Interspeech 2022
18-22 September 2022, Incheon, Korea

Mind the gap: On the value of silence representations to lexical-based speech emotion recognition

Matthew Perez[1], *Mimansa Jaiswal*[1], *Minxue Niu*[1], *Cristina Gorrostieta*[2], *Matthew Roddy*[2], *Kye Taylor*[2], *Reza Lotfian*[2], *John Kane*[2], *Emily Mower Provost*[1]

[1]Computer Science and Engineering, University of Michigan, Ann Arbor, Michigan, USA
[2]Cogito Corporation, Boston, Massachusetts, USA

{mkperez,mimansa,sandymn,emilykmp}@umich.edu,
{cgorrostieta,mroddy,ktaylor,rlotfian,jkane}@cogitocorp.com

Abstract

Speech timing and non-speech regions (here referred to as "silence"), often play a critical role in the perception of spoken language. Silence represents an important paralinguistic component in communication. For example, some of its functions include conveying emphasis, dramatization, or even sarcasm. In speech emotion recognition (SER), there has been relatively little work on investigating the utility of silence and no work regarding the effect of silence on linguistics. In this work, we present a novel framework which investigates fusing linguistic and silence representations for emotion recognition in naturalistic speech using the MSP-Podcast dataset. We investigate two methods to represent silence in SER models; the first approach uses utterance-level statistics, while the second learns a silence token embedding within a transformer language model. Our results show that modeling silence does improve SER performance and that modeling silence as a token in a transformer language model significantly improves performance on MSP-Podcast achieving a concordance correlation coefficient of .191 and .453 for activation and valence respectively. In addition, we perform analyses on the attention of silence and find that silence emphasizes the attention of its surrounding words.

1. Introduction

Recognizing emotion is critical to understanding human communication, wellness, and how users interact with technology. If properly captured, human emotion could lead to more intuitive interfaces for real-world systems and better assistive technology for improving mental health/well-being. Consequently, automatic speech emotion recognition (SER) represents an important area of focus with many downstream applications. However, SER is a complex and challenging task even for humans [1, 2]. Further, these challenges are compounded in scenarios where text is the only available modality, due to privacy considerations and the sensitive nature of acoustics. In these conditions SER estimates must be made from text-features alone. In this paper, we investigate how text-focused SER models can be enhanced by considering low-cost paralinguistic features, like unvoiced segments in transcribed speech.

Privacy is one of the central challenges in deploying SER systems. These concerns arise from the use and sharing of audio data, which is generally considered to be a biomarker containing personally identifiable information (PII). SER systems can address these privacy issues by removing the acoustic information and retaining only language, accessible via automatic speech recognition (ASR). However, this solution is often at the expense of losing critical paralinguistic information. One paralinguistic aspect that can easily be preserved from ASR is silence, which has an important role in non-verbal communication [3,4]. Because silence can provide additional context about an individual's emotion state, we explore how silences embedded within continuous speech can be used to augment text-focused SER systems. Further, silence can be extracted simply and inexpensively given the timing information available from ASR output. Despite this fact, there has been relatively little work on the role of silence in speech emotion recognition, with existing work focusing on acoustic features only [5, 6].

SER literature has demonstrated that language information is particularly informative for predicting valence in emotion [7–11]. Recently, the research community has had success modeling language using transformers, which are models that perform well on linguistic-related tasks due to their ability to learn relationships between different portions of a lexical sequence via self-attention [12–14]. One popular model in particular is BERT, which is often used as an upstream feature extraction tool to capture language representations from text. These BERT embeddings are then combined with other modalities using late-stage fusion for SER. Although many recent studies involving transformer architectures have investigated the fusion of paralinguistics for downstream behavioral prediction tasks [15,16], there has been little work around augmenting these transformer language models with paralinguistic information (i.e. specialized tokens) for speech emotion recognition.

