Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Pitch Contour Model (PCM) with Transformer Cross-Attention for Speech Emotion Recognition

Minji Ryu[1], *Ji-Hyeon Hur*[2], *Sung Heuk Kim*[2], *Gahgene Gweon*[2,3]

[1]Interdisciplinary Program in Cognitive Science, Seoul National University, Republic of Korea
[2]Department of Intelligence and Information, Seoul National University, Republic of Korea
[3]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Republic of Korea
{ten_wins_of_two, jh.hur, tyvod30, ggweon}@snu.ac.kr

Abstract

Pitch is important for distinguishing emotional states through intonation. To incorporate pitch contour patterns into Speech Emotion Recognition (SER) task, we propose the Pitch Contour Model (PCM), which integrates pitch features with Transformer-based speech representations. PCM processes pitch features via linear embedding and combines them with Wav2Vec 2.0 extracted features using cross-attention. Experimental results show that PCM enhances SER performance, achieving state-of-the-art (SOTA) Valence-Arousal-Dominance (V-A-D) scores with V:0.627, avg:0.571 in MSP-Podcast v1.11 and V:0.646, A:0.744, D:0.557 in IEMOCAP datasets. We observe that the effect of z-score normalization on pitch varies across datasets, with lower pitch variability conditions benefiting more from raw pitch values. Furthermore, our study suggests how pretraining and finetuning language mismatches, between English and Korean, affect the choice between CNN-based and linear embeddings for pitch representation.

Index Terms: speech emotion recognition, normalized pitch contour, cross-attention

1. Introduction

The ability of artificial intelligence models to recognize emotions from speech is essential for understanding natural conversations. Speech Emotion Recognition (SER) automatically identifies a speaker's emotional state based on vocal cues, enabling more effective conversational understanding by capturing fundamental aspects of emotional expression. In SER research, emotions are typically classified into two categories: categorical emotions and dimensional emotions [1]. Our study focuses on dimensional emotion recognition based on the Valence-Arousal-Dominance (V-A-D) framework.

In recent SER models, advancements have been made by incorporating Transformer-based speech processing models, such as Wav2Vec 2.0 [2] and HuBERT [3]. In SER tasks, transformers generally extract acoustic features in the lower layers while processing linguistic and contextual information in the higher layers [4, 5]. The final Transformer layers are commonly used in SER, as contextual information contributes significantly to the accuracy of emotion recognition. However, a limitation of Transformer-based models is that acoustic features, which are essential for emotion recognition—particularly pitch information—tend to weaken or diminish in higher layers [6].

The pitch is a crucial cue to distinguish emotional states, providing essential prosodic information such as intonation [7]. However, previous studies suggest that it is not the absolute pitch values, but rather the pitch contour patterns, that play a more significant role in SER [7]. Therefore, we integrated pitch contour patterns into a transformer-based model to improve SER performance.

To further enhance SER model we're trying to build, we used a linear layer instead of CNN layers for pitch processing. In SER research, a previous study incorporated CNN layers to embed pitch values, focusing on capturing localized features across time intervals [8]. While this approach is effective, it may emphasize certain pitch values within segmented regions, potentially altering the pitch contour. In contrast, we considered that a linear layer, by leveraging global features instead of relying solely on local segments, could better preserve the overall shape of the pitch contour. Based on this motivation, we propose the Pitch Contour Model (PCM), which employs a linear layer for pitch embedding instead of CNN layers[1].

To validate the effectiveness of different pitch embedding strategies, we implement three PCM variants: 1) PCM with Linear Embedding and z-score Normalization (PCM-le-Norm), 2) PCM with Linear Embedding without z-score Normalization (PCM-le-noNorm), 3) PCM with CNN-based Embedding (PCM-cnn). By comparing these models, we examine how different embedding methods influence SER performance, particularly in preserving pitch contour patterns. Additionally, we include a Base Model (BM) without explicit pitch integration to assess the direct impact of pitch information on SER performance. We evaluate PCM on two English datasets (IEMOCAP [9] and MSP-Podcast v1.11 [10]) as well as Korean Multimodal Video Dataset (MVD) [11] to explore its applicability in a bilingual setting. The key contributions are as follows:

- We demonstrate that PCM with linear embedding models (PCM-le-noNorm, PCM-le-norm) outperform PCM with CNN-based embedding model (PCM-cnn) in capturing global pitch contour patterns, leading to SOTA performance in V-A-D predictions across multiple datasets. Notably, PCM with linear embedding models achieve the highest CCC scores on MSP-Podcast v1.11 (CCC_V: 0.627, CCC_avg: 0.571), and on IEMOCAP (CCC_V: 0.646, CCC_A: 0.744, CCC_D: 0.557).
- We observe that z-score normalization of pitch does not consistently improve SER performance. Datasets with lower pitch variability (MSP-Podcast) benefit more from non-normalized pitch values, while high-variability datasets (IEMOCAP) do not benefit from normalization.
- We show that CNN-based embedding outperforms linear embedding when the pretraining and finetuning languages are mismatched (English-pretrained PCM-cnn on MVD, CCC_avg: 0.555); Whereas linear embedding outperforms CNN-based embedding when the languages are matched (Korean-pretrained PCM-le-norm on MVD, CCC_avg: 0.553).

