Improving Valence Prediction in Dimensional Speech Emotion Recognition Using Linguistic Information

Bagus Tris Atmaja
Department of Engineering Physics
Sepuluh Nopember Institute of Technology
Surabaya, Indonesia
bagus@ep.its.ac.id

Masato Akagi
School of Information Science
Japan Adv. Inst. of Sci. & Tech.
Nomi, Japan
akagi@jaist.ac.jp

***Abstract*—In dimensional emotion recognition, a model called valence, arousal, and dominance is widely used. The current research in dimensional speech emotion recognition has shown a problem that the performance of valence prediction is lower than arousal and dominance. This paper presents an approach to tackle this problem: improving the low score of valence prediction by utilizing linguistic information. Our approach fuses acoustic features with linguistic features, which is a conversion from words to vectors. The results doubled the performance of valence prediction on both single-task learning single-output (predicting valence only) and multitask learning multi-output (predicting valence, arousal, and dominance). Using a proper combination of acoustic and linguistic features not only improved valence prediction, but also improved arousal and dominance predictions in multitask learning.**

***Index Terms*—valence prediction, linguistic feature, speech emotion recognition, dimensional emotion, affective computing**

I. Introduction

Recent research on dimensional emotion recognition (continuous degree prediction of valence, arousal, and dominance) indicates that the performance of valence prediction is lower than of arousal and dominance [1]–[3]. While arousal and dominance obtained more than 0.4 of concordance correlation coefficient (CCC) scores, valence only obtained around 0.1 of CCC score [1]. It will be difficult to obtain precise emotion categories given the low performance of dimensional emotion recognition, in which Russel argued that categorical emotion can be derived from valence-arousal space [4]. In practice, categorical emotion is widely used than dimensional emotion.

Several techniques have been proposed to tackle this limitation on dimensional speech emotion recognition. Sridhar et al. [5] used higher regularization (dropout rate) for valence than for arousal and dominance. They reported an improvement from CCC score of 0.29 to 0.31 on MSP-Podcast dataset [6]. Other researchers [5] tried different techniques; however, the improvement is still low compared to the scores of arousal and dominance using the same method.

This paper proposes use of linguistic features (also known as text or lexical features), i.e., word embeddings or word vectors, to improve valence prediction in dimensional speech emotion recognition (SER). In most acoustic-based SER approaches, the performance of valence prediction is lower than arousal and dominance predictions. Although the proposed method is not new, our approach differs from reported approaches: we evaluated a network concatenation of *state-of-the-art* word embeddings and an acoustic feature set. The details of our approach is given in Methods section.

The use of linguistic features for valence prediction is borrowed from sentiment analysis research. Sentiment, according to Jurafsky [7], has the same meaning as ”valence” or ”semantic orientation.” Following the success of using word embeddings for sentiment analysis, such as in [8], we incorporate this linguistic information along with acoustic features for dimensional speech emotion recognition. This strategy aims to improve valence prediction in dimensional SER by benefiting semantic knowledge obtained from linguistic features.

The contribution of this paper is the evaluation of the addition of linguistic features to the dimensional SER system for improving the performance of valence prediction. We kept a single acoustic feature set while varying several linguistic features for combination. Acoustic and linguistic features (Acoustic+Linguistic) are trained in different networks in parallel, and both networks are concatenated with a dense network. We perform this valence prediction on single-task learning (STL) to predict the valence dimension only and multitask learning (MTL) to predict valence, arousal, and dominance dimensions. State-of-the-art of word embeddings including word2vec, GloVe, FastText, and BERT models are evaluated to investigate how these models affect valence and dimensional emotion recognition tasks. We report gains up to 104 % for STL and 129% for MTL in improving valence predictions. The best result was obtained by a combination of high-level statistical functions (HSF) from an acoustic feature set with GloVe embedding.

II. Related Work

This section summarizes previous related works and addresses remaining problems on both valence improvement and

978-1-7281-9896-5/20/$31.00 ©2020 IEEE

general speech emotion recognition.