We present one of the first works that investigates approaches to combine silence information with text-based features for downstream SER and provide analyses on how silence impacts learned semantic representations from BERT. We investigate two methods for incorporating silence information. The first uses global silence statistics computed over the utterance along with late-stage fusion with word-level text representations. This method parallels one used in acoustic SER modeling [5]. The second approach involves modeling silence in a transformer-based language model by learning a new token which represents silence. Additionally, we perform analyses on the attention layer of the BERT model in order to interpret how the inclusion of silence affects the prediction of emotion. Our results show performance improvements in activation and valence over a baseline BERT model that does not take silence into account, with significant improvements shown for activation prediction.

2. Related Work

Investigating silence for SER has been a relatively understudied area of research. Notable exceptions include the works by Atmaja and colleagues [5, 6]. They have presented two directions for addressing silence in SER. The first is to omit silence and exclusively focus on portions of voiced speech [6] and the

10.21437/Interspeech.2022-10943

other involves capturing an utterance's silence in a feature that is fused with other extracted features [5].

In the first method presented by Atmaja et. al., the authors filter out silence using voice activity detection (VAD) as well as an attention mechanism in the SER model [6]. The authors show experimental results on The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database and demonstrate that removing silences yielded a 1% absolute improvement and that segmentation with attention yielded a 11% absolute improvement in terms of accuracy on 4-class classification.

Atmaja et. al. build on this work and present an additional approach that highlights silences rather than ignoring them. The authors show that silence features and the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) improve SER performance on IEMOCAP [5]. The authors extract a global silence feature over the utterance by computing the portion of silence frames in a given utterance. This silence feature is then concatenated to the acoustic features, which are the mean and standard deviation of the eGeMAPS feature set. The inclusion of this silence feature results in improved performance for activation and valence regression. They achieve a Concordance Correlation Coefficient (CCC) of 0.561 and 0.214 for activation and valence respectively. The authors also show that silence affects the arousal dimension more than other emotional dimensions. Our work is different from the works of Atmaja et. al. in that we specifically study the effect of silence within the context of language modeling and linguistic features. Instead of IEMOCAP, we use the MSP-Podcast dataset, which contains more naturalistic speech data and doesn't have the same level of lexical overlap seen in the multiple sessions of the IEMOCAP dataset.

Although there's extensive SER literature on using MSP-Podcast, relatively few works that have used linguistic features [17, 18]. One of the barriers to a language-focused approach with MSP-Podcast is the lack of a ground truth transcription. However, a widely used approach is to apply off-the-shelf automatic speech recognition systems to obtain the text content. Pepino et. al. show the performance of categorical emotion classification using language features taken from both GloVe and BERT on MSP-Podcast (version 1.6). The authors transcribe utterances using Google Cloud Speech-to-Text and filter for Angry, Happy, Sad, and Neutral utterances to perform 4-class emotion classification. The authors' primary focus was to investigate a multimodal SER framework and ultimately achieve an unweighted average recall (UAR) score of 0.59 [17].

Similarly, Srinivasan et. al. focused on multimodal SER for MSP-Podcast (version 1.6) by combining BERT representations with acoustic representations from ASR. The authors use an internal ASR system for generating transcriptions as well as extracting speech representations. They perform multitask learning on dimensional emotion labels and achieve CCC of 0.757 and 0.627 for activation and valence respectively using a multimodal student-teacher network [18].

These works provide some insight into the performance of language models on MSP-Podcast, however, it is important to note the version of the dataset which is used as this has a distinct impact on the performance which is presented. Additionally, the previous works focus on extracting broad acoustic representations to combine with linguistic features. Although acoustic features provide a rich source of information, especially in the context of SER, they also come with drawbacks of privacy, size and cost. Paralinguistic features on the other hand are cheaper, smaller, and have fewer concerns regarding personal privacy. Our work specifically focuses on fusing silence features with linguistic features. Silence can be easily extracted from the output of common ASR systems and silence data is small in comparison to acoustics. Once extracted, we investigate both late-stage fusion of a utterance-level silence feature as well as training a new silence token in the BERT framework to allow for semantic representation learning.