[1] https://github.com/snucclab/PCM

10.21437/Interspeech.2025-1233

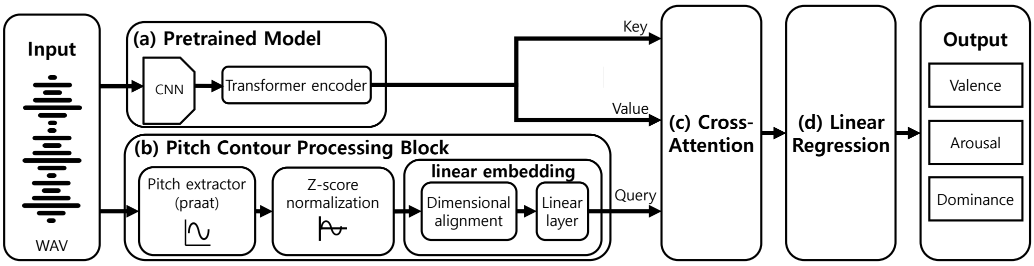

Figure 1: *The PCM model is composed of four main parts: (a) Pretrained Model, (b) Pitch Contour Processing Block, (c) Cross-Attention, and (d) Linear Regression. All models share the same input format (WAV files) and output format (V-A-D scores).*

2. Related Work

Earlier SER models have utilized CNNs [12, 13, 14]. Latif et al. [15] proposed a method to directly process raw speech data to extract low-level speech features through CNNs, and predict emotions based on these features. However, existing CNN-based models have shown limitations in predicting V and D on the IEMOCAP dataset (CCC$_\text{V}$: 0.259, CCC$_\text{A}$: 0.431, CCC$_\text{D}$: 0.272) [16]. To address these limitations, recent research has incorporated transformers into existing CNN architectures. Wagner et al. [16] significantly improved the performance of V prediction in SER by utilizing a transformer model on the IEMOCAP dataset (CCC$_\text{V}$: 0.448, CCC$_\text{A}$: 0.663, CCC$_\text{D}$: 0.518).

Transformer-based SER research has utilized transformer-based self-supervised speech representation models to extract speech features [16, 17, 18, 19, 20]. Transformer-based speech representation models, such as HuBERT [3] and Wav2Vec 2.0 [2], combine CNNs for low-level feature extraction with transformers for capturing global contextual dependencies. Particularly, Wav2Vec 2.0 shows notable performance in various speech tasks and demonstrates outstanding results in SER [16, 19]. Wav2Vec 2.0 is trained on the masked speech representation prediction tasks, with lower layers learning generalized features and the higher layers focusing on task-specific features [4]. Despite the advantages of transformer-based models, such as Wav2Vec 2.0, may lose crucial pitch information for SER during the feature extraction process. Thanh et al. [8] emphasized that pitch plays a vital role in SER, particularly in tonal languages. Additionally, Rodero et al. [7], through research in speech psychology, highlighted that pitch is an important feature for recognizing emotions conveyed through speech. Building on these findings [8, 7], our study proposes integrating pitch information to enhance SER performance in non-tonal languages, such as English and Korean.

3. Pitch Contour Model

3.1. Model Architecture

The proposed Pitch Contour Model (PCM) consists of four main components: (a) Pretrained Model, (b) Pitch Contour Processing Block, (c) Cross-Attention, and (d) Linear regression, as illustrated in Fig. 1. The following subsections provide details of each component.

