# Exploring Foundation Model Fusion Effectiveness and Explainability for Stylistic Analysis of Emotional Podcast Data

Arnab Das[1], Carlos Franzreb[1], Tim Polzehl[1], and Sebastian Möller[1,2]

[1] German Research Center for Artificial Intelligence (DFKI), Germany,
{arnab.das, carlos.franzreb, tim.polzehl}@dfki.de
[2] Technische Universität Berlin, Germany,
sebastian.moeller@tu-berlin.de

**Abstract.** Emotion recognition is one of the crucial research fields for advancing affective computing. Automatic prediction using deep learning models shows poor performance while predicting valence/polarity for spoken utterances. In this paper, we investigate the effectiveness of emotion representations from a recent weakly supervised multilingual large automatic speech recognition (ASR) model along with two other self-supervised pre-trained general purpose foundation models for dimensional emotion recognition tasks from speech. We also propose a fusion architecture and demonstrate that the proposed method can achieve significantly better results compared to a state-of-the-art baseline. Moreover, we train our model with additional pairwise rank loss to further improve the prediction reliability. We further attempt to explain the prediction results using post-hoc occlusion methods demonstrating a strong relationship between the contextual construct of language and valence/polarity. Finally, we perform a comprehensive exploration of the data and labels and identify instances of verbal irony causal for individual prediction failure.

**Keywords:** speech emotion recognition, fusion learning, explainability

## 1 Introduction

Emotion has traditionally been interpreted into the concept of *sentiment*, i.e. an attitude or perception held towards a particular object in the NLP community, and the concepts of *arousal* and *valence* in terms of *emotion*, i.e. a more general instantaneous characteristic of a person's feeling [1]. Further, evidence suggests that sentiment annotations can be decomposed into two components: *intensity* and *polarity* in the NLP community [2], which roughly correspond to the concepts of *arousal* and *valence* as defined in the speech community, here further referred to as Speech Emotion Recognition (SER) [1]. Though these dimensions of sentiment and emotion are not fundamentally the same they are certainly co-related and overlapping.

Speech is a fundamental medium of human-to-human communication and we humans have an innate ability to detect a speaker's emotion through the linguistic and paralinguistic cues in speech. That allows us to construct or modify our responses accordingly. The growing prevalence of speech devices, such as voice assistants interacting with users, necessitates automated and precise emotion detection by machines, emphasizing the importance of emotion recognition (ER) as a critical research area contributing to the overall goal of improving affective computing. Research on ER techniques has mainly relied on two conceptual emotion representation models: categorical and dimensional. Methods related to the categorical concept define the task of ER as a multi-class classification problem and categorize the speech with labels like "angry", "sad", etc. In contrast, methods based on the dimensional model attempt to regress the values of emotion-defining dimensions in a continuous range, such as "arousal/intensity" and "valence/polarity" [3].

Early SER methods extracted several low-level acoustic descriptors (LLD), such as pitch [4], loudness [5], root mean square (RMS) energy [6], zero crossing rate (ZCR), duration as well as spectral features like spectral kurtosis [7], spectral flux [3, 8], etc, from the speech stimuli to categorize them into several discrete classes. The classification algorithms mostly involved the Gaussian Mixture Model (GMM) [9], Hidden Markov Model (HMM) [10, 11], Support Vector Machine (SVM) [9], etc. Other researchers have shown that several speech signal representations, such as the Mel Frequency Cepstrum Coefficient (MFCC) [9], Linear Prediction Cepstrum Coefficient (LPCC) [12], and Log Frequency Power Coefficients (LFPC) [10], perform well to encode emotional cues from speech signals and successfully aid in identification. Later, numerous researchers began to adapt artificial neural networks (ANN) [13], convolution neural networks (CNN) [14, 15], or long-short-term memory (LSTM)-based [16, 17] recurrent networks to solve the SER task. In addition to the ways to accurately classify speech in discrete emotion categories, numerous efforts were made to predict or regress emotion dimensions [18, 19, 20, 21] such as arousal/intensity, valence/polarity, and dominance. [18] uses several LLDs and MFCCs as input features and tries to regress the value of arousal, valence, and dominance in a multi-task learning scenario and compares it with single-task learning. These methods are often criticized for poor valence prediction performance [22].

In recent years, in comparison to task-specific training, fine-tuning large self-supervised (SSL) foundation models on downstream tasks has resulted in considerable performance improvements in several audio-specific tasks [23, 24], including SER [1]. [25] fine-tuned Wav2vec 2.0 [26] and HuBERT [27] models on SER task and compared their performance for categorical emotion recognition. In contrast to categorical emotion prediction, [28] fine-tuned Wav2vec 2.0/HuBERT for arousal, valence, and dominance regression tasks and archives state-of-the-art results. The authors demonstrate that the representations learned by these SSL models considerably enhance valence predictions. Similar findings can be observed in the work by [1] where they compare a CNN model to these transformer-based general-purpose self-supervised models, demonstrating that

the latter significantly outperforms the task-specific CNN model. The authors argue that the information acquired from linguistic cues, i.e. the content of the speech, is appropriate for predicting valence, but paralinguistic cues are better suited for predicting arousal.