In recent years, there have been several attempts to improve valence prediction on speech emotion recognition. In addition to the aforementioned proposal of using higher regularization, a similar approach using lexical feature was proposed by Aldeneh et al. [9] using pretrained word2vec with Mel Filterbank for the acoustic feature. However, they converted the regression task into the classification task by dividing valence scores into negative, neutral, and positive category. The authors improved unweighted average recall (UAR) from 0.59 with acoustic modality to 0.694 with acoustic-lexical modalities on the IEMOCAP dataset. Using a similar idea, Zhang et al. used acoustic and lexical features to recognize valence from the speech on three valence categories from the IEMOCAP dataset. Instead of extracting lexical features of words, the authors extracted lexical features of phonemes, i.e., 40-dimensional unique phoneme including an additional "out of vocabulary" label. The proposed method improved UAR from 0.64 with acoustic-only modality to 0.74 with acoustic-lexical (phonemes) modality.

Instead of predicting categories of valence only, research on predicting categories and continuous degrees of emotion is more familiar. In [10], the authors used semantic features from the affective dictionary along with the MFCC features to predict valence and arousal. A deduction of mean absolute error (MAE) from 1.98 to 1.40 for valence and from to 1.29 to 1.28 was reported using the proposed acoustic-semantic combination over acoustic only. In other research [11]–[14], the authors used different deep learning architectures to predict categorical emotion from both speech and text. Some authors used phonemes instead of text for predicting emotion category [15], [16], and some authors compared text feature from automatic speech recognition (ASR) with manual transcription to investigate its effectiveness its combination with acoustic features for categorical emotion recognition [17].

We know of no authors who reported valence improvement on continuous-degree dimensional emotion recognition. When valence is used as the target of the prediction task, its continuous values are converted to discrete categories. When acoustic and text features were combined, different text features (e.g., semantics from an affective dictionary) is used with. Most approaches on that acoustic-linguistic combination also evaluated categorical emotion instead of dimensional emotion. This research reports an evaluation of various word embedding models, including state-of-the-art models, in combination with an acoustic feature set for continuous-degree dimensional emotion recognition, particularly on valence dimension.

III. Methods

A. Dataset

The IEMOCAP dataset developed by Busso et al. [18] was used to evaluate the proposed method. This dataset contains speech and text modalities. The proposed method used utterance-based audio files (apart from dialog-based audio files) to extract acoustic features and dimensional labels. The original labels are in 5-scale range, and we normalized those labels to floating points in the range [-1, 1] when feeding them into classifiers, following the work of [1]. All data were used, totaling 100039 utterances and labels. A MinMax preprocessing method removed the outliers by inflating labels below 1 to 1 and deflating labels above 5 to 5 [19]. This processing step followed the scale described in the reference paper [18] because we found some labels are inconsistent with the stated 5-level scale. We used the manual text transcription provided in the dataset to generate the linguistic features, i.e., word vectors, which are evaluated to improve valence prediction. Sahu et al. [17] reported that the difference in performance between automatic and manual transcriptions for categorical emotion with Acoustic+Linguistic in that dataset is 4 % using Google ASR, thanks to advancement in speech recognition technology.

We split the dataset into training and test partition; session 1 – 4 are used for training while session 5 is left for a test. This scenario is speaker independent strategy for a cross validation. In training set, 20% of data were used for validation.

B. Feature sets

Our method utilized two types of feature sets, acoustic and linguistic features.

Acoustic features: pyAudioAnalysis (pAA), an open-source Python library for audio signal analysis, was used to extract 34 low-level descriptors (LLD). Although this tool is not designed for affective analysis, several authors reported the effectiveness of the acoustic features extracted from that tool in speech emotion task [17], [20]. In [17], the authors obtained comparable results between that pAA and affective-designed GeMAPS feature sets [21]. By default, pAA extracts 34 LLDs for a given audio file. In this research, we did not use those LLDs directly. Instead, we used high-level statistical functions (HSF), which are extracted from those LLDs. Note that the definition of HSF here only refer to mean and standard deviation (Mean+Std) of LLDs, since a report suggested that both functionals performed better than the whole set of audio features (LLDs + HSFs) and audiovisual features [22]. Instead of extracting HSF from GeMAPS, we extracted HSF from the pAA feature set. Hence, the input feature set is a 68-dimensional feature vector for each utterance. Table I shows a list of LLDs in the pAA feature set. Note that only HSF features are used in this research.