3.1.1. Pretrained Model

(a) Pretrained Model is the component that extracts Transformer-based speech features. The pretrained model used in PCM is wav2vec2-large-robust [21], chosen based on its superior performance in prior SER studies [16]. The model consists of a CNN feature extractor followed by a Transformer encoder, which processes the input waveform and extracts contextualized speech representations. These representations serve as the Key $K \in \mathbb{R}^{\text{seq\_len} \times d}$ and Value $V \in \mathbb{R}^{\text{seq\_len} \times d}$ for the cross-attention mechanism, as formulated in Eq. 1. The pretrained Wav2Vec2 model was used as a feature extractor, and its parameters remained frozen during training.

$$K, V = f_{\text{transformer}}(X) \tag{1}$$

3.1.2. Pitch Contour Processing Block

Parallel to the pretrained model, (b) Pitch Contour Processing Block extracts and embeds pitch features to be used as the Query in the cross-attention. The block consists of three stages: pitch extraction, z-score normalization, and linear embedding. Pitch features are extracted using the Praat-based Parselmouth library, which computes pitch values from the input waveform. The extracted pitch values undergo z-score normalization to reduce speaker- and language-dependent variations while preserving pitch contour patterns. The normalized pitch features are then embedded using a two-step transformation. First, a linear transformation aligns the pitch features with the Transformer-based speech representations as shown in Eq. 2. where $P \in \mathbb{R}^{\text{seq\_len} \times 1}$ is the original pitch vector, and $P' \in \mathbb{R}^{d \times 1}$ is the transformed pitch representation. Second, the transformed pitch features are passed through another linear layer to generate the Query (Q) as shown in Eq. 3. where $W_{\text{enc}} \in \mathbb{R}^{d \times d}$ and $b_{\text{enc}} \in \mathbb{R}^{d \times 1}$ are the weight and bias parameters, respectively.

$$P' = W_{\text{align}} P \tag{2}$$

$$Q = W_{\text{enc}} P' + b_{\text{enc}} \tag{3}$$

3.1.3. Cross-Attention

(c) Cross-Attention integrates pitch and Transformer-based speech features by treating the pitch-embedded *Q* as the Query and the Transformer-extracted *K, V* as the Key and Value, respectively. This process is formulated as shown in Eq. 4. where d_k is the scaling factor for dot-product attention. This mechanism ensures that pitch information guides the refinement of speech representations without altering the inherent structure of speech representations.

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{4}$$

3.1.4. Linear Regression

The final component of PCM is (d) Linear Regression, which maps the attention-weighted output $Z \in \mathbb{R}^{1 \times d}$ to continuous V-A-D scores. This layer enables the model to generate predictions for Valence, Arousal, and Dominance dimensions.

3.2. PCM Variants

To investigate the effects of different pitch embedding strategies and normalization methods, we evaluate three PCM variants against a Base Model (BM):

- **Base Model (BM)**: A model without explicit pitch integration consisting only of (a) Pretrained Model and (d) Linear Regression.
- **PCM with Linear Embedding and z-score Normalization (PCM-le-norm)**: Our primary model, which incorporates z-score normalization before applying linear embedding in (b) Pitch Contour Processing Block.
- **PCM with Linear Embedding without z-score Normalization (PCM-le-noNorm)**: Identical to PCM-le-norm but omits z-score normalization in (b) Pitch Contour Processing Block, directly applying linear embedding to raw pitch values.
- **PCM with CNN-based Embedding (PCM-cnn)**: Replaces the linear embedding in (b) Pitch Contour Processing Block with a CNN-based embedding to extract pitch features before cross-attention.

By comparing these variants, we analyze the impact of pitch integration, embedding strategy, and normalization on SER performance.

4. Experiments

4.1. Datasets

We used three datasets in our experiment: MSP-Podcast [10] and IEMOCAP [9] for English, and the Multimodal Video Dataset (MVD) [11] for Korean. The IEMOCAP dataset contains 10,039 utterances from 10 actors, recorded over five sessions. Each utterance is annotated with V-A-D dimensions. To ensure a speaker-independent setup, we used Sessions 1–3 for training, Session 4 for validation, and Session 5 for testing. The MSP-Podcast v1.11 dataset consists of 151,654 utterances from over 2,172 speakers, with each utterance labeled with V-A-D dimensions. The dataset is pre-divided into train, validation, and test sets, and we used the provided splits as is, maintaining speaker independence. The MVD dataset comprises 85,077 utterances from 300 Korean actors and provides annotations for V and A dimensions, but excludes the D dimension. To evaluate the model's applicability in a Korean context, we sampled 10,000 utterances from MVD. The original V-A scores (ranging from 1 to 10) were divided into four intervals, ensuring an equal number of samples from each range. Table 1 summarizes each dataset, including sample counts, emotion labels, and language.

4.2. Experimental Settings

All models were trained on three datasets with a batch size of 1, using the AdamW optimizer and a learning rate of 1e-5. The MVD and IEMOCAP datasets were trained for 100 and 500 epochs, respectively, while the MSP-Podcast dataset was trained for 10 epochs. We used Mean Squared Error (MSE) loss to predict V-A-D scores as floating-point values, with performance evaluated using the Concordance Correlation Coefficient (CCC), a widely adopted metric in dimensional SER tasks.