In connection with the findings of these studies: 1) We explore the efficacy of speech representations, learned by a state-of-the-art weakly supervised multilingual large ASR model and two other recently proposed pre-trained general-purpose SSL foundation speech encoders, by assessing their ability to predict arousal/intensity and valence/polarity accurately. We train individual regressors using these speech representations. We propose to train these SER models using a pairwise rank loss to guarantee that they can accurately predict valence and arousal scores by learning from comparing two utterances. 2) We introduce a novel late fusion architecture combining representations acquired from all three supervised ASR and self-supervised models. This architecture explores the synergies and complementarity between representations extracted from supervised ASR and SSL models in SER tasks. The evaluation results show that the proposed fusion method improves the valence/polarity and arousal/intensity prediction tasks by nearly 19.3% and 2.6% respectively, over the baseline. 3) We investigate the limitations and challenges of such methods involving representations from foundation models for the SER task by assessing the sanctity of the SER annotations and analyzing their efficacy while detecting verbal irony. 4) Finally, we explain the prediction results using frequency band occlusion and n-gram-based temporal parts-of-speech occlusion in a model-agnostic manner, demonstrating a strong relationship between these models' prior knowledge of contextual language semantics and the final emotion prediction.

## 2 Method

In this section, we describe the overall architecture of our proposed method along with the different loss functions required for training the models.

### 2.1 Architecture

Figure 1 depicts a schematic diagram of our proposed architecture. The three speech encoders used in the proposed method are Whisper [29], WavLM [30] and W2v-BERT2.0 [31]. The speech encodings from different transformer layers are passed through a Time and layer-wise Transformer (TL-Tr) layer as proposed in [32] to generate layer-wise and time-wise attention-weighted average speech embeddings before passing it on to a regression head. Finally, a regression head, comprising feed-forward layers, gives predictions for arousal/intensity and valence/polarity. The proposed fusion model relies on the late fusion of speech encoding obtained from three pre-trained transformer-based large model encoders.

**Whisper:** [29] proposed a general-purpose end-to-end ASR model that was trained with 680,000 hours of multilingual speech data in a multitask supervision

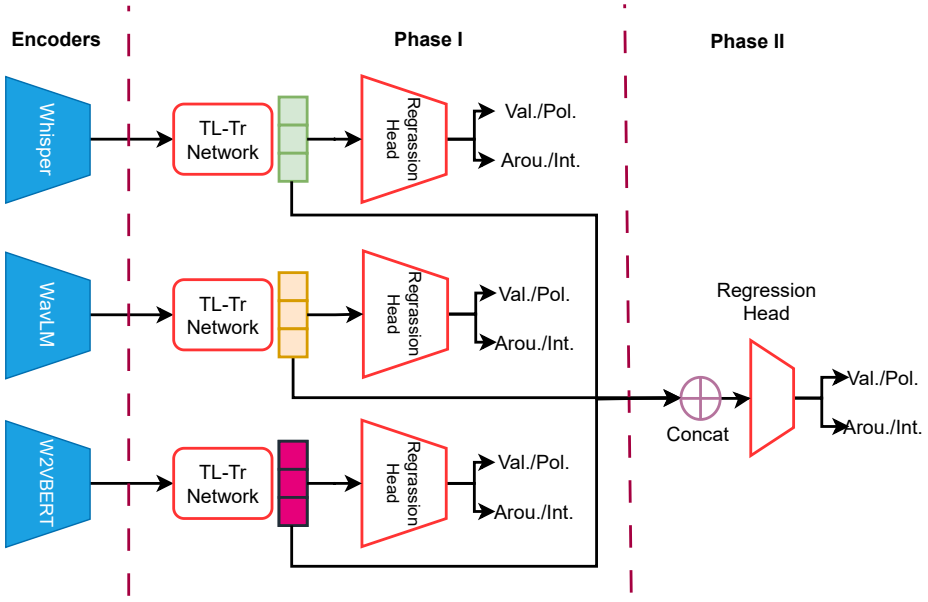


Fig. 1: Schematic diagram of the proposed method. The blue modules are fixed and the red-bordered boxes are trainable. *Val./Pol.* refers to valence/polarity and *Arou./Int.* refers to arousal/intensity.

setting including speech recognition, speech translation, language identification, etc. The encoder module contains 24 transformer layers. We choose the medium-sized model, which encodes each frame with an embedding of length 1024.

**WavLM:** A self-supervised approach for extracting universal speech representations is proposed in [30]. The model is available in a variety of configurations, and we choose the large version, which is trained on 96,000 hours of speech data. The encoder consists of 24 transformer layers as well with 1024-dimensional hidden states.

**W2v-BERT 2.0:** The SeamlessM4T framework [31] proposes a large self - supervised encoder architecture with 24 Conformer layers [33] pre-trained using a combination of contrastive learning and masked prediction learning techniques utilizing 1,000,000 hours of audio data in 143 languages. Similar to Whisper and WavLM, the hidden dimension is also 1024.

**TL-Tr:** In different encoder layers, these large foundation models encode various acoustic and linguistic information, as well as non-trivial semantic language information [34]. SSL-based SER approaches [1, 28] frequently use the encoding from the final encoder layer, ignoring any relevant information from earlier layers. Furthermore, they employ average pooling across all temporal frames, which means giving equal weight to all temporal segments, which may not be appropriate for the specific task. We adapt the TL-Tr layer proposed in [32], which uses a combination of temporal and layer-wise transformer blocks in tan-

dem to provide higher attention weights to sample specific important segments and layers before applying average pooling over temporal sequences and layers, respectively.
The TL-Tr layer outputs a 1024 dimensional embedding from the entire utterance which is passed through the regression head layer to generate final predictions.