Linguistic features: Linguistic information/features are extracted from speech data. We defined linguistic feature here as vector representation of every word in each sentence in the dataset. In addition to directly using word embeddings (word vectors) from the conversion of vocabularies to vectors in the dataset [23], we also used pretrained word embeddings from other models, that are trained in larger datasets (e.g., Wikipedia), to weigh original word embeddings. Those pretrained models are Word2Vec [24], GloVe embeddings [25], FastText [26], and BERT [27]. All pretrained models have a 300-dimensional vector per token including the original word embedding, excluding the BERT model. The BERT pretrained model has a 768-dimensional vector for each word.

TABLE I
PYAUDIOANALYSIS FEATURE SET (PAA) [28]; ONLY HSFS ARE USED AS INPUT FEATURES.

LLDs	zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectra flux, spectral roll-off, 13 MFCCs, 12 chroma vectors, chroma deviation.
HSFs	Mean, Std

The maximum sequence length was limited to 554 based on the longest sequence; utterances with number of words lower than this number is padded with zeros to achieve the same length.

C. Dimensional Speech Emotion Recognition System

Several authors have proposed utilizing text with acoustic features to improve performance of speech emotion recognition, including a multimodal feature pooling [9] and an early-late fusion technique [13]. The first report converted continuous valence degree to discrete (categorical) classes while the latter methods only applied to categorical emotion. The proposed approach in this paper maintains continuous prediction as the target (floating-point numbers) of regression task. The proposed method used some state-of-the-art of word embeddings, including BERT embedding, while others only used one or two types of pretrained embeddings (such as in [9] using Word2Vec, [29] using FastText, and [12] using GloVe embedding). Figure 1 shows an overview of our approach on using acoustic and text features for predicting valence (V), arousal (A), and dominance (D).

Figure 1 shows our system processed acoustic and text features in parallel, and concatenated both networks with a dense network. We used a similar network for both acoustic and text networks with three LSTM layers stacked in rows. The difference is the number of nodes in the first layer, in which we used the same dimension as the input feature. A part of Figure 1 bounded by the dashed line shows the acoustic-only dimension emotion recognition, the baseline of this research. A single-task learning with single output, which predicts valence degree only, is shown with a gray background.

Acoustic Network. We transmitted a 68-dimensional acoustic vector from mean and std of pyAudioAnalysis acoustic feature set to the acoustic network: three stacked LSTM layers. Before entering the first LSTM stack, we performed batch normalization on those acoustic features to accelerate the computation process [30]. The batch normalized inputs are then fed into an LSTM layer with 68 units, the same as the size of the acoustic features. The other two 256-unit LSTM layers are added into the acoustic network shaping a stack of three LSTM layers. The final output (instead of full sequence output) of the last LSTM layer is connected to a dense network with 64 units as a final layer of the acoustic network. We do not use the dropout layer on this network since the size of the input features is small.

Text Network. Word embeddings as linguistic features are the input of the text network. All word embeddings have 300-dimensional vectors, except for the BERT model, which has 768-dimensional vectors. All word embeddings are fed into an embedding layer before entering the first LSTM stack. Except for the original word embeddings (WE), all embeddings are weighted by pretrained models. The size of the first LSTM layer is the same as the size word embeddings, i.e., 300 or 768. The other two 256-unit LSTM layers are added with and coupled with a dense layer with 64 units, same as the acoustic network. We used a dropout rate of $p = 0.4$ based on our experiment results for this text network.

