.1. The Mask Sub-challenges

Following the standard setting, for each individual class of features, we first use Support Vector Machine (SVM) with linear kernel for classification. For early-stage fusion, we combine all the individual feature sets together and train SVM with linear kernel for prediction. The SVM parameters are optimized on the development set and the SVM classification models are trained on the training set. After that, we employed late-stage fusion to combine the different feature sets together based on their decision scores in terms of majority vote, harmonic mean, or train a second-level SVM for model combination.The SVM fusion models are trained with RBF kernels with the decision scores from individual features on the development set. The results on all the individual features, as well as the fusion results, are shown in Table 1.

First we discuss the prediction power of the individual features. We can see that the proposed FV-MFCC features achieve 69.4% UAR on the development set, significantly outperform all the other individual acoustic features provided by the challenge baseline, including ComParE functionals, BoAWs, Deep Spectrum and AUDEEP features. On the other hand, both the FV-MFCC and FV-ComParE features outperform the MFCC and ComParE functional features by a large margin. This demonstrates the benefit of the fisher vector encoding paradigm.

Next we turn to discuss the fusion results. First we notice that the early-fusion results (69.0%) on Dev. set could not beat the best single feature (69.4% achieved by FV-MFCC). Late-stage fusion seems more promising except for the simple MV techniques. Fusion based on harmonic mean is slightly better than the best single feature, and the SVM fusion achieves 71.3% UAR, which is the highest among all the fusion results. The SVM fusion achieves consistently high performance on the testing set. The UAR on the testing set reaches **75.1%**, which is 3.3% higher than the official baseline. Note that the baseline performance was also achieved via a combination of the four baseline acoustic predictions based on majority vote.

Table 1: *Results obtained for the Mask Sub-Challenge; the performance is measured in terms of the UAR (%) on Dev. and/or Test set.*

	Dev	Test
Individual Features		
ComParE functionals	62.6	66.9
Bag-of-Audio-Words (BoAWs 2000)	64.2	67.7
DEEP SPECTRUM (ResNet50)	63.4	70.8
AUDEEP (Fused)	64.4	66.6
MFCC functionals	60.4	–
FV-ComParE	67.0	–
FV-MFCC	**69.4**	–
Fusion		
Early Fusion	69.0	–
Majority Vote (MV)	69.4	–
Harmonic Mean	70.3	–
SVM Fusion	71.3	**75.1**
Official Baseline	–	71.8

5.2. The Elderly Sub-challenges

In the Elderly sub-challenges, for each individual class of acoustic features, we first use SVM with linear kernel to generate the chunk-level predictions. The SVM parameters are optimized on the development set and the SVM classification models are trained on the training set. After that, we train SVM, RNN, and LSTM models based on chunk-level scores to aggregate the chunk-level predictions to narrative-level predictions on the development set. Finally, we employ various late-stage fusion techniques (MV, DNN, SVM) to combine the best narrative-level decision scores from different acoustic feature sets together. The DNN and SVM fusion models are trained and optimized on the development set.

The linguistic features are extracted directly on the narrative-level. For each type of linguistic features, we train SVM with linear kernel. The SVM parameters are optimized on the development set and the SVM classification models are trained on the training set. The results on the Elderly sub-challenge are shown in Table 2.

Table 2: *Results obtained for the Elderly Sub-Challenge; the performance is measured in terms of the UAR (%) on Dev. and/or Test set.*

Chunk to Narrative Aggregation with Acoustic Only: UAR (%) on Dev.				
	MV	SVM	RNN	LSTM
	A/V	A/V	A/V	A/V
ComParE functionals	39.1/40.1	48.1/42.4	34.5/36.6	36.8/39.6
Bag-of-Audio-Words	42.0/40.6	**48.2**/43.9	45.5/42.2	40.4/45.9
DEEP SPECTRUM	34.9/31.6	41.3/39.1	39.3/41.5	38.6/41.5
AUDEEP	40.4/32.7	45.2/45.9	45.1/41.1	43.8/43.7
FV-MFCC	42.0/45.5	41.9/**47.3**	44.7/41.6	33.8/45.8
Fusion of the Narrative Predictions with Acoustic Only: UAR (%)				
	Dev		Test	
	A	V	A	V
Early Fusion	44.9	50.3	–	–
Majority Vote (MV)	51.2	53.7	**54.3**	31.4
DNN Fusion	37.4	35.6	–	–
SVM Fusion	58.5	45.2	38.7	40.6
Narrative Predictions with Linguistic Features: UAR (%)				
	Dev		Test	
	A	V	A	V
BoW	36.5	53.3	–	–
TF.IDF	39.2	**64.9**	38.8	54.6
Sparse PMI	38.2	55.0	–	–
Sparse NGD	34.1	58.7	–	–
PMI–BoW	35.0	55.7	–	–
NGD–BoW	33.5	58.7	–	–
PMI–TF.IDF	37.5	55.7	–	–
NGD–TF.IDF	37.8	56.6	–	–
Fusion of the Narrative Predictions with Linguistic Features: UAR (%)				
	Dev		Test	
	A	V	A	V
Majority Vote (MV)	41.2	65.9	–	–
Narrative Predictions with Acoustic + Linguistic Features: UAR (%)				
	Dev		Test	
	A	V	A	V
Acoustic+Linguistic (MV)	51.9	70.8	38.9	**59.0**
Official Baseline	–	–	50.4	49.0