### 2.2 Two-phase training

Our proposed framework is trained in two phases as shown in Figure 1. In phase I, individual regressors based on the three different encoders are trained separately. Unlike [1], the weights of the pre-trained encoders remain fixed during training and we only train the TL-Tr layer along with the regression head. To enable the model to learn from comparison, we propose to apply the pairwise rank loss in addition to regression loss. We create a dynamic pair for each sample, by randomly shuffling the batch. Then, the model is trained by optimizing the joint loss $L_{total}$, which is a combination of $L1$ losses and rank loss $L_{rank}$, as depicted in Eqn. 1, where $\beta$ is a hyperparameter.

$$L_{total} = \beta * (L_{val} + L_{arou}) + (1 - \beta) * L_{rank} \tag{1}$$

The model predicts arousal/intensity and valence/polarity for a pair of samples $x_1$ and $x_2$, with $L_{val1}$ and $L_{val2}$ representing the $L1$ loss for valence/polarity predictions ($v_{x1}$, $v_{x2}$) and $L_{arou1}$ and $L_{arou2}$ for arousal/intensity predictions ($a_{x1}$, $a_{x2}$). Then $L_{val}$ and $L_{arou}$ are just the additive loss for valence/polarity and arousal/intensity respectively for the pair of samples as depicted in Eqn. 2.

$$\begin{aligned} L_{val} &= L_{val1} + L_{val2} \\ L_{arou} &= L_{arou1} + L_{arou2} \end{aligned} \tag{2}$$

We implement the pair-wise rank loss as described in [35] and [36], where $L_{rank}$ is the sum of valence rank loss $L_{v\_rank}$ and arousal rank loss $L_{a\_rank}$.

$$\begin{aligned} Pr_{val} &= \frac{e^{v_{x1} - v_{x2}}}{1 + e^{v_{x1} - v_{x2}}} \\ L_{v\_rank} &= -\alpha * \ln(Pr_{val}) \\ &\quad - (1 - \alpha) * \ln(1 - Pr_{val}) \end{aligned} \tag{3}$$

The valence/polarity predictions are first mapped to a probability measure $Pr_{val}$ using a logistic transformation as depicted in Eqn. 3. Afterward, valence/polarity rank loss $L_{v\_rank}$ is computed as a negative log-likelihood loss where $\alpha$ depicts the soft ground truth rank label, such that $\alpha \in \{0, 0.5, 1\}$ depending on the value of valence/polarity ground truth of the pair. Pair-wise arousal/intensity rank loss $L_{a\_rank}$ is also calculated similarly, by replacing the valence/polarity predictions with arousal/intensity predictions in Eqn. 3. The model learns to predict the arousal/intensity and valence/polarity by optimizing L1 losses, at the same time it acquires a comparative knowledge of the predictions by optimizing the rank losses.

| Model | Valence/Polarity | Arousal/Intensity |
|---|---|---|
| W2V-L-Robust [1] | 53.86 (52.97↔54.71) | 64.73 (63.96↔65.5) |
| Whisper | **65.36** (64.64↔66.05) | 64.82 (64.04↔65.56) |
| WavLM | 58.96 (58.16↔59.75) | 65.07 (64.35↔65.71) |
| W2V-BERT 2.0 | 49.72 (48.85↔50.61) | <u>65.13</u> (64.31↔65.9) |
| Model Fusion | <u>64.25</u> (63.51↔ 65.01) | **66.42** (65.66↔ 67.13) |

Table 1: Evaluation results based on speaker-independent overall CCC (multiplied by 100) value on MSP-Podcast version 1.11 test set 1 along with bootstrap 95% confidence interval values. Best scores in bold, second underlined.

In Phase II, we concatenate the outputs of the TL-Tr layers connected to all three encoders and train a single regression head jointly with the TL-Tr layers as shown in Figure 1, producing a late fusion setup.

## 3 Experiments and ER Results

This section covers the specifics of the conducted experiments, the dataset that was utilized, and the outcomes of the tests.

### 3.1 Dataset and training details

We train and evaluate our model on the emotional podcast utterances from the MSP-Podcast corpus [37] version 1.11. The corpus comes with pre-defined train, development, and several test partitions. The train partition contains 84,030 segments from 1409 different speakers and the development set comprises 19,815 utterances from 454 speakers. We use the test1 (largest) partition as the evaluation set consists of 30,647 samples from 237 speakers. The length of the utterances varies from around 1.9 seconds to almost 11.9 seconds. For training, we pad all the samples to 12 seconds and sample them at 16 kHz frequency. In the corpus, ground truth for valence/polarity and arousal/intensity are provided in a range of 1-7, we normalize them to a range of 0-1 both for training and inference. We train the models with an H100 GPU of 80 GB memory, using a batch size of 64. For model training, Adam optimizer [38] is used with a fixed learning rate of $1e^{-4}$. For phase II training, we lower the learning rate to $5e^{-5}$. For evaluation, we use the concordance correlation coefficient (CCC) metric [39] similar to [1] separately for valence/polarity and arousal/intensity, which measures the amount of agreement between the predicted values and the ground truth. Each

model is trained for 50 epochs and we choose the best model with the lowest valence CCC loss on the development set. We also perform hyperparameter tuning on the development set and empirically choose $\beta = 0.6$. For evaluation, we calculate the overall CCC on the entire test set. The code is available online [3].