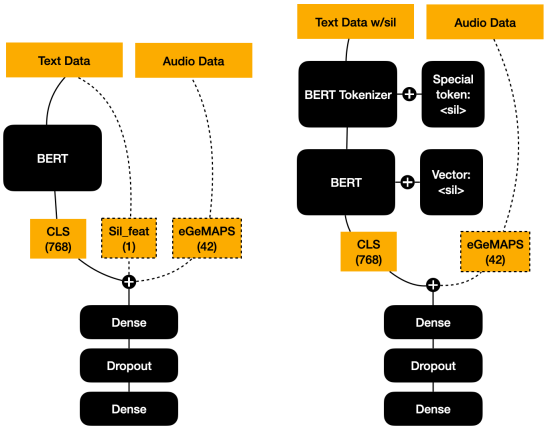

Figure 1: *Model architecture for silence fusion*

3. Dataset

We use the MSP-Podcast corpus (version 1.7), which contains naturalistic speech and is annotated with dimensional and categorical emotion labels [19]. The dataset has roughly 100 hours of speech and is categorized into 38,179 training utterances, 7,539 validation utterances, and 12,902 test utterances. In our work, we use these predefined partitions for training, validation, and testing. Each utterance was rated by multiple annotators and in our work we use the average rating across all annotators. We create normalized continuous activation labels by performing z-normalization over the training data. We filter out utterances that had no annotator agreement on the categorical emotion label (following [17]), leaving 35,952 training, 6,341 validation, and 12,355 test samples. In addition to using the continuous emotion label, we also create binned valence and activation labels by taking the 3-quantiles computed over the training set. This method has been used in other SER works for more robust modeling [9, 11]. We transcribe the MSP-Podcast dataset using the Microsoft Azure speech-to-text API. Output transcriptions contain word timing information which we later use to determine where silences occur.

In addition to MSP-Podcast we also use the Switchboard dataset (SWBD), which contains approximately 260 hours of conversational telephone speech from 543 speakers. We use Switchboard for fine-tuning BERT to the stylistic differences in transcribed speech compared to the online written text it was originally trained on. We selected the Switchboard dataset due to its size (large amounts of naturalitstic speech data, similar to MSP-Podcast) and well-established recipes for training an ASR model. We follow the Kaldi recipe for training a deep neural network (DNN) ASR model [20] and perform forced alignment using the ground truth transcriptions.

4. Method

4.1. Model Architecture

Linguistic representations are obtained by feeding the transcripts through a pretrained BERT language model provided by

HuggingFace [1]. This model consists of 12 layers, 12 attention heads and 110M parameters. The linguistic representations that are output from BERT are then distilled to a CLS token, which is a fixed feature vector of size 768 representing the entire utterance. The CLS token has been used for sentence classification tasks and shown to be effective in SER applications with multimodal late-stage fusion techniques [21].

In all of our experiments we finetune the pretrained BERT model by performing masked language modeling (MLM) using first the Switchboard dataset and then MSP-Podcast. Our goal here is to adapt BERT to the stylistic nuances of spoken text. We perform MLM by randomly masking 15% of the tokens in each utterance and training for a maximum of 30 epochs with an initial learning rate of 5e-5. We perform model checkpointing and early stopping with a patience of 10 by monitoring the loss of the MSP-Podcast validation set.

For our downstream SER model, we replicate the model architecture for SequenceClassification in the HuggingFace API (shown in Figure 1). The CLS output from BERT is passed to a dense layer that has the same input/output shape with ReLU activation. We then apply dropout (p=0.5) before passing representations through a final output layer. We train the model using an Adam optimizer with a learning rate of 1e-5 and reduce the rate by a factor of 0.75 when the validation loss increases between epochs. We refer to this as the baseline BERT model.

We train two separate models: one that predicts a continuous score (activation or valence regression) and the other that predicts a binned activation or valence label (see Section 3). We train the regression model with the Concordance Correlation Coefficient (CCC) metric and the classification model with the Unweighted Average Recall (UAR) metric. We use model checkpointing and early stopping based on the validation set performance. After training has completed, we evaluate the model on the test set using the best model identified over the validation data. This process is repeated using 10 different seeds, which means using the established train/validation/test partitions and repeating the experiment 10 times, comparing the performance of the models over the 10 runs. In both cases, we present our results in terms of the mean and standard deviation across the 10 runs. We perform significance tests using a repeated paired t-test and note significance when $p<0.05$.