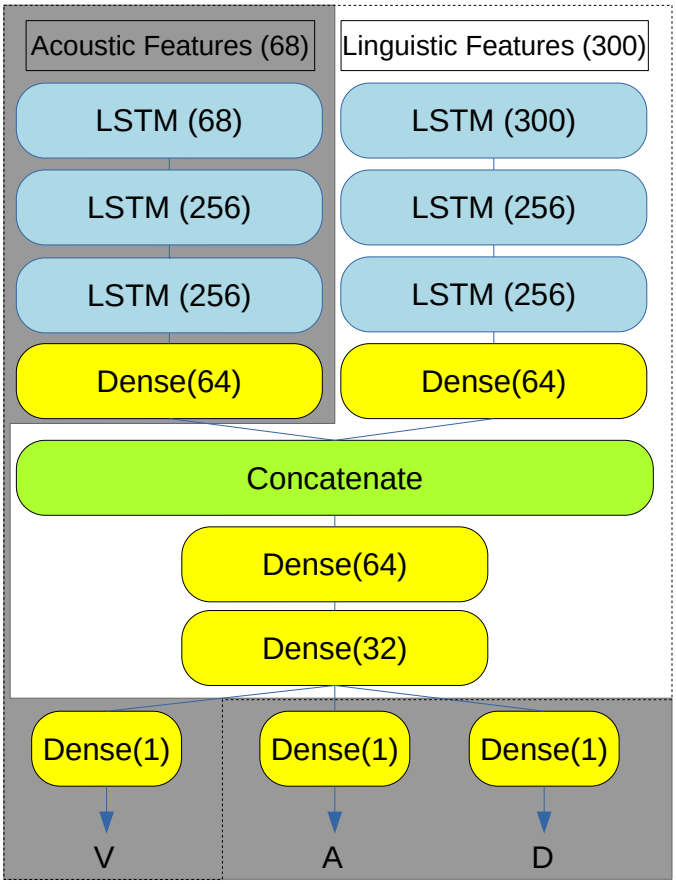

Fig. 1. Dimensional speech emotion recognition system using acoustic and text features; numbers in brackets represent the number of nodes; dashed line: valence-only prediction; gray background: acoustic-only emotion recognition.

Concatenation Network. We concatenated both acoustic and text networks and fed acoustic and linguistic features into those networks. Mathematically, the combined Acoustic+Text network, the "Dense(64)" after "Concatenate" in Figure 1, was formulated as in equation 1. Here, $f(y)$ denotes the output of the corresponding layer; W_1, W_2 denote the weights from previous layers (a: acoustic; t: text), i.e., dense layer after LSTM for each network, and the current hidden layer, respectively; x_a and x_t are the acoustic features and word embeddings, respectively; b is a bias; and g is an activation function. Thus, the output of the that dense layer was

$$f(y) = W_2 g([W_{1a}^\top xa + b_{1a}; W_{1t}^\top xt + b_{1t}]) + b_2. \quad (1)$$

The next dense layer works in similar way. Three dense layers with one unit will output the prediction score of valence, arousal, and dominance. The error between the predicted emotion attributes and the gold-standard labels are minimized using CCC loss function (CCCL), i.e., for STL, CCCL for valence ($CCCL_V$) is defined as,

$$CCCL_V = 1 - CCC_V. \tag{2}$$

A CCC to measure concordance correlation between two variables is formulated as follows [31],

$$CCC = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2\sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{3}$$

where ρ_{xy} is the Pearson correlation coefficient (PCC) between two variables, σ is a standard deviation, and μ is a mean value.

For MTL, the total loss function is the sum of CCC losses from three emotion dimensions. We weighted each emotion dimension in MTL with a parameter modified from [32] and [1],

$$CCCL_T = \alpha CCCL_V + \beta CCCL_A + (1 - \alpha - \beta)CCCL_D. \tag{4}$$

Using a linear search, we found the optimum parameters are $\alpha = 0.7$ and $\beta = 0.2$ for both linguistic-only and acoustic-linguistic networks, while $\alpha = 0.1$ and $\beta = 0.4$ worked best for the acoustic-only networks.

We optimized acoustic-only emotion recognition using an Adam optimizer with Nesterov momentum (nadam) [33]. For Acoustic+Text emotion recognition, we optimized the model with RMSprop [34], since this optimizer gave better results than the previous Adam optimizer in our experiments. This model was implemented in Keras toolkit [35] with TensorFlow backend [36]. The Python codes to run these experiments are available at https://github.com/bagustris/dser_with_text.