### 3.2 ER model results

The results of our experiments are summarized in Table 1. We present overall speaker-independent CCC values (multiplied by 100 for readability) along with a 95% confidence interval (CI) measured by randomly choosing 1000 bootstraps with 50% test samples each. The baseline model based on the W2v-Large-Robust encoder achieves CCC values of 53.86 and 64.73 for valence/polarity and arousal/intensity prediction, respectively, similar to the results reported in [1]. Using our suggested strategy, the weakly supervised Whisper ASR-based model achieves a CCC score of 65.36 for valence/polarity prediction which significantly outperforms the baseline ($p < 0.05$ for paired t-test based on bootstrapped CCC values) and 64.82 for arousal/intensity prediction. WavLM and W2v-BERT-based models perform similarly to the Whisper-based model in terms of arousal/intensity prediction, with CCC values of 65.07 and 65.13, respectively, although with an overlapping CI with the baseline model. For valence/polarity prediction, the WavLM-based model obtains a CCC value of 58.96, however, the W2v-BERT-based model only achieves a value of 49.72, which is lower than the baseline model.

Our proposed fusion framework significantly outperforms the baseline framework described by [1] for both valence/polarity and arousal/intensity predictions (completely non-overlapping CI and $p < 0.05$ for paired t-test based on bootstrapped CCC values for both valence/polarity and arousal/intensity). It obtains a CCC value of 66.42 for arousal/intensity prediction, which is higher than the baseline. Similarly, also for valence/polarity prediction, the fusion model surpasses the baseline significantly by attaining a CCC value of 64.25, though it is somewhat similar to the supervised Whisper ASR-based model as the CIs overlap.

### 3.3 Discussion of ER model results

The results shown in Table 1 are consistent with findings from the SER literature. The Whisper ASR-based model has the highest CCC (65.36) for valence/polarity prediction, indicating that models with prior knowledge of speech recognition, speech translation, or language processing perform better than general-purpose models, as valence/polarity is heavily influenced by linguistic semantics [1]. On the other hand, general-purpose self-supervised models are not better at predicting valence/polarity than the ASR-based model indicating that they comparatively lack linguistic semantic information in their encoding. However, the other two models based on the general purpose representation extracted from

[3] https://github.com/arnabdas8901/MSPPodcast_ContinuousEmotionRecognition

the WavLM and W2v-BERT 2.0, are as good as (overlapping CI) the Whisper-based model in terms of CCC values for arousal/intensity prediction. Our proposed fusion model combines the benefits of both paradigms, supervised ASR and general-purpose SSL. It achieves the highest overall CCC value (66.42) for arousal/intensity prediction, with results significantly higher than both the baseline and the whisper-based model. Similarly, the valence/polarity prediction performance (64.25) is also comparable with Whisper and significantly higher(non-overlapping CI and $p < 0.05$) than the WavLM and W2v-BERT-based models. The results also highlight that the latest foundation models are better in SER tasks than relatively older large models as both Whisper and WavLM-based models show improved performance compared to the W2V-Large-Robust model proposed as part of the baseline.

## 4 Post-Hoc Data Exploration

In order to gain more insights into the nature and quality of the data, as well as in order to provide explanations of model behavior we further extend our experiments to post-hoc analyses and visualizations.

### 4.1 Annotation inconsistency

Based on the results presented in Table 1, further investigation reveals that the dataset contains a lot of inconsistent annotation both for arousal/intensity and valence/polarity ratings and for primary emotion categories even after considering aleatoric uncertainty for annotations.

```
WORKER00006951; Sad; Neutral; A:3.000000; V:2.000000; D:3.000000;
WORKER00008706; Sad; Happy,Amused,Excited; A:3.000000; V:2.000000; D:3.000000;
WORKER00008722; Sad; Happy,Amused,Excited; A:3.000000; V:2.000000; D:3.000000;
WORKER00008714; Sad; Happy,Excited; A:3.000000; V:2.000000; D:3.000000;
WORKER00008725; Sad; Happy,Amused,Surprise,Excited; A:3.000000; V:2.000000; D:3.000000;
WORKER00009045; Sad; Neutral,Excited; A:3.000000; V:2.000000; D:3.000000;
WORKER00003609; Sad; Neutral; A:3.000000; V:2.000000; D:3.000000;
WORKER00009220; Sad; Happy,Amused,Excited; A:3.000000; V:2.000000; D:3.000000;
WORKER00009308; Sad; Happy,Amused,Excited; A:3.000000; V:2.000000; D:3.000000;
WORKER00006768; Sad; Other-humorous ; A:3.000000; V:2.000000; D:3.000000;
```

Fig. 2: Inconsistent emotion annotations with primary and secondary emotion labels are contradictory, while arousal/intensity and valence/polarity ratings indicate more towards neutral. A listening test reveals that the utterance is not *sad* at all.

Figure 2 shows an example in which a sample is annotated by 10 annotators where most of the annotators have indicated mutually contradictory emotions like *"Sad"* and *"Happy"* at the same time. Upon listening, it is clear that the utterance is not sad at all as someone is welcoming guest hosts in a show with the spoken content as *"we've got some really great guest hosts, if you hate me, you'll*

*love them"*. The ground truth emotion category is provided as *"Sad"* considering majority voting. In another example, a lady can be heard giving a eulogy with the message: *"rest in peace. he was a great coach. he was one of my favorite coaches. again, got me into basketball. it's got me wanting to play basketball"*. The sample's primary emotion is provided as *angry*, with the highest possible arousal/intensity ground truth rating of 1. A hearing test reveals that the ground truth rating is incorrect. Such inconsistencies in the training set may prevent the model from reaching its full potential, and the existence of such examples in the test set may imply an incorrect assessment result. To check whether the annotations are correct, a simple solution could be to inspect a subset of samples manually in both the training and development splits, where the trained model makes high prediction errors.