We first discuss the performance of acoustic prediction. We notice that the prediction power of various acoustic features are very different, on both A and V affective states. Moreover, the chunk to narrative aggregation seems playing a very important role on system performance. The final prediction accuracy (UAR) with different aggregation approaches vary a lot, even with the same acoustic features. The proposed SVM-based aggregation models trained with chunk-level decision functionals consistently outperform the other aggregation techniques, on most of the acoustic features. Our best prediction with single acoustic feature set are 48.2% (with BoAW) on Arousal (A) and 47.3% (with FV-MFCC) on Valence (V), which significantly outperform the best single acoustic baseline (42.0% on A with BoAW and 45.7% on V with ComParE functionals). Note that the baseline uses majority vote to aggregate the chunk-level prediction to narrative. The RNN and LSTM aggregation models are not very robust. This maybe because we have very few training data – only 87 narratives in the training set.

Next we turn to discuss the fusion with various acoustic features. Generally speaking, fusion of five acoustic modalities with majority vote achieves very promising performance. The UAR on A and V on the Dev. set are 51.2% and 53.7% respectively, which is 3% and 6.4% higher than the best single acoustic features. We evaluate this fusion model on the test set as well, which achieves very encouraging results on A. The UAR is **54.3%**, which outperforms the official baseline of 50.4% by a large margin. Not surprisingly, our acoustic fusion model doesn't work well on V, since V is usually modeled better with linguistics acoustics.

Then we turn to discuss the prediction power of different linguistic features. The prediction results of all the proposed linguistic features on the development set can be found in Table 2. As expected, the linguistic features perform much better on V than A, and vise versa. The state-of-the-art TF.IDF feature achieves the best performance on V. The UAR on the Dev set is 64.9%. Note that the best baseline linguistic features has a 56.1% UAR on the development set. We evaluate the TF.IDF features on test set as well, the UAR reaches 54.6%, significantly outperforming all the baseline linguistic features extracted with frozen BERT models. We further combine various linguistic features together by using majority vote. Marginal improvements have been obtained after fusion on both A and V on the development set.

Finally we discuss the potential for improving prediction by combining the lexical and acoustic modalities together. We find that fusion of lexical and acoustic modalities do not lead to consistent improvements over Arousal but substantially improve over the Valence. Combining various acoustic and linguistic modalities together consistently achieves the best performance on Valence. The UAR is 70.8% on the development set and **59.0%** on the test set, which significantly outperforms the official baseline of 49% achieved by the BLAtt features.

6. Conclusions

In this paper, we investigated various acoustic and linguistic representations on the Mask and Elderly Sub-challenges of INTERSPEECH 2020 ComParE challenge. We have shown that the proposed FV-MFCC feature is very promising. It has very strong prediction power on its own, and can also provide complementary information when fused with other acoustic features. We further explored several model fusion techniques to combine different modalities together, and proposed novel SVM models to aggregate the chunk-level predictions to the narrative-level predictions based on the chunk-level decision functionals. Finally we discussed the potential for improving prediction by combining the lexical and acoustic modalities together. Our methods significantly outperform the official baselines on the test set in the Mask Sub-challenges, and on both Arousal and Valence dimension in the Elderly Sub-challenges. We obtain an UAR of 75.1%, 54.3%, and 59.0% on the Mask and Elderly Arousal & Valence prediction tasks respectively, which are 3.3%, 3.9% and 10.0% higher than the official baseline UARs of the challenge.

7. Acknowledgements

This work is partially supported by the NSF of USA under EAGER Grant IIS-2034791.

8. References

[1] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," *Proceedings INTERSPEECH. Shanghai, China: ISCA*, 2020.

[2] 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.

[3] M. Schmitt and B. Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3370–3374, 2017.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. W. Schuller, "Snore sound classification using image-based deep spectrum features." in *INTERSPEECH*, vol. 434, 2017, pp. 3512–3516.

[6] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. of the DCASE 2017 Workshop*, 2017.

[7] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340–6344, 2017.

[8] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Advances in neural information processing systems*, 1999, pp. 487–493.

[9] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[10] S. Fiel and R. Sablatnig, "Writer identification and writer retrieval using the fisher vector on visual vocabularies," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 545–549.

[11] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 255–259.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[14] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, A. N. Vembu, and R. Prasad, "Emotion recognition using acoustic and lexical features," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[15] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4749–4753.

[16] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, A. N. Vembu, and R. Prasad, "Emotion recognition using acoustic and lexical features," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[17] H. Cao, A. Savran, R. Verma, and A. Nenkova, "Acoustic and lexical representations for affect prediction in spontaneous conversations," *Computer speech & language*, vol. 29, no. 1, pp. 203–217, 2015.

[18] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012, pp. 485–492.

[19] R. L. Cilibrasi and P. M. Vitanyi, "The google similarity distance," *IEEE Transactions on knowledge and data engineering*, vol. 19, no. 3, pp. 370–383, 2007.

[20] A. Cleeremans, D. Servan-Schreiber, and J. L. McClelland, "Finite state automata and simple recurrent networks," *Neural Computation*, vol. 1, no. 3, pp. 372–381, 1989.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735