4.2. Global Silence Feature

We define silence as unvoiced segments between word boundaries provided by ASR. For both MSP-Podcast and Switchboard we use ASR to identify periods between words and insert a sil token at these locations in the transcription.

We then compute an utterance-level silence feature following the work of Atmaja et. al. [5], which is the portion of silence in the utterance. This feature is then concatenated to the linguistic representations prior to classification. This approach allows us to represent silence information captured over the utterance and investigate a late-stage fusion approach with the CLS token extracted from the BERT language model. See Figure 1 (left), dashed lines show an optional inclusion of the silence or acoustic features (Section 4.4).

4.3. Learned Silence Token

Our second approach of encoding silence information involves learning silence within the BERT language model framework. We implement this by augmenting the tokenizer and embedding matrix to account for the new sil token. The added sil token is first learned through masked language modeling (described earlier). This finetuned BERT model with the learned silence token is then used in our downstream pipeline, allowing us to investigate the effect of encoding silence information into the language model itself with the intention of learning semantic representations between words and silence. To the best of our knowledge, this is the first study to incorporate silence representations with conventional lexical embeddings in this fashion for SER. See Figure 1 (right), dashed lines show an optional inclusion of acoustic features (Section 4.4)

[1] https://huggingface.co/bert-base-uncased

Method	Activation		Valence	
	UAR	CCC	UAR	CCC
BERT	.419±.002	.168±.012	**.527±.003**	.447±.015
BERT+sil_feat	.418±.003	.167±.011	.526±.002	.430±.02
BERT+sil_token	**.427±.003***	**.191±.014***	**.527±.003**	**.453±.014**
Multimodal - eGeMAPS fusion				
BERT	.507±.004	.488±.015	.537±.005	.466±.029
BERT+sil_feat	.507±.004	.490±.009	.538±.004	**.484±.015**
BERT+sil_token	**.510±.003**	**.492±.006**	**.546±.003***	.478±.011

Table 1: *SER results on the MSP-Podcast, including the mean±std across the 10 seeds. * indicates statistical significance as defined by paired t-test against BERT where p<0.05*

4.4. Multimodal

In our work, we also experiment with multimodal SER using both standard SER acoustic features and BERT features. We are interested in investigating the impact of silence when a full suite of emotion-focused acoustic-based features are included. In this study, we use the eGeMAPS feature set, which has been used in a variety of works to capture emotion and investigate the effects of silence on acoustics [5, 6, 22]. Inspired by [5] we perform late-stage fusion with the CLS token and/or silence feature followed by layer normalization (outlined in Figure 1).

5. Results

5.1. Silence Performance

Our results, shown in Table 1, demonstrate that when considering text-only approaches, BERT + silence token significantly outperforms both the baseline (finetuned BERT) and silence feature approaches at modeling activation for the MSP-Podcast dataset. Our results show that BERT with silence features does slightly worse in both activation and valence when comparing against the baseline BERT model. However, a BERT model with a learned silence token performs on par or better in valence and significantly better in activation when compared against a finetuned BERT model without a silence token. The proposed BERT model with the learned silence token achieves an average performance of .427 UAR and .191 CCC for activation and .527 UAR and .453 CCC for valence.

These results demonstrate that silence can provide useful information for SER when modeled properly. The most prominent results are seen when silence is modeled in the language model framework using a silence token. The largest impact seems to be on activation, achieving a relative improvement of 14.6% for CCC, which is expected since paralinguistics typically perform best on activation. The performance of BERT + silence token suggests there are semantic relationships between silence and words, which we investigate further in Section 6.

5.2. Multimodal

As a follow up, we investigate the impact of a silence token in BERT for multimodal SER (shown in Table 1). We look at using utterance based features such as eGeMAPS, which have been useful in a variety of SER applications, to capture meaningful paralinguistic properties from acoustics.