| Dimensions | with phrase *"fu**ing"* & val. $> 0.5$ | others |
|---|---|---|
| Valence/Polarity | 20.19 | 64.35 |
| Arousal/Intensity | 53.57 | 66.41 |

Table 2: Comparative valence/polarity evaluation result (CCC$\times$100) for samples' group containing a slang word *"fu**ing"* and high valence/polarity in to other samples

### 4.2 Verbal irony

We also conducted a post-hoc root cause analysis of the results for utterances in which our suggested model did not perform well and discovered some intriguing findings. The dataset comprises podcast talks, where slang phrases like "fu**ing" [4] are often used. Native speakers frequently use such slang phrases to express verbal irony, which occurs when the content of a speech is negative but the overall sentiment or emotion is positive, or vice versa. For example, the test set contains a sample, whose content is *"seth mcfarland's voice is fu**ing amazing"*, having a ground truth valence/polarity rating of 0.7, but our model predicts the valence/polarity as 0.35. To further analyze the influence of verbal irony in valence/polarity prediction, we divide the test samples into two categories: samples containing the word *"fu**ing"* with a valence/polarity ground truth $> 0.5$ and other samples. The evaluation results are presented in Table 2. The results reveal that for the group of samples containing verbal irony, the valence/polarity prediction performance declines dramatically to a CCC value of 20.19, but for the others, it remains as high as 64.35. In contrast, the decrease in arousal/polarity prediction performance is not as drastic.

[4] The characters "**" are introduced to maintain the decorum of the paper, not present in actual dataset

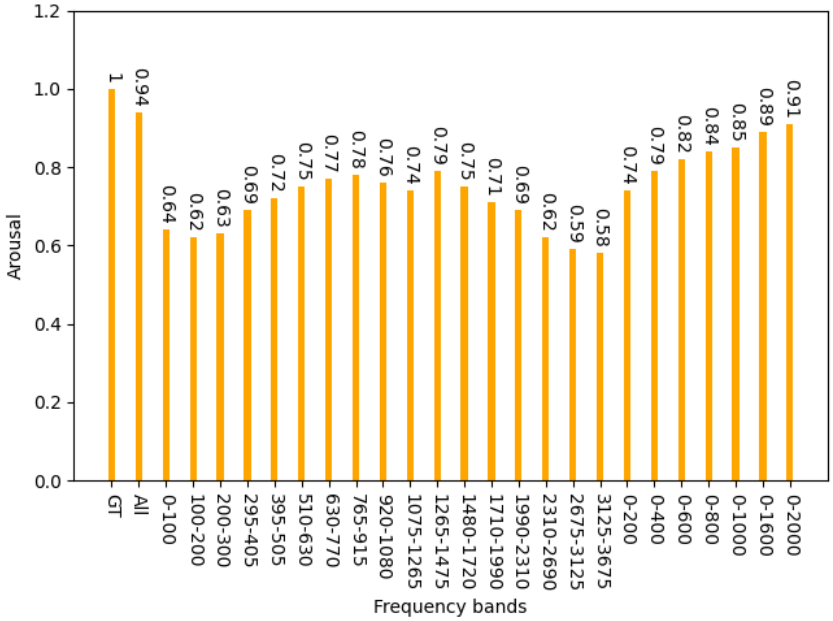


Fig. 3: Impact of critical frequency pass-bands on arousal/intensity prediction. Only the bands mentioned in the x-axis ticks are passed and other frequencies are suppressed. *All* refers to no suppression, passing all frequencies. *GT* refers to ground truth annotation.

### 4.3 Occlusion-based explainability

In addition to improving model performance, explainability of model behavior becomes crucial for SER tasks, as it does for other tasks. Concurrent work on explaining audio classification explored backpropagation-based [40] approaches or used LIME (surrogate linear models) for explaining phoneme recognition [41]. [42] used input perturbation techniques to measure the contribution of word-level content paralinguistic features, resulting in more plausible explanations. Saliency-based approaches are frequently criticized, because the attribution maps generated by these methods are unreliable and do not provide a genuine assessment of model behavior [43]. Patch-based occlusion methods, which are used in the image domain [44], are similarly unsuitable for the audio domain, because audio representation differs from pixelated picture representations. Hence, we tailor the occlusion sensitivity analysis to suit the bimodal properties (linguistic and paralinguistic) of speech data.

We investigate changes in model prediction utilizing spectral and temporal occlusion on our proposed fusion model. First, we examine the sensitivity of arousal/intensity regression by just filtering the stimuli over the critical frequency bands [45] and their combinations. We accomplish this on a sample with a valence/polarity ground truth of 0 and an arousal/intensity ground truth of

