EXPLOITING ACOUSTIC AND LEXICAL PROPERTIES OF PHONEMES TO RECOGNIZE VALENCE FROM SPEECH

Biqiao Zhang[⋆] *Soheil Khorram*[†] *Emily Mower Provost*[⋆]

[⋆] University of Michigan, Ann Arbor, Michigan
[†] University of Texas, Dallas, Texas

ABSTRACT

Emotions modulate speech acoustics as well as language. The latter influences the sequences of phonemes that are produced, which in turn further modulate the acoustics. Therefore, phonemes impact emotion recognition in two ways: (1) they introduce an additional source of variability in speech signals and (2) they provide information about the emotion expressed in speech content. Previous work in speech emotion recognition has considered (1) or (2), individually. In this paper, we investigate how we can jointly consider both factors to improve the prediction of emotional valence (positive vs. negative), and the relationship between improved prediction and the emotion elicitation process (e.g., fixed script, improvisation, natural interaction). We present a network that exploits both the acoustic and the lexical properties of phonetic information using multi-stage fusion. Our results on the IEMOCAP and MSP-Improv datasets show that our approach outperforms systems that either do not consider the influence of phonetic information or that only consider a single aspect of this influence.

***Index Terms*—** Speech Emotion Recognition; Audio and Phonemes; Convolutional Neural Networks

1. INTRODUCTION

Emotions modulate acoustic signals both explicitly, through paralinguistic characteristics (e.g., the tone and tempo of speech), and implicitly, through the alteration of the content of speech. Therefore, speech content is a double-edged sword in emotion recognition: the variability it introduces to the acoustic signals makes it harder to distill emotion-related cues, yet the content itself is reflective of emotion. In this paper, we explicitly consider both roles of speech content and demonstrate that, in so doing, we are able to make more accurate predictions of emotional valence (positive vs. negative).

We present a speech emotion recognition (SER) system that considers: (1) the acoustic variability in terms of both emotion and speech content, here defined as sequences of phonemes, and (2) the direct connection between emotion and phoneme sequences. We investigate whether leveraging both (1) and (2) leads to improved performance. We concentrate on predicting valence (the positive vs. negative aspect of an emotional display [1, 2]) because it has been shown to be difficult given only acoustic signals [3].

Previous research has investigated how phonemes modulate acoustics together with emotion by exploring phoneme-level emotion classification methods [4–6], or designing acoustic features [7–11] or labels that incorporate phonetic knowledge [12]. The results of these studies showed that phonemes vary in how they are modulated by emotion and that features designed based on phonetic knowledge work well in emotion recognition. Recent works have shown that emotion can be predicted directly from sequences of phonemes without acoustic information, by modeling phoneme sequences like word sequences [13], using LSTM networks [14], or multi-channel CNN networks [15]. These works have also shown that combining utterance-level phonetic and acoustic representations brings further improvement. However, work that considers both the phonetic modulation of acoustics and the link between phoneme sequences and emotions is still missing. In addition, we do not yet know how models that exploit the acoustic and/or phonetic properties of phonemes is influenced by emotion elicitation method (i.e., fixed, improvised under targeted scene, spontaneous).

In this work, we seek to improve valence prediction by leveraging the dual-functionality of phonemes, using temporal Convolutional Neural Networks. We hypothesize that adding phonetic information at different stages has different effects and that we can exploit both the acoustic and the lexical properties using a *multi-stage fusion* model that combines acoustic and phonetic information at both feature-level (*feature fusion, FF*) and utterance-level (*intermediate fusion, IF*). We investigate how models leveraging phonetic information at different stages are influenced by the emotion elicitation process of the data. We test our hypothesis on the IEMOCAP dataset [16] and the MSP-Improv dataset [17].

