INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Temporal Context in Speech Emotion Recognition

Yangyang Xia[1*], *Li-Wei Chen*[2*], *Alexander Rudnicky*[2], *Richard M. Stern*[1,2]

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, USA
[2]Language Technologies Institute, Carnegie Mellon University, USA
raymondxia@cmu.edu, {liweiche, air, rms}@cs.cmu.edu

Abstract

We investigate the importance of temporal context for speech emotion recognition (SER). Two SER systems trained on traditional and learned features, respectively, are developed to predict categorical labels of emotion. For traditional acoustical features, we study the combination of filterbank features and prosodic features and the impact on SER when the temporal context of these features is expanded by learnable spectro-temporal receptive fields (STRFs). Experiments show that the system trained on learnable STRFs outperforms other reported systems evaluated with a similar setup. We also demonstrate that the wav2vec features, pretrained with long temporal context, are superior to traditional features. We then introduce a novel segment-based learning objective to constrain our classifier to extract local emotion features from the large temporal context. Combined with the learning objective and fine-tuning strategy, our top-line system using wav2vec features reaches state-of-the-art performance on the IEMOCAP dataset.

Index Terms: speech emotion recognition, deep neural networks, prosodic features, wav2vec, learnable spectro-temporal receptive fields

1. Introduction

Automatic speech emotion recognition (SER) has received increasing attention in recent years. Studies have shown that emotion recognition can be achieved using frame-level acoustical features as well as para-linguistics features calculated from a much larger time span [1, 2]. In this paper we focus on creating an accurate emotion classifier that uses a combination of relevant acoustical cues and the right time span over which to integrate extracted features.

Various knowledge-based and learned features have been explored for SER. Many of these methods have based their approach on short-time spectral features (*e.g.,* [3], [4], among others). Medium-time modulation features [5] and prosodic features [6, 7] have also been explored. In addition to individual features, a combination of features of various form and resolution was found to be helpful for predicting valence and arousal [8] or for categorical emotions [4].

Learned features trained on large-scale corpora in an unsupervised manner provide an attractive alternative to those mentioned above. Wav2vec [9] is one example of such learned features, as it computes features that span a long temporal context window from a raw waveform. Despite its success in automatic speech recognition (ASR) and speaker identification tasks, only a few studies [10, 11] have explored the use of this feature for SER. Boigne et al. [11] show that wav2vec performs better than low-level descriptors and spectrogram on SER. Macary et al. [10] find that the joint use of wav2vec and BERT-like [12] features outperforms short-time cepstral features with word2vec [13] on a continuous SER task. These studies show that representations learned from wav2vec can be beneficial for SER.

In addition to using features extracted over different time spans, the use of context-aware classifiers has been a popular approach for SER. While many studies rely on DNN architectures such as long short-term memory (LSTM) for implicitly learning from long temporal context, making such procedure explicit with clearer objectives has shown to be beneficial. Aldeneh et al. [14] explicitly models the context span (called "regional saliency" by the authors) of their system by varying the filter size of a convolutional neural network (CNN). In contrast, some works [3, 15, 16] found it beneficial to perform frame-level classification and use a simple averaging [3] or training another classifier on the top of extracted frame-level features [15, 16] to produce the utterance-level prediction. While there is not yet a dominant theory of the optimal way to aggregate frame-level decisions for utterance-level decisions, these methods show that both approaches are useful for SER.

In this paper, we examine several SER features and the effect of expanding the temporal context of these features. We develop two systems based on acoustical features and learned features, respectively. For acoustical features, we select three representative features that span different levels of acoustical context. The modulation feature is automatically extracted using a constrained CNN that we have shown can discriminate voice quality [17]. For learned features, we focused on wav2vec [9], which spans a much longer temporal context than traditional features. However, long temporal context tends to include excessive information and makes the classifier vulnerable to spurious correlations. Motivated by existing segment-based classifiers for SER [3, 15, 16], we design a novel training objective that constrains the learned representation to predict emotion from each learned feature frame. Combining this objective with fine-tuning strategies for wav2vec, our system outperforms previous methods on the IEMOCAP dataset.