The results show the inclusion of eGeMAPS led to overall improved activation and valence performance across all methods. We specifically note significant improved activation performance, which in the case of BERT + silence token, increased from .191 to .492 CCC. This increase highlights the strong arousal patterns captured by paralinguistics. We also note that despite the inclusion of eGeMAPS, BERT + silence token still performs the best with respect to activation and attains a statistically significant valence improvement for binned classification over the baseline BERT model.

6. Discussion

We investigate the effect of the silence token on regression. We focus on the BERT + silence token model, as described in Section 4.3 in order to better understand when silence improves activation performance and how it affects surrounding word tokens in BERT. We downsample our data to include only samples with at least three word tokens and fewer than 18 silence tokens (removing 86 out of 12,902 samples).

6.1. What are the attribution of the silence token

First, we assess the contribution of silences to the overall prediction for the sample using integrated gradients [23] from the Captum library [24]. Our analysis focuses on model behavior by observing the different relationships between input (i.e. sentence composition, utterance length, etc.) and output (i.e. predicted and ground truth labels). Integrated gradients provides an attribution weight for each token, indicating the token's effect on the final prediction. A large positive value *(green)* indicates that the presence of the silence token pushes the prediction of a sample towards a higher activation, while a large negative value *(blue)* indicates that the presence of the silence token pushes the prediction of a sample towards a lower activation. We define an attribution as "high" when the absolute value of the attribution is above 0.3, which is the 85^{th} percentile of all the attribution values obtained over the test dataset, as in prior work [25].

In 92% of the samples that see a performance improvement, the silence token itself has a high attribution value. For example, the silence token attributions in the following utterance leads to a more accurate activation prediction. *it* SIL was it *was* *SIL* it *was* *SIL* *a* *watershed* moment . However, the attribution of the silence token in the following utterance does not, *maybe* *SIL* that **s *a* lesson *in* *reality* . This supports the notion that the attribution of the silence token is impactful to the overall activation prediction and generally improves performance. Further, we observe that the attribution for the majority of silence tokens (95%) is negative. This means that the presence of a silence token generally improves performance by muting or lowering the predicted activation for a given sample.

We observe that silence (*SIL*) tokens are most impactful when they are near a highly attributed general word token. For example, consider the following sentence: *there* *is* nothing but *emptiness* *SIL* ***(1)*** called i move *SIL* *on* *SIL* ***(2)*** *us*. Here, *SIL* token 1 is followed by another imapctful token ('called') with a high contribution, whereas *SIL* token 2, which precedes a non-impactful word, does not have a significant contribution.

To help measure this relationship, we collect the maximally attributed silence token in a sentence and its distance to the nearest highly attributed token (absolute attribution value higher than 0.3). We then calculate the maximal silence attribution per unit distance and find the correlation between this value and overall improvement in performance, to be 0.54. This suggests that the distance between highly attributed silence and non-silence tokens plays a role in measured activation improvement. This analysis highlights the fact that the location of these silence tokens matters in the context of activation performance.

6.2. When do silence tokens improve prediction?

Next, we study the characteristics of the test samples that are most strongly impacted by the silence tokens. We first start by investigating the label of each utterance. We z-normalize the activation predictions for the test data using the mean and standard deviation of the ground truth labels from the training set. In the samples with labels < -1, the presence of the silence token improves performance 86.5% of the time. In the samples with labels >1, the presence of the silence token improves performance 67.2% of the time. This contrasts with those within the range of $\{-1, 1\}$, where the presence of the silence token improves performance only 48.17% of the time. We find that the most impacted test samples are those which are "emotional outliers", compared to the training data.

Lastly, we find that samples that have a lower ratio of silence tokens to non-silence tokens generally see improved performance. When the ratio of silences is below 0.18, the performance increases in 82% of the samples. When the ratio of silences is higher, the performance improves in only 8% of the samples. We believe that this may occur because the larger number of silence tokens could instead be capturing aspects of speaking style, which may not correspond to any particular emotion. For example, in the following sentence, yep *SIL* good SIL *for* *SIL* you *SIL* don *SIL* henry *SIL* *wherever* you; the ground truth label is 1.13, and the prediction with/without silence is $-0.32/0.44$, i.e., the ratio of silence is 0.4 and all silence tokens with high attribution precede those with lower attribution, leading to an undesirable drop in predicted activation value.