IV. Results and Discussion

In the experiments, we set fixed random numbers for each evaluated method in Python files. However, the use of linguistic features resulted in different CCC scores on each run. The reported scores below are an average of 20 experiments. The acoustic-only method (HSF) obtained the same score for every run thanks to seed number initialization.

A. Performance of Acoustic vs. Acoustic+Linguistic

Table II shows our result using linguistic features to improve valence prediction. We have divided our results on that table into two parts: results obtained by STL to predict valence, and results obtained by MTL to predict valence, arousal, and dominance, simultaneously. The use of text feature improves valence prediction by a remarkable margin: the best result from Acoustic+Linguistic features, i.e., HSF+GloVe, doubled the valence prediction of the acoustic-only feature (HSF). All combinations of acoustic features with any word embedding demonstrated improvement on both STL and MTL approaches.

Since there is only one parameter to be predicted in STL, the valence prediction may result better than MTL on the same size input features. Surprisingly, there are no remarkable differences between the results obtained by STL and MTL in our results. Generally, STL still obtained higher performance by a very small margin than MTL. However, given the same running time and resources (including the same input size), obtaining valence, arousal, and dominance is more beneficial than obtaining valence only. Our results suggest that instead of predicting valence only using STL, predicting all emotion dimensions using MTL is recommended for future research direction. There is no significant difference between results (STL vs. MTL) for improving valence prediction by incorporating text features in dimensional SER.

In MTL, the result obtained by HSF+GloVe improves not only the prediction of valence, but also prediction of arousal and dominance. All combinations of Acoustic+Linguistic features improved the prediction for valence and dominance. If we used averaged CCC scores from valence, arousal, and dominance as a single metric to determine overall performance among all emotion dimensions, all combinations of Acoustic+Linguistic features gain performance improvements from acoustic-only dimensional speech emotion recognition.

In comparing pretrained word embeddings to weigh the original tokens, we found that the model trained by GloVe embeddings achieved higher CCC scores than other models. A newer word embedding model than GloVe, i.e., FastText and BERT embeddings, cannot surpass the result obtained by GloVe embedding in this affect recognition. Note that in our implementation, we did not perform the fine-tuning of the BERT model. Instead, we used the same architectures as other Acoustic+Text networks used for the BERT pretrained model [27]. Although the dimension of the BERT model is larger than other word embedding models (i.e., 768 vs. 300), the result is still lower than GloVe embedding, but higher than that of any other model. Fine-tuning the BERT model may improve the performance of valence prediction as suggested in other classification tasks [37].

To this end, it is shown that linguistic information helps to improve valence prediction. The semantic of spoken word contains emotional contents which are translated into word embeddings and their pretrained models. Fusing acoustic and linguistic information is a straight forward way to improve dimensional emotion recognition since linguistic information also can be extracted from speech.

B. Relative Improvements

Since the goal of this paper is to report improvements of valence prediction, we have included relative improvements obtained by Acoustic+Linguistic over acoustic-only dimensional speech emotion recognition. Figure 2 shows these results. All Acoustic+Linguistic concatenation results with pretrained word embedding obtained relative improvement more than 80%. The highest performance, obtained by HSF+GloVe, doubled the performance on both STL and MTL valence predictions. The relative improvements obtained by MTL are

TABLE II
CCC SCORES OF VALENCE (V), AROUSAL (A), AND DOMINANCE (D) FROM ACOUSTIC FEATURE (HSF) VS. ACOUSTIC AND LINGUISTIC FEATURES (HSF+); STL: SINGLE-TASK LEARNING SINGLE OUTPUT; MTL: MULTITASK LEARNING MULTI OUTPUT.