Organization of this paper. In the next section, we introduce our method for speech emotion recognition. We then describe the evaluation procedure. Finally, we discuss our experimental results.

2. Method

We explore four frame-level features in which each frame is extracted from speech with significantly different contexts. These context spans are shown in Table 1. We then discuss the implementation of each, as well as the DNN classifier for SER.

2.1. Features

Filterbank feature and augmentation. We use the log-mel power spectrogram as the baseline filterbank feature for SER. Given the short-time Fourier transform (STFT) of a speech signal, the mel weighting is a frequency integration function that

*Equal contribution

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-1840

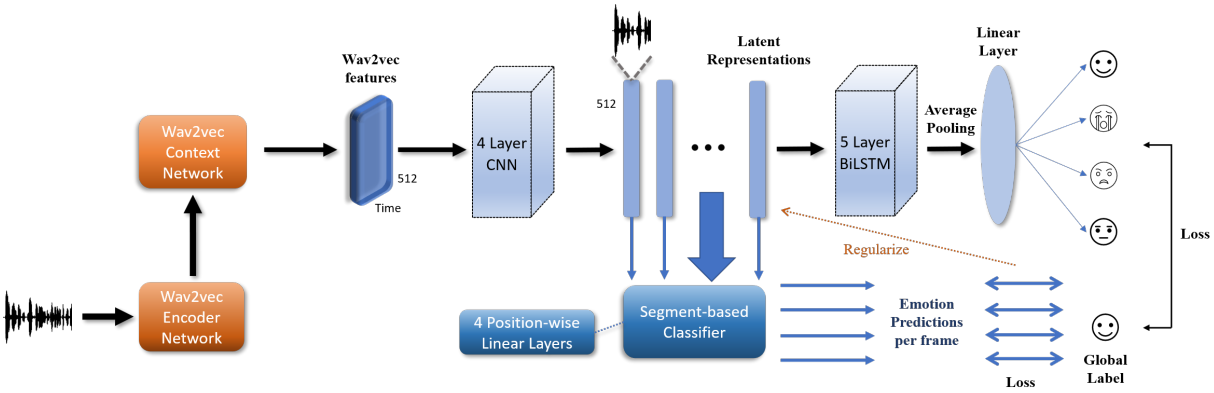


Figure 1: *Illustration of our best performing SER system. The CNN block consists of four layers of one-dimensional CNN with 512 filters. The length and stride size of each filter for each layer are (4, 4, 3, 3) and (2, 2, 1, 1), respectively. The BiLSTM block consists of five layers of bidirectional LSTM with 256 dimensions in the hidden layer. A global average pooling is then taken across the time dimension. The position-wise linear layers have 512 dimensions in the hidden layer. The segment-based classifier is not included in the baseline system.*

transforms the linear frequency scale to the mel scale which approximates the ear's critical bandwidth [18]. We found through experiments that 30 mel bands are sufficient for SER, while having higher frequency resolution produced overfitting in the classifier.

We apply two types of online data augmentation to the spectrogram feature during training: SpecAugment [19] and frequency warping. While the application of SpecAugment [19] is motivated by the observation that no single energy bin should determine the emotion of an utterance, frequency warping is used to normalize the effects of age and gender. Due to resource limitations, we approximate frequency warping through the transformation of the mel weightings [20] instead of transforming the raw time-domain signal [21]. The warping function is based on the bilinear transformation [22] as follows:

$$\omega_t = \omega + 2\arctan\left[\frac{\alpha\sin(\omega)}{1-\alpha\cos(\omega)}\right] \quad (1)$$

where ω and ω_t are the original and target center frequencies of a mel band, respectively, and $-1 < \alpha < 1$ is a constant.