Our results show that our multi-stage fusion model outperforms both FF and IF models, especially on data produced using improvisations and natural interactions. We also find that both FF and IF are beneficial compared to unimodal models, and that IF outperforms FF. However, the advantage of modeling phoneme sequences independently, either in the unimodal phonetic model or IF, decreases as the lexical content becomes more spontaneous, indicating that this advantage may come from memorizing emotionally salient patterns in speech content. The novelty of this paper includes the presentation of: (1) a multi-stage fusion approach that exploits the dual-functionality of phonemes and (2) an investigation into the influence of the type of lexical content on the performance of the models leveraging different functions of phonemes.

2. DATA

2.1. Datasets

We use two English dyadic emotion datasets: IEMOCAP and MSP-Improv. We choose these datasets because: (1) their sizes allow us to train neural networks; (2) they provide evaluations of valence; (3) they contain varying lexical patterns due to the use of different emotion elicitation methods, allowing us to conduct relevant analyses.

IEMOCAP: The IEMOCAP dataset consists of five sessions of dyadic interactions, each between a male and a female actor. The 12 hours of data were segmented into 10,039 utterances according to speaker turns. The emotions of the speakers were elicited through scripted and improvised scenes. The lexical content of

scripted scenes and the improvisation targets of the improvised scenes were shared across sessions. The scripted and improvised portions of IEMOCAP consist of 5,255 and 4,784 utterances, respectively. The valence of each utterance was assessed by at least two evaluators using a 5-point scale [16]. We create phone-level transcriptions by force aligning the provided manual transcriptions to the audio files (see Section 2.2). We exclude six utterances for which forced alignment failed. We conduct experiments: over the entire dataset (*IEMOCAP-all*), on only the scripted utterances (*IEMOCAP-scripted*), and on only the improvised utterances (*IEMOCAP-improv*).

MSP-Improv: The MSP-Improv corpus contains six sessions of dyadic interactions between pairs of male-female actors. There are nine hours of speech and 8,438 utterances. The data elicitation includes both improvisations and target sentences embedded in designed scenarios. The valence of each utterance is assessed using a 5-point scale by at least five evaluators [17]. We use the automatic transcriptions produced by the Bing Speech API[1], provided by the creator of the dataset. We focus on the improvisations and the natural interactions and only use utterances that have transcriptions in our experiments. This decreased our data to 5,650 utterances, which we refer to as *MSP-I+N*. We choose to exclude target sentences and not to perform experiments for the improvised and natural partitions separately due to the limited size of the partitions.

2.2. Data Preprocessing

Labels: We convert the 5-point ratings into three categories: negative, neutral, and positive, and generate fuzzy labels for each utterance as in [18, 19]. We represent each evaluation as a 3-dimensional one-hot vector by keeping 3 as "neutral" and merging 1-2 and 4-5 as "negative" and "positive", respectively. We then use the mean over the evaluations for each utterance as the ground truth. For instance, given an utterance with three evaluations, 3, 4, and 5, we first convert the evaluations to [0, 1, 0], [0, 0, 1], and [0, 0, 1], respectively. After taking the mean, the ground truth label for this utterance is [0, 1/3, 2/3]. In this way, we form the problem of valence recognition as a three-way classification task.

Acoustic Features: We extract 40-dimensional log Mel-frequency Filterbank energy (MFB) using Kaldi [20]. The MFBs are computed over frames of 25ms, with a step size of 10ms, as in [19, 21, 22]. We perform speaker-dependent z-normalization at the frame-level.

Phonemes: We acquire the start and end time of each phoneme by using forced alignment between the audio and the manual (IEMOCAP) or automatic (MSP-Improv) transcriptions. We use Gentle[2], a Kaldi-based forced aligner. It identifies 39 unique phonemes and an additional "out of vocabulary" label for unrecognized sounds, resulting in a 40-dimensional one-hot vector for each phoneme. The phonetic representations are used in two different ways: (1) independently without repetition, and (2) repeated and with the same step-size as acoustic features. See more details in Section 3.1.