Methods	**STL**	**MTL**			
	V	V	A	D	Mean
HSF	0.208	0.183	0.577	0.444	0.401
HSF+WE	0.363 ± 0.007	0.364 ± 0.010	0.565 ± 0.022	0.474 ± 0.014	0.468
HSF+Word2Vec	0.380 ± 0.013	0.387 ± 0.012	0.558 ± 0.019	0.471 ± 0.016	0.472
HSF+FastText	0.380 ± 0.011	0.374 ± 0.011	0.561 ± 0.021	0.475 ± 0.013	0.470
HSF+GloVe	0.424 ± 0.010	0.421 ± 0.008	0.590 ± 0.008	0.484 ± 0.009	0.498
HSF+BERT	0.380 ± 0.016	0.377 ± 0.026	0.574 ± 0.021	0.482 ± 0.017	0.478

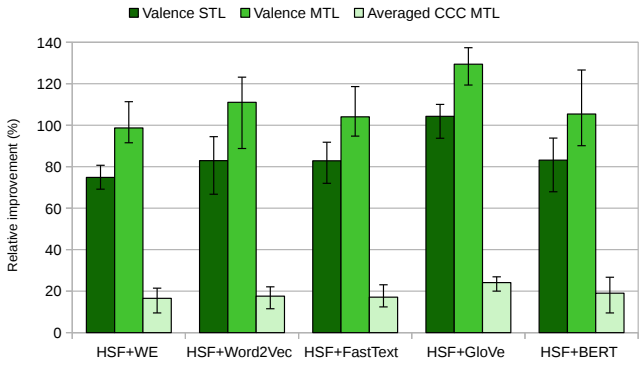

Fig. 2. Relative improvement of Acoustic+Linguistic features over acoustic-only feature with different pretrained word embedding weights in single-task and multitask learnings; the error bars shows variances.

higher than the relative improvements obtained by STL. The proposed dimensional emotion recognition utilizing acoustic and linguistic features improved not only valence prediction but also overall dimensional emotion prediction, as indicated by averaged CCC scores. Although they are not shown in the table, these averaged CCC scores can be obtained by averaging the scores of V, A, and D in the last three columns of Table II. Note that although all Acoustic+Linguistic pairs obtained averaged CCC scores, only the pair of HSF+GloVe obtained improvement on all emotion dimensions. The rest of Acoustic+Linguistic pairs improved only on valence and dominance predictions.

We observe that our approach using linguistic feature for dimensional SER improved continuous-degree valence prediction than any of previous reported result. Although the performance of STL is higher than MTL (Table II), the relative improvements in MTL are higher than in STL on predicting valence (Figure 2). Hence, the MTL approach is suggested for future research for improving the prediction of all emotional dimensions.

V. CONCLUSIONS

This paper reported on the use of linguistic features to improve valence prediction in dimensional speech emotion recognition. The proposed approach doubled the previous level of performance of valence prediction in single-task and multitask learnings. When using the latter learning method, not only was performance of valence improved, but also the performance of dominance by combinations of the acoustic features with any evaluated word embedding. By using a proper word embedding weight, i.e., GloVe embedding, improvements of valence, arousal, and dominance measures were obtained from acoustic-only approach. By including the linguistic features, our approach doubled the performance of the baseline valence prediction on both STL and MTL approaches. Although the performance was improved, the current results on dimensional speech emotion recognition still lack advancement compared to other machine learning implementations. Further studies are required to improve these performances to close to human annotations.

REFERENCES

[1] S. Parthasarathy and C. Busso, "Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning," in *Interspeech*, 2017, pp. 1103–1107.

[2] M. AbdelWahab and C. Busso, "Domain Adversarial for Acoustic Emotion Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 12, pp. 2423–2435, dec 2018.

[3] N. Tits, K. E. Haddad, and T. Dutoit, "ASR-based Features for Emotion Recognition: A Transfer Learning Approach," in *Proc. ofthe First Gd. Chall. Work. Hum. Multimodal Lang.*, 2018, pp. 48–52.

[4] J. A. Russell, "Affective space is bipolar," *J. Pers. Soc. Psychol.*, 1979.

[5] K. Sridhar, S. Parthasarathy, and C. Busso, "Role of Regularization in the Prediction of Valence from Speech," in *Interspeech 2018*. ISCA: ISCA, sep 2018, pp. 941–945.