Pitch pulse. Prosodic features based on voice quality measurements have been explored for SER [7, 23], but the temporal evolution of the features was not taken into account. Features such as global pitch might be biased by the speaker gender, while the extraction method of jitter and shimmer could be sensitive to environmental degradation [24]. For these reasons, we use the pitch pulse as the prosodic feature in our experiments. Specifically, pitch pulse locations N_k of each utterance are extracted using *Praat* [25], which uses a global path finder for robust estimation of pitch [26]. Instantaneous amplitude of a speech waveform $x[n]$ is then taken for each pulse, resulting in a train of pulses whose heights are the same as the heights in the original waveform:

$$\text{PitchPulse}[n] = \sum_k x[N_k]\delta[n - N_k]. \quad (2)$$

Note that the pitch pulse preserves all the information needed to extract jitter, shimmer, and pitch, and removes spectral information that relate to phonetic information. Finally, we take the log-mel spectrogram of the pitch pulses, as mentioned previously.

Automatic extraction of modulation patterns. Spectro-temporal modulation patterns have been shown to effectively discriminate between speech and nonspeech [27] and more recently speech with different qualities [17]. We believe that emotive speech is correlated with variations in speaking rates, which can be extracted from temporal modulation [28, 29], and pitch, whose spectral modulation pattern may be important. These observations motivate us to apply the front end of the STRFNet [17] for automatic learning of the modulation filterbank that directly optimizes SER. Using the filterbank feature, we found that 200 ms is a good time support for the learned STRFs. In our experiments, we interpret the extracted modulation from these STRFs as either a feature by itself, or an auxiliary feature that expands the temporal (and spectral) context of the two previously described short-time features.

Learned features. Wav2vec [9] is a model trained to extract meaningful representations out of speech. It consists of two multilayer CNN modules, an encoder network and a context network. The encoder network extracts low-level representations directly from a raw waveform. Based on this representation, the context network is trained to predict future representations from past ones using a contrastive loss. The output layer of the context network is the learned high-level representation. While not being directly interpretable, this learned feature has been shown to be helpful for ASR and for speaker identification [9]. This indicates the learned representation is not only able to capture phonetic information, but also retains para-linguistic information, a key feature for SER.

2.2. Neural network classifiers and learning objectives

We next describe two baseline classifiers for SER. We then describe a learning objective designed to support the long context window of wav2vec features.

Baseline systems. Our baseline system for the engineered features is a combination of CNNs, recurrent neural networks (RNNs), and self-attention as described in [17]. Briefly, the CNN is used to extract 2-D patterns in the time-frequency domain. The single-layer gated recurrent unit and self-attention serve to extract long-term temporal patterns and reduce the time dimension, respectively. Finally, a mutli-layer perceptron (MLP) is used to produce the posterior probability of each emotion category. The reader is directed to [17] for the details of

the architecture. The training objective for the baseline system is the cross-entropy between the predicted posterior probability distribution and the true distribution of emotion categories.

Our baseline system for learned features, shown in Figure 1, is largely similar to the system for the engineered features. Due to the difference in dimensionality between the engineered features and the learned features, the size of the classifiers was scaled accordingly. We have confirmed empirically that the difference in performance is not due to difference in classifier.

Segment-based classification objective. One persistent issue with training a deep neural classifier is that the model may be coding spurious correlations [30]. Given the large temporal context of wav2vec, it is likely that our classifier encodes unwanted relationship that is unrelated to emotion. To address this, we considered an additional objective to regularize the learned latent representations. Specifically, given a length-T sequence of latent representations learned from the main classifier with parameter Θ: $[\mathbf{x_1^\theta}, \mathbf{x_2^\theta}, \cdots, \mathbf{x_T^\theta}]$, we add an additional objective $\mathcal{L}_s$ trained jointly with our main objective:

$$\mathcal{L}_s = \frac{1}{T}\sum_{t=1}^{T} \ell(f_s^\phi(\mathbf{x_t^\theta}), y), \tag{3}$$

where ℓ is the cross-entropy loss function. f_s^ϕ, the segment-based classifier shown in Figure 1, is an independent DNN that predicts the utterance-level emotion y from a single feature frame $\mathbf{x_t^\theta}$. Studies [3, 15] have shown that classifiers trained in this fashion are still able to learn emotion transitions across time. By optimizing this objective, each $\mathbf{x_t^\theta}$ is trained to contain information about local emotion rather than other information from the large temporal context of wav2vec. To enhance this objective further, we also experiment with multiple segment-based classifiers.