3. METHODOLOGY

3.1. Network Structures

We design our models based on the temporal Convolutional Neural Network with global pooling (*Conv-Pool*) structure, which has been demonstrated to perform well in [21, 22]. Figure 1 shows the architectures of our networks. These networks consist of the following components (Figure 1(a)):

- A Conv-Pool sub-network (i.e., a 1D convolutional layer over time and a global max-pooling layer) that generates a fixed-length utterance-level representation from the variable-length input of acoustic and/or phonetic features.
- A concatenation of the multiple utterance-level representations (denoted as "Cat" in Figure 1 and "+" in model names).
- An optional dropout layer, two fully-connected layers and a softmax layer (denoted as *FC*).

There are three Conv-Pool branches in our networks: the acoustic branch (*Ab*), the phonetic branch (*Pb*), and the feature-fusion branch (*APb*). *Ab* and *Pb* operate on variable-length MFB features and phoneme sequences, respectively. In *APb*, we aim to capture the phonetic modulations of acoustic features. We concatenate the phoneme label with the MFBs at each frame. For example, if a specific phoneme lasts 0.1 seconds, the same one-hot vector is concatenated with the MFB features of the ten corresponding frames. For audio frames with no matching phoneme, a zero-vector is used instead. The number of input channels of the convolutional layer is 40, 40, and 80 for *Ab*, *Pb*, and *APb*, respectively. Feeding the output of a single branch to the *FC* sub-network results in three models (Figure 1(b)): two unimodal models (i.e., *Ab_FC* and *Pb_FC*), and a multimodal single-stage feature-fusion model (*APb_FC*).

We concatenate the outputs of *Ab* and *Pb* for joint modeling in *FC* to captures the high-level interaction between the learned acoustic and phonetic representations. This results in our multimodal single-stage intermediate-fusion model, *Ab+Pb_FC* (Figure 1(b)).

We hypothesize feature fusion and intermediate fusion play different roles in the network. Feature fusion allows our network to capture how phonemes modulate acoustics. However, it may not be effective in linking speech content and emotional state, specifically, in extracting phoneme sequences that are informative identifiers of valence. This is because: (1) each single phoneme may be repeated several times in order to have the same step-size with the MFBs, resulting in insufficient temporal context for the phoneme sequences in the convolution layer; (2) the input phoneme sequences are much more sparse than the MFBs, resulting in representations dominated by acoustic information. Intermediate fusion, on the other hand, can more efficiently leverage the complementary emotionally salient information learned from audio and phoneme sequences. Because of the dual-functionality of phonemes, we propose to combine them into a multi-stage fusion model to exploit the advantages of both techniques. This network, *APb+Pb_FC*, concatenates the representations from *APb* and *Pb* and jointly models them in *FC*. In addition, we explore another multi-stage fusion network, *APb+Ab_FC*, which concatenates *APb* and *Ab* for comparison. Both *APb+Pb_FC* and *APb+Ab_FC* are shown in Figure 1(b).

3.2. Hyper-parameters and Training Strategy

We use ReLU as the activation function in all layers but the output layer, where softmax is used. We select the layer size of the convolutional and fully-connected layers from $\{128, 256\}$ as in [22]. The layer size is kept consistent throughout each model. We fix the kernel width to 16 for *Ab* and *APb*, which is shown to perform well on both IEMOCAP and MSP-Improv in [21]. For *Pb*, we select a kernel width of 6, based on the average number of phonemes (6.38) per English word, according to the CMU pronunciation dictionary[3].