7. Conclusion

In this work, we investigate two methods for incorporating silence information with text features for SER. The first uses an utterance-level silence feature and the second models silence in a BERT model by learning a new token representation. We show that learning a silence token in a BERT model achieves up to a 14.6% relative performance improvement for activation CCC on MSP-Podcast. We perform secondary analyses to understand the relationship between the silence token and the model prediction. We find that silence not only has a strong impact on activation but that the location and frequency are also important characteristics to consider. This work is especially important for SER applications where privacy, space, or cost issues are raised from sharing audio/acoustic data.

8. Acknowledgements

This material is based in part upon work supported by the NSF-GRFP. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding sources listed above.

9. References

[1] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[2] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[3] M. Ephratt, "Linguistic, paralinguistic and extralinguistic speech and silence," *Journal of pragmatics*, vol. 43, no. 9, pp. 2286–2307, 2011.

[4] D. Matsumoto and H.-S. Hwang, "The messages of emotion, action, space, and silence," *The Routledge handbook of language and intercultural communication*, pp. 130–147, 2012.

[5] B. T. Atmaja and M. Akagi, "The effect of silence feature in dimensional speech emotion recognition," *arXiv preprint arXiv:2003.01277*, 2020.

[6] ——, "Speech emotion recognition based on speech segment using lstm with attention model," in *IEEE International Conference on Signals and Systems (ICSigSys)*. IEEE, 2019, pp. 40–44.

[7] B. Vlasenko, R. Prasad, and M. Magimai.-Doss, "Fusion of acoustic and linguistic information using supervised autoencoder for improved emotion recognition," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 2021, pp. 51–59.

[8] Z. Aldeneh, S. Khorram, D. Dimitriadis, and E. M. Provost, "Pooling acoustic and lexical features for the prediction of valence," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 68–72.

[9] M. Jaiswal and E. M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7985–7993.

[10] G. Soganc ıoglu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadée, A. A. Salah, and A. Karpov, "Is everything fine, grandma? acoustic and linguistic modeling for robust elderly speech emotion recognition," 2020.

[11] M. Jaiswal, Z. Aldeneh, and E. Mower Provost, "Controlling for confounders in multimodal emotion classification via adversarial learning," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 174–184.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *arXiv preprint arXiv:2203.07378*, 2022.

[15] J. Biggiogera, G. Boateng, P. Hilpert, M. Vowels, G. Bodenmann, M. Neysari, F. Nussbeck, and T. Kowatsch, "Bert meets liwc: Exploring state-of-the-art language models for predicting communication behavior in couples' conflict interactions," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 385–389.

[16] G. Boateng, P. Hilpert, G. Bodenmann, M. Neysari, and T. Kowatsch, ""you made me feel this way": Investigating partners' influence in predicting emotions in couples' conflict interactions using speech data," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 390–394.

[17] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6484–6488.

[18] S. Srinivasan, Z. Huang, and K. Kirchhoff, "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition," *arXiv preprint arXiv:2112.00158*, 2021.

[19] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE Signal Processing Society*. IEEE Signal Processing Society, 2011.

[21] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition," *arXiv preprint arXiv:2008.06682*, 2020.

[22] B. T. Atmaja and M. Akagi, "Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning," *APSIPA Transactions on Signal and Information Processing*, vol. 9, 2020.

[23] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017.

[24] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan *et al.*, "Captum: A unified and generic model interpretability library for pytorch," *arXiv preprint arXiv:2009.07896*, 2020.

[25] P. Sturmfels, S. Lundberg, and S.-I. Lee, "Visualizing the impact of feature attribution baselines," *Distill*, vol. 5, no. 1, p. e22, 2020.