3. Experimental Setup

3.1. Dataset

We use the IEMOCAP dataset [31] for all our evaluations. IEMOCAP has five sessions of recordings, in which each session contains recordings from one male speaker and one female speaker. We follow majority of the research community and use four emotion classes: neutral, happy, sad, and angry, in which the excited emotion is merged with the happy emotion. We use the default labels provided by the IEMOCAP, which only retains the data if there is a unique majority agreement across the annotators.

3.2. Training procedure

All systems mentioned in the previous section are trained in the following way. For each iteration, a mini-batch of 64 3-second speech signals is selected. They were scaled by a random factor that spans a 40 dB dynamic range before passed to the feature extraction module. For experiments involving data augmentation using frequency warping, an alpha value is uniformly drawn from $[-0.15, 0.15]$ for each utterance. To mitigate the issue of unbalanced number of data between emotion classes, we weighted the loss of each class by a constant that is inversely proportional to the population of that class. We used the Adam optimizer [32] with an initial learning rate of 1×10^{-4} for updating all learnable parameters. For the wav2vec model, we use the pretrained model developed by Facebook AI[1], which is pretrained on 960 hours of English speech from Librispeech [33].

[1]https://github.com/pytorch/fairseq/tree/master/examples/wav2vec

3.3. Evaluation procedure

We evaluate our performance in terms of weighted accuracy (WA) and unweighted accuracy (UA) [6], two common metrics for SER. We performed 8-fold cross validation as used by [3], using Session 5 for validation and testing on each of the rest speakers at a time. We noticed that the details of cross validation are not well-described in some studies, and some use session-based cross validation for evaluation. We still compare our results with prior works at Table 3. However, one should bear in mind that the CV procedure may be different when comparing performance across studies.

4. Experimental Results and Discussions

4.1. Comparison of features using the baseline model

Our results for using different features on the baseline model are presented in Table 1. We first give an analysis on the results of the engineered features. Recall that the learnable STRFs are used to either extract the modulation feature of a signal or extend the temporal context of other short-time features. We found that the removal of linguistic information resulted in a heavy drop in performance for the pitch pulse feature. The inclusion of modulation features did not improve on the performance of that of the spectrogram feature. However, when paired with online frequency warping, the combination of spectrogram and modulation features resulted in the best performance of 63.2% unweighted accuracy among engineered features. We believe that frequency warping has added variations to the spectro-temporal patterns that were only picked up by the constrained convolutional layer of STRFNet.

Table 1: *SER accuracies by features. Systems are trained without or with online frequency warping (in parentheses).*

Feature	Context span [ms]	UA (%)	WA(%)
Pitch pulse	75	46.2 (47.5)	43.7 (45.5)
Spectrogram	75	61.7 (61.8)	60.7 (59.6)
Modulation	200	60.3 (59.2)	57.4 (56.7)
P. p. + Spec.	75-75	62.2 (62.0)	59.9 (59.2)
P. p. + Mod. [2]	75-200	46.8 (47.1)	44.7 (45.2)
Spec. + Mod.[3]	75-200	61.4 (**63.2**)	59.5 (**61.0**)
All three	75-75-200	61.1 (61.8)	58.7 (59.4)
Wav2vec	810	**64.1** (-)	**62.0** (-)

Context span and classification performance. We observed an interesting trend in that performance actually scales with context span. Nevertheless, it is not generally true that long-context-span features are better for SER. The features should be informative enough to provide emotion details for that given context. This explains why the modulation feature alone provides worse performance than the spectrogram features. While the context span is longer, it loses fine-grained spectral detail. But when it is combined with the original spectrogram and equipped with frequency warping it provides additional information outside the context window of the spectrogram. This also indicates that simply increasing the analysis

[2]Referring to a combination of the pitch-pulse spectrogram and its modulation features

[3]Referring to a combination of spectrogram and its modulation features

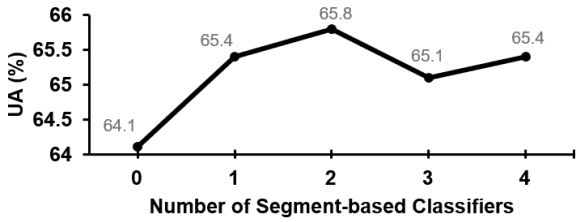


Figure 2: *Number of segment-based classifiers N_s v.s. the performance on IEMOCAP. The experiments were performed without fine-tuning the wav2vec model.*

window size of spectral features may not result in better performance because of reduced time resolution.