[1] https://azure.microsoft.com/en-us/services/cognitive-services/speech/

[2] https://lowerquality.com/gentle/

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

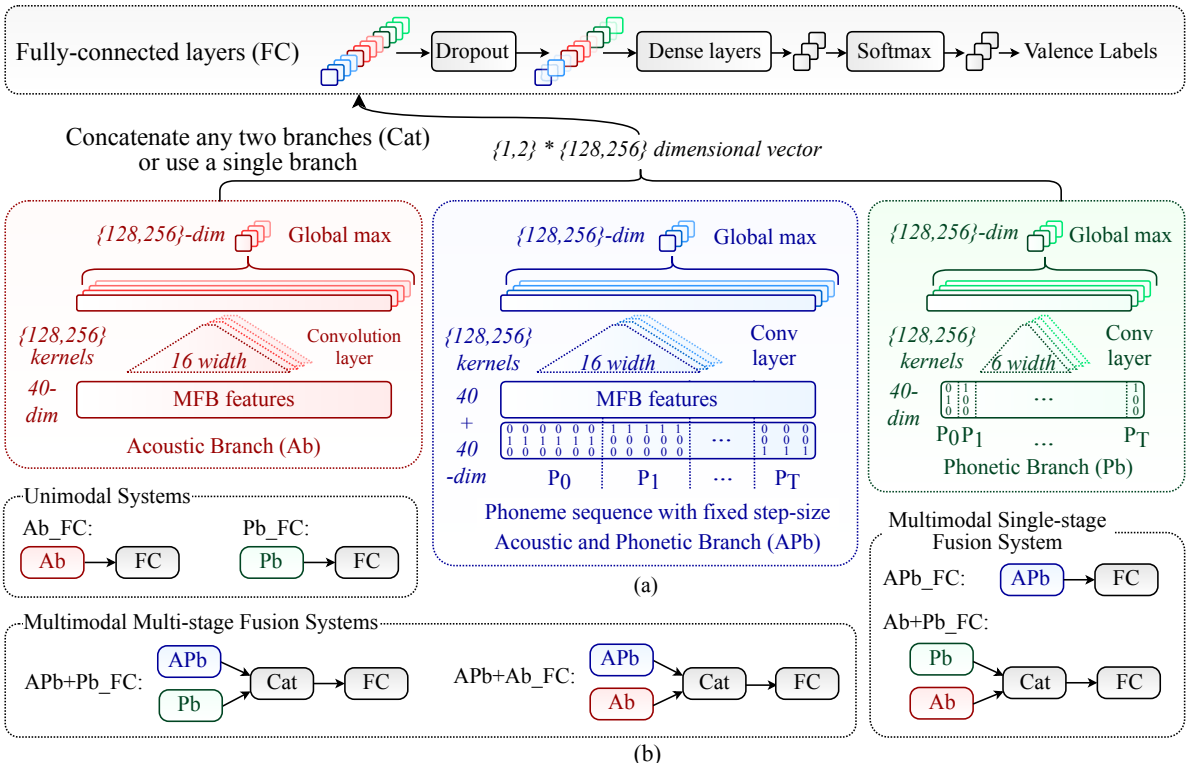

Fig. 1. (a) A general network that illustrates all the components, including: the acoustic branch (*Ab*), the phonetic branch (*Pb*), the acoustic and phonetic branch (*APb*) that combines the features of the two modalities; the concatenation of the utterance-level representations (*Cat*), and a stack of dropout, fully-connected and softmax layers (*FC*). (b) Architectures for all models.

Besides, we incorporate an optional dropout layer after the global max-pooling to improve generalization of the networks. The dropout probability is selected from $\{0, 0.2, 0.5\}$, where 0 corresponds to no dropout, and 0.2 and 0.5 are from the suggested range in [23].

We experimented using PyTorch version 0.2.0. The loss function is cross-entropy computed using the fuzzy labels. We weigh the classes using $N/(3 * \sum_{j=1}^{N} gt_j^i)$ in the loss calculation, where N is the total number of training utterances, gt_j^i is the value for class i in the fuzzy ground truth label for utterance j. We train the models using a learning rate of 0.0001 with the Adam optimizer [24].

We use Unweighted Average Recall (UAR) as the performance measure due to unbalanced data [25]. When the ground truth has ties, we deem predictions for any of the tied positions as correct, as in [19]. For instance, when the ground truth is [0.5, 0.5, 0], prediction of either 0 or 1 are correct. As a result, the chance performance of making predictions uniformly at random is higher than 33.33%.