To understand how different pitch embedding methods affect model performance, we compared four model configurations (BM, PCM-le-noNorm, PCM-le-norm, PCM-cnn) across the IEMOCAP, MSP-Podcast, and MVD datasets. These models represent PCM variants, designed to test the impact of embedding strategies. All models were initialized with the same randomly initialized parameters for a fair comparison. We also evaluated an alternative pretrained model trained on Korean speech [22] and used it to train and compare BM and the three PCM variants on the MVD dataset. This experiment aimed to analyze how pretraining language matching and pitch embedding strategies influence SER performance. We then compared the performance of our PCM models against existing SOTA models to demonstrate their advantages.

To validate our PCM variants, we compared the performance of our four models with the top three models from the Odyssey 2024 - Speech Emotion Recognition Challenge [23] on the MSP-Podcast dataset. The top model, Goncalves.a, utilized the pretrained WavLM [24] model and attentive statistics pooling [25], while Goncalves.b and Goncalves.c were extended versions of Goncalves.a. For the IEMOCAP dataset, we compared our four models with Atmaja's multimodal model [26], which combines text and audio. We also compared them with Srinivasan's transfer learning-based audio-only models (Srinivasan.a) [27] and its improved version (Srinivasan.b), which uses residual-based filtering of low-quality teacher predictions.

Table 1: *Overview of SER Datasets, including sample counts, emotion labels, language and difference in pitch standard deviation (SD) between high and low V-A-D scores for each dataset*

	MSP-Podcast (English)	IEMOCAP (English)	MVD (Korean)
Train	84030	5766	8000
Validation	19815	2103	1000
Test	30647	2170	1000
SD diff. V	4.56	16.74	16.01
SD diff. A	15.50	28.87	13.04
SD diff. D	9.88	13.69	-
SD diff. avg	9.98	19.77	14.53

5. Results and Discussion

The proposed Pitch Contour Model (PCM) consistently outperformed both the Base Model (BM) and prior studies [23, 26, 27], across multiple datasets. As shown in Table 2, the highest performance values for CCC_V, CCC_A, CCC_D, and CCC_{avg} in each dataset are highlighted in bold. By observing the bold values in Table 2, it is evident that PCM models (PCM-le-noNorm, PCM-le-norm, PCM-cnn) achieved superior results in most cases, demonstrating the effectiveness of integrating pitch information. Notably, PCM-le-noNorm exhibited the highest performance, including achieving the best CCC_{avg} score (0.571) on MSP-Podcast. In IEMOCAP, PCM with linear embedding (PCM-le-noNorm, PCM-le-norm) outperformed prior studies [23, 26, 27] and PCM-cnn. PCM-le-norm achieved a high CCC_V score (0.646) and CCC_A score (0.744), while PCM-le-noNorm demonstrated superior performance with a CCC_D score (0.557) and CCC_{avg} score (0.646). These results highlight the advantage of preserving overall pitch contour patterns. The specific results for each dataset are as follows.

For MSP-Podcast v1.11, the PCM-le-noNorm model yielded the highest CCC_{avg} score (0.571), surpassing both PCM-cnn and the BM. The CCC_V score (0.627) achieved SOTA results, suggesting the effectiveness of linear embedding in capturing global pitch trends. PCM-cnn recorded the second-highest performance, showing notable improvements in CCC_A score (0.621) and CCC_D score (0.536). This suggests that localized feature extraction can be beneficial for certain dimensions.

Table 2: *Performance comparison between the proposed PCM variants (PCM-le-noNorm, PCM-le-norm, and PCM-cnn models) and existing models on the MSP-Podcast, IEMOCAP, and Korean MVD datasets. BM = Base Model; PCM-le-noNorm = PCM with Linear Embedding without z-score Normalization; PCM-le-norm = PCM with Linear Embedding and z-score Normalization; PCM-cnn = PCM with CNN-based Embedding.*