Wav2vec representation on SER. As mentioned in Section 2.1, wav2vec features provide learned fine-grained information from a large unsupervised speech corpus. Equipped with a large context span, it is not surprising that it outperforms all the other features. However, we actually found that the performance of wav2vec representations could be improved further by adding the segment-based classification objective mentioned in Section 2.2. This indicates the classifier needs to be regularized in order to capture good features from wav2vec representation. We analyze the experiment results in the following section.

4.2. The wav2vec classifier

In this section we provide empirical evidence that segment-based classification and a good fine-tuning strategy noticeably boosts performance. Based on these observations, we believe that while wav2vec is more informative than other features, the classifier may also capture noisy information not correlated with emotion, resulting in sub-optimal performance.

Effect of segment-based classification objective. In this experiment, we explore how additional segment-based classifiers affect performance. Here we denote the number of segment-based classifiers as N_s, and $N_s = 0$ corresponds to the baseline model without regularization. From Figure 2 we can see that the baseline model performs noticeably worse than the ones with additional classifiers. This implies the original $\mathbf{x}_\mathbf{t}^\theta$ we learned without the segment-based classification objective may not be informative of the emotion states at time t. Otherwise the additional objective would be trivial and would not impact the overall performance. With the additional objective, we first learn a sequence of emotion states, and the goal of the BiLSTM layers will be simply predicting the utterance label on top of that. Our experiment result indicates that this strategy is effective for SER. It also shows that additional care needs to be taken when dealing with learned features despite having a large temporal context. When exploring the number of additional classifier, we found the optimal performance using two classifiers as shown in Figure 2. Therefore we use $N_s = 2$ for our top-line model.

Fine-tuning strategy for wav2vec. While fine-tuning is considered an essential part when using pretrained embeddings [12, 34], previous works using wav2vec on SER [10, 11] uses the embeddings without fine-tuning. Here we inspect how fine-tuning different parts of the wav2vec model changes the performance. Table 2 shows that fine-tuning only the context network outperforms the other strategies by a large margin. Fine-tuning the whole wav2vec or only the encoder network results in almost the same performance as not doing fine-tuning. This is actually not surprising since fine-tuning a pretrained model directly could be brittle [12, 34, 35], especially if the model is fine-tuned on a small dataset. Our experiment result suggests that the pretrained encoder network may have successfully captured low-level representations that contains information about the emotion states. However, by fine-tuning the context network, we can extract better high-level features that focus on emotion building upon the given low-level features.

Table 2: *Performance comparison between 4 fine-tuning strategies of wav2vec. Experiments were done on $N_s = 2$.*

Model	UA (%)	WA (%)
Fine-tune All	65.8	64.1
Fine-tune Context Network	**66.9**	**65.4**
Fine-tune Encoder Network	65.5	63.9
No Fine-tune	65.8	63.8

4.3. Comparison with prior works

Our top-line result is achieved by using $N_s = 2$ and fine-tuning only on the context network of wav2vec. Table 3 presents a comparison with popular methods reported in this literature. Our baseline model on spectral features has already achieved a competitive result. By changing the spectral feature to wav2vec, we gain an 1% absolute improvement, similar to [11], who also use wav2vec. By adding the segment-based classifcation objective (+Seg) and also fine-tuning the context network (+Seq+FineTune), we are able to outperform other methods.