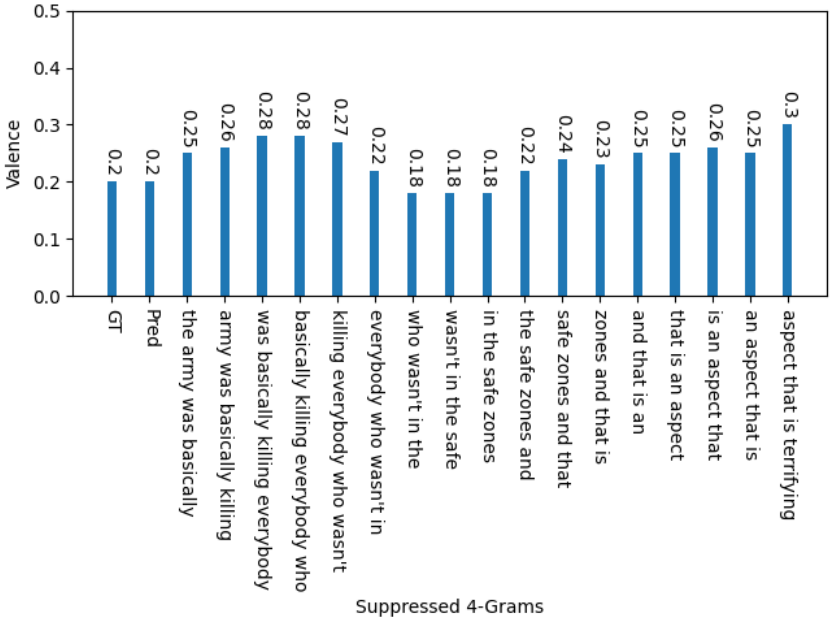


Fig. 4: Low valence/polarity prediction explainability using 4-gram part of speech occlusion. *GT* refers to ground truth annotation and *Pred* refers to overall model prediction without any occlusion.

1, iteratively passing it through critical bandpass filters as a pre-processing step before capturing the model's reaction. The findings are depicted in Figure 3. The results reveal that the arousal/intensity value decreases as we move through extremely low or very high-frequency bands of the spectrum. When all frequencies are fed into the model, the arousal/intensity value is most accurately predicted. Since the occlusion-based method is sample-specific, the contribution of each frequency band may vary for different samples depending on the paralinguistic factors (crying, laughing, or yelling) that are present.

As part of temporal occlusion, we also use a 4-gram-based part of speech occlusion to better understand the language reliance on the valence/polarity predictions as well as contextual awareness. The MSP-Podcast dataset used in our experiment includes textual transcriptions and a text grid file for each audio, which provides the start and end times for each spoken word. For our 4-gram-based occlusion method, we used the start time of the first word and the end time of the last word in a 4-gram. We then zeroed out the actual audio within these selected timestamps to create an occluded version of the audio. This occluded audio is subsequently passed through our arousal/valence prediction pipeline. This process is repeated for all the 4-grams, with only the section belonging to each specific 4-gram being occluded in each iteration. We used a sample from test set 1 with the content *"the army was basically killing everybody who wasn't*

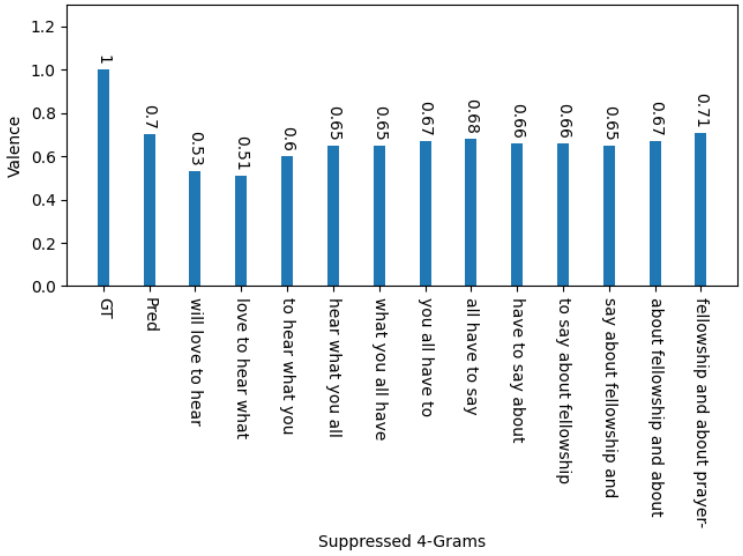


Fig. 5: High valence/polarity prediction explainability using 4-gram part of speech occlusion. *GT* refers to ground truth annotation and *Pred* refers to overall model prediction without any occlusion.

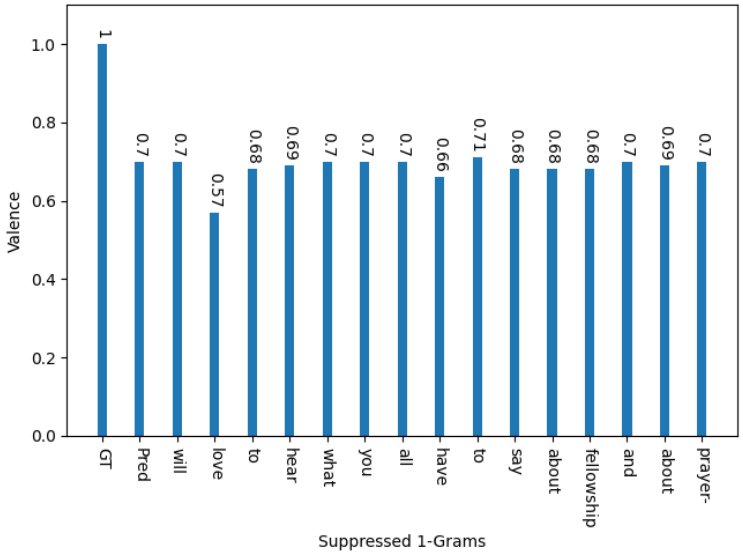


Fig. 6: Valence/polarity prediction explainability using 1-gram part of speech occlusion. *GT* refers to ground truth annotation.