We use the leave-one-speaker-out evaluation setting for our experiment. Both IEMOCAP and MSP-Improv are organized by sessions. At each round, we left out data from a single speaker as the test set, use data from the other speaker in the same session for validation, and use data from the remaining sessions for training. We run each experiment five times to reduce performance fluctuation. For each training-validation-testing combination, we select the number of training epoch ($\in [1, 30]$) by maximizing the validation UAR for each run separately and select the layer size and dropout probability by maximizing the validation UAR averaged over five runs. We report the test UAR corresponding to the chosen hyper-parameters, averaged over speakers and runs. We set the batch size to 100, and zero-pad the features to the maximum length of each batch.

4. RESULTS AND DISCUSSIONS

We present the average test UAR for the experiments on IEMOCAP-all, IEMOCAP-scripted, IEMOCAP-improv, and MSP-I+N in Table 1, together with the chance performance calculated by making predictions uniformly at random. For the results of each experiment, we first test if the influence of model is significant by using a repeated-measure ANOVA (RANOVA). We treat the per-speaker performance as the "subject" and model as the within-subject factor. We report the statistics in Table 1. We find that the influence of model is significant in all experiments when asserting significance at $p<0.05$, even with the lower bound correction. We compare pairs of models across experiments to understand the effect of each approach and the influence of the type of lexical content. We use Tukey's honest test based on the RANOVA model for these pairwise comparisons and assert significance at $p<0.05$.

4.1. Unimodal Results

We find that *Pb_FC* significantly outperforms *Ab_FC* on IEMOCAP-all and IEMOCAP-scripted, while *Ab_FC* significantly outperforms *Pb_FC* on MSP-I+N. It is clear that *Pb_FC* performs better than *Ab_FC* when all the data or a large portion of the data are scripted, while the opposite is true when there is less control on the lexical content of the data (i.e., improvisations and natural interactions). In fact, *Pb_FC* achieved the highest performance among all models on IEMOCAP-scripted. It is interesting to see that when emotion-related scripts are repeated across training, validation, and testing data, additional information from the acoustic modality brings more harm than good. This indicates that Conv-Pool with phoneme se-

Model	IEMOCAP -all	IEMOCAP -scripted	IEMOCAP -improv	MSP -I+N
Chance	45.40	46.91	44.55	36.09
Ab_FC	64.04	61.18	65.00	51.84†
Pb_FC	69.18∗	**78.42**∗◇	62.50	47.54
APb_FC	67.17∗	67.21∗	67.68∗†	53.98∗†
Ab+Pb_FC	73.33∗ † ◇	75.09∗◇	69.13∗†	54.99∗†
APb+Pb_FC	**73.79**∗ † ◇	75.34∗◇	**70.05**∗ † ◇	**55.98**∗ † ◇
APb+Ab_FC	67.09∗	65.54∗	67.44∗	54.34∗†
$F(5, 45/55)$	70.3	55.4	19.6	25.6
p_{LB}	1.52e-5	3.92e-5	1.66e-3	3.67e-4

Table 1. The average test UAR and the statistics of RANOVA (F and p_{LB}) for the influence of model. The best result in each experiment is bolded. $F(5, 45)$ and $F(5, 55)$ are for experiments on IEMOCAP and MSP-I+N, respectively. p_{LB} is the p-value with lower bound correction. ∗, †, and ◇ represent that the marked model significantly outperforms *Ab_FC*, *Pb_FC*, and *APb_FC*, respectively, using Tukey's honest test and asserting significance at p$<$0.05.

quence can learn and memorize speech-content-related patterns that are strongly associated with emotion classes, but does not work as well as acoustics on unscripted/natural data.