[6] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, 2019.

[7] D. Jurafsky and J. H. Martin, "Lexicons for Sentiment and Affect Extraction," in *Speech Lang. Process. An Introd. to Nat. Lang. Process. Comput. Linguist. Speech Recognit.*, 3rd ed., 2017, pp. 326–345.

[8] Y. Li, ·. Quan Pan, T. Yang, S. Wang, ·. J. Tang, E. Cambria, Q. Pan, and J. Tang, "Learning Word Representations for Sentiment Analysis," vol. 9, pp. 843–851, 2017. [Online]. Available: https://sentic.net/learning-word-representations-for-sentiment-analysis.pdf

[9] Z. Aldeneh, S. Khorram, D. Dimitriadis, and E. M. Provost, "Pooling acoustic and lexical features for the prediction of valence," in *ICMI 2017 - Proc. 19th ACM Int. Conf. Multimodal Interact.* ACM, 2017, pp. 68–72.

[10] S. G. Karadogan and J. Larsen, "Combining semantic and acoustic features for valence and arousal recognition in speech," in *2012 3rd Int. Work. Cogn. Inf. Process.* IEEE, may 2012, pp. 1–6.

[11] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, no. September, pp. 247–251, 2018.

[12] S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text," *Spok. Lang. Technol. Work.*, pp. 112–118, oct 2018. [Online]. Available: http://arxiv.org/abs/1810.04635

[13] J. Sebastian, P. Pierucci, and T. L. Gmbh, "Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts," *Interspeech*, pp. 51–55, 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3201

[14] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech Emotion Recognition Using Speech Feature and Word Embedding," in *2019 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Lanzhou, 2019, pp. 519—-523.

[15] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Interspeech 2018*, 2018, pp. 3688–3692.

[16] B. Zhang, S. Khorram, and E. M. Provost, "Exploiting Acoustic and Lexical Properties of Phonemes to Recognize Valence from Speech," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May. Institute of Electrical and Electronics Engineers Inc., may 2019, pp. 5871–5875.

[17] S. Sahu, V. Mitra, N. Seneviratne, and C. Espy-Wilson, "Multi-modal learning for speech emotion recognition: An analysis and comparison of ASR outputs with ground truth transcription," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2019, pp. 3302–3306.

[18] C. Busso, M. Bulut, C.-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," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825—-2830, 2011.

[20] S. Tripathi and H. Beigi, "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning," *CoRR*, apr 2018. [Online]. Available: http://arxiv.org/abs/1804.05788

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, apr 2016.

[22] M. Schmitt and B. Schuller, "Deep Recurrent Neural Networks for Emotion Recognition in Speech," in *DAGA*, 2018, pp. 1537–1540.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," jan 2013.

[25] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Conf. Empir. Methods Nat. Lang. Process.*, 2014, pp. 1532–1543.

[26] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval.*, dec 2019, pp. 52–55.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv Prepr. arXiv*, oct 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[28] T. Giannakopoulos, "pyAudioAnalysis: An open-source python library for audio signal analysis," *PLoS One*, vol. 10, no. 12, pp. 1–17, 2015. [Online]. Available: https://github.com/tyiannak/pyAudioAnalysis/

[29] M. I. Torres, R. Justo, M. Carrilero, M. de Velasco, and J. Antón, "Emotion Detection from Speech and Text," no. November, pp. 68–71, 2018.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, 2015, pp. 448–456.

[31] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–68, 1989.

[32] B. T. Atmaja and M. Akagi, "Multitask Learning and Multistage Fusion for Dimensional Audiovisual Emotion Recognition," in *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2020, pp. 4482—-4486. [Online]. Available: https://ieeexplore.ieee.org/document/9052916/

[33] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, dec 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[34] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural networks Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[35] F. Chollet and Others, "Keras," https://keras.io, 2015.

[36] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and Others, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symp. Oper. Syst. Des. Implement. (OSDI '16)*, 2016, pp. 265–283.

[37] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?" in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2019.