*in the safe zones, and that is an aspect that is terrifying"* having a ground truth valence/polarity rating of 0.2. Before handing input to the model, we mask 4-gram phrases by sliding a word each time. The result is illustrated in Figure 4. The outcome demonstrates that, in the absence of any occlusion, the overall valence/polarity prediction and the ground truth are equal. However, when the 4-gram phrases containing the word *killing* are suppressed the valence score increases towards the neutral valence/polarity score. Additionally, when the last 4-gram phrase containing the word *terrifying* is suppressed the prediction goes even further up towards neutral with a score of 0.3. On the contrary, in the middle part of the utterance when the 4-grams with a positive word *safe* is suppressed, the valence/polarity prediction goes further negative to 0.18.

We experimented with the same 4-gram based temporal occlusion on another example with content *"will love to hear what you all have to say about fellowship and about prayer"* having ground truth valence/polarity rating of 1 and the result is demonstrated in Figure 5. The overall prediction without occlusion is 0.7. The result shows that the 4-gram phrases containing the word *love* when suppressed drastically change the model's prediction and bring down the valence score (0.53 and 0.51 ) very close to the neutral scale, However, a minimal change is observed when other 4-grams are suppressed. In contrast, when we perform the same experiment with 1-gram-based part of speech masking, the predictions hardly vary except for single word *love*, as depicted in Figure 6. This shows a strong linguistic contextual dependency of the model's valence/polarity prediction.

A similar experiment using temporal occlusion is also performed for arousal or intensity prediction using a sample containing content *"right. it's almost like a dig. person just walks"* having arousal/intensity ground truth of 0.83. The speaker can be heard laughing at the start of the utterance which constitutes a positive arousal/intensity rating as the spoken content is otherwise not very expressive. The result of the experiment, as depicted in Figure 7, shows that the overall arousal/intensity prediction by the model is 0.75. However, when the first 4-gram phrase, which overlaps with the speaker's smiling, is suppressed the arousal/intensity substantially comes down to 0.66. The remaining 4 grams do not affect the model's prediction when suppressed. This explainability result re-affirms that paralinguistic cues have a far greater role in predicting arousal/intensity than linguistic content.

These trends regarding arousal/intensity and valence/polarity predictions and their relation with semantic or paralinguistic aspects of the utterance are observed in many other samples, only a few are presented in this section as examples.

## 5 Conclusion

In this paper, we investigated the effectiveness of the speech representations learned by foundation models for ER using arousal/intensity and valence/polarity prediction tasks. The result demonstrates that representation from a large multilingual ASR model like Whisper is beneficial for valence/polarity regression as

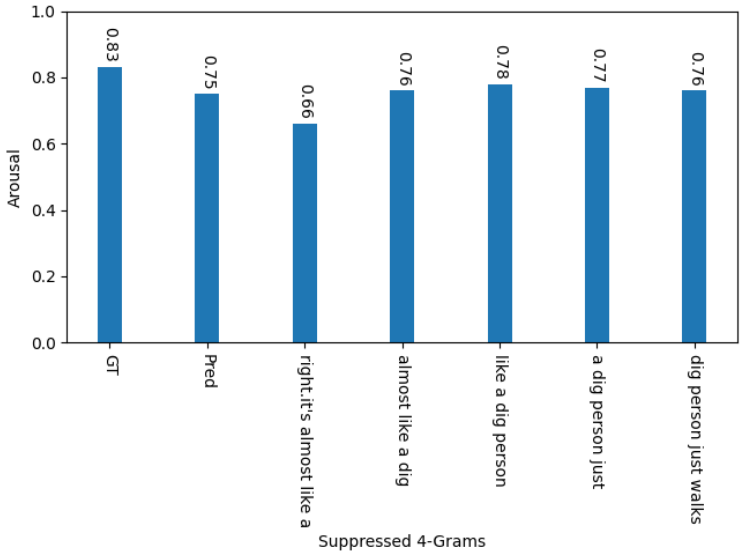


Fig. 7: Arousal/intensity prediction explainability using 4-gram part of speech occlusion. *GT* refers to ground truth annotation and *Pred* refers to model prediction without any occlusion.

it achieves a 17.6% higher CCC value over the baseline. On the other hand, it is comparable to models based on representations learned by other recent general-purpose SSL models for arousal/intensity prediction. The detailed evaluation results also depict that our proposed fusion model can combine and complement both types of representations to accurately predict valence/polarity and arousal/intensity as it significantly outperforms both the baseline and the ASR-based model. However, the foundation models remain limited in their ability to fully encode all the complexities of a spoken language, such as verbal irony. Further research is necessary to develop foundation models for general-purpose speech encoding that are better able to handle downstream tasks and understand the complexities and minute nuances of speech. These models should be trained with speech data that demonstrates a wider range of compound emotions. We also showed that large datasets like MSP-Podcast, often used for SER model training, contain incorrect annotations which is detrimental to model training and evaluation. In addition, we suggest speech-appropriate occlusion-based techniques to elucidate the model's performance in SER tasks and offer insights regarding the reason for the predictions in a model-agnostic way. Our future research endeavors will expand upon explainability methodologies by producing significance scores for distinct utterance segments that substantiate the model's prediction. Additionally, by incorporating a cross-attention mechanism, we will investigate more advanced fusion techniques to capture the advantages of different representations better. We also intend to work on integrating our va-

lence/polarity and arousal/intensity detection model with large language models (LLMs) to identify verbal irony and other complex sentiments from spoken speech data.

## 6 Acknowledgements

This research has been partly funded by the Federal Ministry of Education and Research Germany (BMBF 16KISA007, project Medinym) and partly by the Volkswagen Foundation.