Table 3: *Comparison of methods on IEMOCAP*

Model	Feature	UA (%)
Regional Saliency [14]	Mel Spectrogram	61.8
Frame-based CNN [3]	Spectrogram	58.3
FCN+Attention [36]	Spectrogram	63.9
MLP with mean pooling [11]	Wav2vec	64.3
Baseline (*ours*)	Spec. + Mod.	63.2
Baseline (*ours*)	Wav2vec	64.1
+Seg (*ours*)	Wav2vec	65.4
+Seg+FineTune (*ours*)	Wav2vec	**66.9**

5. Conclusion

In this paper, we investigate multiple acoustical and learned features for the speech emotion recognition task. We found that features extracted with fine resolution and larger context window improve the accuracy of emotion classification. For acoustical features, we found that the combination of short-time spectral features and medium-time modulation features was able to improve performance relative to features used separately. We show that the wav2vec feature is superior to traditional features, and that a good regularization objective function and fine-tuning strategy help boosting the performance even further. Based on this work, we intend to continue exploring the combination of acoustical and learned features with contextual constraints. We are also interested in the role of these features for emotion recognition in more natural conversational settings.

6. Acknowledgements

We are grateful to PwC USA for funding this research through The Digital Transformation and Innovation Center at Carnegie Mellon University sponsored by PwC.

7. References

[1] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[3] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[4] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence LSTM architecture," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6474–6478.

[5] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.

[6] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[7] A. Jacob, "Speech emotion recognition based on minimal voice quality features," in *2016 International Conference on Communication and Signal Processing (ICCSP)*, 2016, pp. 0886–0890.

[8] Z. Yang and J. Hirschberg, "Predicting Arousal and Valence from Waveforms and Spectrograms Using Deep Neural Networks," in *Proc. Interspeech 2018*, 2018, pp. 3092–3096.

[9] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition." in *INTERSPEECH*, 2019, pp. 3465–3469.

[10] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 373–380.

[11] J. Boigne, B. Liyanage, and T. Östrem, "Recognizing more emotions with less data using self-supervised transfer learning," 2020.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR*, 2013.

[14] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 2741–2745.

[15] S. Mao, P. Ching, C.-C. J. Kuo, and T. Lee, "Advancing Multiple Instance Learning with Attention Modeling for Categorical Speech Emotion Recognition," in *Proc. Interspeech 2020*, 2020, pp. 2357–2361.

[16] S. Mao, P. Ching, and T. Lee, "EigenEmo: Spectral Utterance Representation Using Dynamic Mode Decomposition for Speech Emotion Classification," in *Proc. Interspeech 2020*, 2020, pp. 2352–2356.

[17] T. Vuong, Y. Xia, and R. M. Stern, "Learnable Spectro-Temporal Receptive Fields for Robust Voice Type Discrimination," in *Proc. Interspeech 2020*, 2020, pp. 1957–1961.

[18] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[19] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[20] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1. IEEE, 1996, pp. 353–356.

[21] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," *Proceedings of the IEEE*, vol. 60, no. 6, pp. 681–691, 1972.

[22] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*. Springer Science & Business Media, 2012, vol. 201.

[23] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, "Automatic emotion recognition using prosodic parameters," in *Ninth European conference on speech communication and technology*, 2005.

[24] P. Boersma, "Should jitter be measured by peak picking or by waveform matching?" *Folia Phoniatrica et Logopaedica: Official Organ of the International Association of Logopedics and Phoniatrics (IALP)*, vol. 61, no. 5, pp. 305–308, 2009.

[25] P. Boersma and V. Van Heuven, "Speak and unspeak with PRAAT," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[26] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193. Citeseer, 1993, pp. 97–110.

[27] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[28] M. Elhilali, *Modulation Representations for Speech and Music*. Cham: Springer International Publishing, 2019, pp. 335–359.

[29] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech communication*, vol. 25, no. 1-3, pp. 117–132, 1998.

[30] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *CoRR*, vol. abs/1907.02893, 2019. [Online]. Available: http://arxiv.org/abs/1907.02893

[31] 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. 335–359, 2008.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[34] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. A. Smith, "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping," *CoRR*, vol. abs/2002.06305, 2020. [Online]. Available: https://arxiv.org/abs/2002.06305

[35] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. J. Liu, "FreeLB: Enhanced Adversarial Training for Natural Language Understanding," in *Eighth International Conference on Learning Representations (ICLR)*, April 2020.

[36] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1771–1775.