4.2. Single-stage Fusion Results

The feature-fusion model (*APb_FC*) significantly outperforms *Ab_FC* in all four experiments. However, *APb_FC* only significantly outperforms *Pb_FC* on IEMOCAP-improv and MSP-I+N, while shows significant performance loss on IEMOCAP-scripted. In addition, the performance of *APb_FC* is very stable across the different portions of IEMOCAP. These results support our hypothesis that in feature fusion, the phonetic information is helpful for learning emotion-salient acoustic representations, but cannot effectively capture the emotion-related patterns in speech content.

The intermediate-fusion model (*Ab+Pb_FC*), on the other hand, shows significant improvement compared to both *Ab_FC* and *Pb_FC* in all experiments except for *Pb_FC* on IEMOCAP-scripted. This indicates that there is complementary information from representations learned separately from the audio and phoneme modalities.

The advantage of *Ab+Pb_FC* over *APb_FC* decreases with the flexibility of the lexical content. *Ab+Pb_FC* significantly outperforms *APb_FC* on IEMOCAP-scripted and IEMOCAP-all, but is only comparable to *APb_FC* on IEMOCAP-improv and MSP-I+N. This presents additional evidence that the memorization of patterns in phoneme sequences is most beneficial when the elicitation relies upon scripts. This suggests that there are multiple causes behind the improvements over the unimodal models, via feature fusion and intermediate fusion, and that we may achieve further performance gain by combining them using multi-stage fusion.

4.3. Multi-stage Fusion Results

Our proposed multi-stage fusion model, *APb+Pb_FC*, aims to exploit the dual-functionality of phonemes. It significantly outperforms *APb_FC* in all four experiments. *APb+Pb_FC* also shows consistent performance improvement over *Ab+Pb_FC*, and the advantage is larger on data with less control over the lexical content (i.e., IEMOCAP-improv and MSP-I+N). This result supports our hypothesis that the consideration of both the phonetic modulation of acoustics and the connection between phoneme sequences and emotions allows us to improve the performance of valence prediction.

We investigate the performance of another multi-stage fusion model, *APb+Ab_FC*, which merges the outputs of the feature fusion branch and the unimodal acoustic branch. We find that *APb+Ab_FC* is comparable to *APb_FC* in all experiments, and significantly outperformed by *Ab+Pb_FC* on IEMOCAP-all and IEMOCAP-scripted. The fact that repeatedly adding the acoustic modality does not improve performance is in line with our hypothesis that the learned representation from fused acoustic and phonetic features is dominated by the audio modality.

We compare our best UAR with the state-of-the-art result using the same label processing, training-validation-testing folds, and evaluation method [19]. We find that *APb+Pb_FC* outperforms the intermediate-fusion of the acoustic and lexical modalities using outer-product in [19] by 4.4% in UAR on IEMOCAP-all. This further demonstrates the effectiveness of our method. We note, however, that we cannot attribute the performance gain completely to the use of phoneme sequences and multi-stage fusion. The differences in network structure (e.g., Conv-Pool vs. GRU, dropout, activation function), hyper-parameters (e.g., layer size, kernel width), optimizer, and training paradigm all have important influence on the final results.

5. CONCLUSIONS

In this paper, we explore the impact of incorporating phonetic knowledge into acoustic valence recognition. We propose to repeatedly add phonetic features, at both feature-level and utterance-level, into a single temporal convolutional neural network. We show that this multi-stage fusion model outperforms all other models on IEMOCAP-all, IEMOCAP-improv, and MSP-I+N, even when the transcriptions are estimated using ASR systems (i.e., MSP). The gain over the most accurate single-stage fusion network is the greatest given improvised and natural interactions. This demonstrates efficacy of this approach given imperfect transcriptions and speech data that are collected without reliance upon a script. Finally, the proposed system outperforms the state-of-the-art approach from the literature.

Our results also show that the phonetic branch helps the network leverage the direct link between emotion and speech content contained in phoneme sequences. Feature fusion can capture the phonetic modulation of acoustics, but the resulting representation is dominated by the acoustic modality. The advantage of intermediate fusion over feature fusion decreases when the lexical content becomes more spontaneous. These findings support our assumption that feature fusion and intermediate fusion exploit acoustic and lexical properties of phonemes, respectively. Future work will explore the feasibility of performing integrated phone recognition coupled with emotion recognition.