## 7 Limitations

Despite the encouraging results of this study on continuous emotion prediction using speech features from pre-trained acoustic foundation models and the MSP-Podcast dataset, certain limitations must be recognized. The podcast recordings that make up the MSP-Podcast dataset may display a particular range of emotions and speech patterns that aren't typical of other contexts like conversational speech, call centers, or clinical settings. This domain exclusivity can somewhat constrain the generalizability of our approach and make it more difficult for the model to be applied to out-of-domain spoken content without additional training. Furthermore, continuous real-time emotion prediction requires low latency and significant processing power, especially for deep learning models. This restriction could make it more difficult to implement the suggested method in real-time applications on devices with constrained memory or computing power.

# Bibliography


[1] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[2] Leimin Tian, Catherine Lai, and Johanna Moore. Polarity and intensity: the two aspects of sentiment analysis. In Amir Zadeh, Paul Pu Liang, Louis-Philippe Morency, Soujanya Poria, Erik Cambria, and Stefan Scherer, editors, *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 40–47, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[3] Felix Weninger, Florian Eyben, Björn W Schuller, Marcello Mortillaro, and Klaus R Scherer. On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology*, 4:292, 2013.

[4] Björn Schuller, Martin Wöllmer, Florian Eyben, and Gerhard Rigoll. Prosodic, spectral or voice quality? feature type relevance for the discrimination of emotion pairs. 2009.

[5] Florian Eyben. *Real-time speech and music classification by large audio feature space extraction*. Springer, 2015.

[6] Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Zhigang Deng, Sungbok Lee, Shrikanth Narayanan, and Carlos Busso. An acoustic study of emotions expressed in speech. In *Eighth International Conference on Spoken Language Processing*, 2004.

[7] Björn Schuller, Dejan Arsic, Frank Wallhoff, and Gerhard Rigoll. Emotion recognition in the noise applying large acoustic feature sets. 2006.

[8] Purnima Chandrasekar, Santosh Chapaneri, and Deepak Jayaswal. Emotion recognition from speech using discriminative features. *International Journal of Computer Applications*, 101(16):31–36, 2014.

[9] Iker Luengo, Eva Navas, Inmaculada Hernáez, and Jon Sánchez. Automatic emotion recognition using prosodic parameters. In *Ninth European conference on speech communication and technology*. Citeseer, 2005.

[10] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.

[11] Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B Mariño. Speech emotion recognition using hidden markov models. In *Seventh European conference on speech communication and technology*, 2001.

[12] Peipei Shen, Zhou Changjun, and Xiong Chen. Automatic speech emotion recognition using support vector machine. In *Proceedings of 2011 international conference on electronic & mechanical engineering and information technology*, volume 2, pages 621–625. IEEE, 2011.

[13] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181, 2006.

[14] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014.

[15] WQ Zheng, JS Yu, and YX Zou. An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 international conference on affective computing and intelligent interaction (ACII)*, pages 827–831. IEEE, 2015.

[16] Wootaek Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*, pages 1–4. IEEE, 2016.

[17] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2227–2231. IEEE, 2017.

[18] Srinivas Parthasarathy and Carlos Busso. Jointly predicting arousal, valence and dominance with multi-task learning. In *Interspeech*, volume 2017, pages 1103–1107, 2017.

[19] Meysam Asgari, Géza Kiss, Jan Van Santen, Izhak Shafran, and Xubo Song. Automatic measurement of affective valence and arousal in speech. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 965–969. IEEE, 2014.

[20] Khiet P Truong, David A van Leeuwen, Mark A Neerincx, and FM Jong. Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion. 2009.

[21] Bagus Tris Atmaja, Yasuhiro Hamada, and Masato Akagi. Predicting valence and arousal by aggregating acoustic features for acoustic-linguistic information fusion. In *2020 IEEE REGION 10 CONFERENCE (TENCON)*, pages 1081–1085. IEEE, 2020.

[22] Jonathan Chang and Stefan Scherer. Learning representations of emotional speech with deep convolutional generative adversarial networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2746–2750. IEEE, 2017.

[23] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Bjoern W Schuller. Audio self-supervised learning: A survey. *Patterns*, 3(12), 2022.

[24] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

[25] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition,

speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*, 2021.

[26] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[27] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[28] Sundararajan Srinivasan, Zhaocheng Huang, and Katrin Kirchhoff. Representation learning through cross-modal conditional teacher-student training for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6442–6446. IEEE, 2022.

[29] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[30] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[31] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023.

[32] Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. *arXiv preprint arXiv:2307.03183*, 2023.

[33] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[34] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE, 2021.

[35] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[36] Kexin Wang, Yunlong Zhao, Qianqian Dong, Tom Ko, and Mingxuan Wang. Mospc: Mos prediction based on pairwise comparison. *arXiv preprint arXiv:2306.10493*, 2023.

[37] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2017.

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

[39] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

[40] Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428, 2024.

[41] Xiaoliang Wu, Peter Bell, and Ajitha Rajan. Can we trust explainable ai methods on asr? an evaluation on phoneme recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10296–10300, 2024.

[42] Eliana Pastor, Alkis Koudounas, Giuseppe Attanasio, Dirk Hovy, and Elena Baralis. Explaining speech classification models via word-level audio segments and paralinguistic features. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2221–2238, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[43] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[44] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[45] Mandar A Rahurkar, John HL Hansen, James Meyerhoff, George Saviolakis, and Michael Koenig. Frequency band analysis for stress detection using a teager energy operator based feature. In *INTERSPEECH*, 2002.