Acknowledgements

This material is based in part upon work supported by the Michigan Institute for Data Science ("MIDAS"), by Toyota Research Institute ("TRI"), and by the National Science Foundation (NSF CAREER 1651740). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the MIDAS, NSF, TRI, or any other Toyota entity. We thank Prof. Carlos Busso for sharing the ASR transcripts of the MSP-Improv dataset with us.

References

[1] JA Russel, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–78, 1980.

[2] James A Russell, "Core affect and the psychological construction of emotion," *Psychological review*, vol. 110, no. 1, pp. 145, 2003.

[3] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Interspeech*, 2008, pp. 597–600.

[4] Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan, "Emotion recognition based on phoneme classes," in *ICASSP*, 2004.

[5] Carlos Busso, Sungbok Lee, and Shrikanth S Narayanan, "Using neutral speech models for emotional speech analysis," in *Interspeech*, 2007.

[6] Bogdan Vlasenko, Dmytro Prylipko, Ronald Böck, and Andreas Wendemuth, "Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications," *Computer Speech & Language*, vol. 28, no. 2, pp. 483–500, 2014.

[7] Kyung Hak Hyun, Eun Ho Kim, and Yoon Keun Kwak, "Emotional feature extraction based on phoneme information for speech emotion recognition," in *RO-MAN*, 2007, pp. 802–806.

[8] Dmitri Bitouk, Ragini Verma, and Ani Nenkova, "Class-level spectral features for emotion recognition," *Speech communication*, vol. 52, no. 7-8, pp. 613–625, 2010.

[9] Bogdan Vlasenko, David Philippou-Hübner, Dmytro Prylipko, Ronald Böck, Ingo Siegert, and Andreas Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal emotions," in *ICME*, 2011, pp. 1–6.

[10] Saurav Sahay, Shachi H Kumar, Rui Xia, Jonathan Huang, and Lama Nachman, "Multimodal relational tensor network for sentiment and emotion classification," *arXiv preprint arXiv:1806.02923*, 2018.

[11] Zhaocheng Huang and Julien Epps, "An investigation of partition-based and phonetically-aware acoustic features for continuous emotion prediction from speech," *IEEE Transactions on Affective Computing*, 2018.

[12] Wenjing Han, Huabin Ruan, Xiaomin Chen, Zhixiang Wang, Haifeng Li, and Björn Schuller, "Towards temporal modelling of categorical speech emotion recognition," *Interspeech*, pp. 932–936, 2018.

[13] Kalani Wataraka Gamage, Vidhyasaharan Sethu, and Eliathamby Ambikairajah, "Salience based lexical features for emotion recognition," in *ICASSP*, 2017, pp. 5830–5834.

[14] Kalani Wataraka Gamage, Vidhyasaharan Sethu, and Eliathamby Ambikairajah, "Modeling variable length phoneme sequences - a step towards linguistic information for speech emotion recognition in wider world," in *ACII*, 2017, pp. 518–523.

[15] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," *Interspeech*, pp. 3688–3692, 2018.

[16] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[17] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[18] Jonathan Chang and Stefan Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *ICASSP*, 2017, pp. 2746–2750.

[19] Zakaria Aldeneh, Soheil Khorram, Dimitrios Dimitriadis, and Emily Mower Provost, "Pooling acoustic and lexical features for the prediction of valence," in *ICMI*, 2017, pp. 68–72.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[21] Zakaria Aldeneh and Emily Mower Provost, "Using regional saliency for speech emotion recognition," in *ICASSP*, 2017, pp. 2741–2745.

[22] Biqiao Zhang, Georg Essl, and Emily Mower Provost, "Predicting the distribution of emotion perception: capturing inter-rater variability," in *ICMI*, 2017, pp. 51–59.

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learing Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] Andrew Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," in *Interspeech*, 2012.