MSP-Podcast ver. 1.11 dataset					**IEMOCAP dataset**				
Model	$\mathrm{CCC_V}$	$\mathrm{CCC_A}$	$\mathrm{CCC_D}$	$\mathrm{CCC_{avg}}$	Model	$\mathrm{CCC_V}$	$\mathrm{CCC_A}$	$\mathrm{CCC_D}$	$\mathrm{CCC_{avg}}$
Goncalves.a [23]	0.607	0.567	0.424	0.533	Atmaja [26]	0.446	0.594	0.485	0.508
Goncalves.b [23]	0.603	0.582	0.450	0.530	Srinivasan.a [27]	0.582	0.667	0.545	0.598
Goncalves.c [23]	0.583	0.585	0.419	0.529	Srinivasan.b [27]	0.481	0.645	0.528	0.551
BM	0.575	0.581	0.506	0.564	BM	0.635	0.673	0.512	0.610
PCM-le-noNorm	**0.627**	0.581	0.506	**0.571**	PCM-le-noNorm	0.640	0.740	**0.557**	**0.646**
PCM-le-norm	0.561	0.588	0.507	0.552	PCM-le-norm	**0.646**	**0.744**	0.546	0.645
PCM-cnn	0.512	**0.621**	**0.536**	0.556	PCM-cnn	0.563	0.709	0.507	0.593

Korean MVD dataset (using the English-pretrained model)				**Korean MVD dataset (using the Korean-pretrained model)**			
Model	$\mathrm{CCC_V}$	$\mathrm{CCC_A}$	$\mathrm{CCC_{avg}}$	Model	$\mathrm{CCC_V}$	$\mathrm{CCC_A}$	$\mathrm{CCC_{avg}}$
BM	0.508	0.499	0.504	BM	0.562	0.477	0.520
PCM-le-noNorm	0.538	0.501	0.535	PCM-le-noNorm	0.579	0.503	0.541
PCM-le-norm	0.560	0.509	0.520	PCM-le-norm	**0.594**	**0.512**	**0.553**
PCM-cnn	**0.591**	**0.518**	**0.555**	PCM-cnn	0.569	0.501	0.535

However, the PCM-le-norm model, which we primarily investigated, showed relatively lower performance on MSP-Podcast. This phenomenon can be attributed to the dataset's characteristics. Table 1 shows that the Difference in Pitch Standard Deviation (SD) between High and Low Valence Scores (SD diff. V). MSP-Podcast (SD: 4.56) is significantly smaller than those in IEMOCAP (SD: 16.74) and MVD (SD: 16.01). Since MSP-Podcast exhibits minimal pitch variability in V, applying z-score normalization may have suppressed crucial information, making the raw pitch values in PCM-le-noNorm more effective.

For IEMOCAP, PCM with linear embedding (PCM-le-noNorm, PCM-le-norm) achieved the highest scores across all emotion dimensions, with $\mathrm{CCC_V}$ (0.646), $\mathrm{CCC_A}$ (0.744), and $\mathrm{CCC_D}$ (0.557) outperforming prior studies and BM. These results suggest that linear embedding effectively retains pitch contour patterns, leading to superior SER performance. Interestingly, z-score normalization had little impact on performance in IEMOCAP, as PCM-le-noNorm and PCM-le-norm exhibited nearly identical $\mathrm{CCC_{avg}}$ scores (0.646 and 0.645). This suggests that, unlike MSP-Podcast, IEMOCAP provides sufficient pitch variability across emotional states, reducing the necessity for normalization. Meanwhile, PCM-cnn consistently lagged behind linear embedding models, suggesting the notion that CNN's localized feature extraction may lead to a loss of fine-grained intonation cues necessary for V-A-D prediction.

For Korean MVD, results varied depending on the pretrained model. When using the English-pretrained model, PCM-cnn achieved the highest performance ($\mathrm{CCC_{avg}}$: 0.555). In contrast, with the Korean-pretrained model, PCM-le-norm outperformed all other variants ($\mathrm{CCC_{avg}}$: 0.553). This pattern suggests that the matching between the pretraining and finetuning languages influence the effectiveness of pitch embedding strategies. When the pretrained model was trained in a different language (English) from the finetuning data (Korean), CNN-based embedding performed better, possibly because CNN extracts localized pitch variations, which might be more robust to cross-linguistic differences. A similar trend was reported by Thanh et al. [8], where CNN-based pitch encoding improved SER performance in tonal languages (Chinese, Thai, and Vietnamese) when finetuning an English-pretrained model. This suggests CNN-based embeddings may be beneficial in mismatched language settings. On the other hand, when both pretraining and finetuning languages matched (Korean-Korean), linear embedding exhibited higher performance, likely due to its ability to capture broader pitch contour patterns that are more meaningful in a familiar linguistic environment. One possible explanation for this trend is inspired by language acquisition theories. Non-native speakers often struggle to perceive and reproduce global pitch patterns accurately, whereas native speakers naturally recognize and utilize broader intonation structures [28]. In a similar vein, when processing an unfamiliar language, the model may have faced challenges in capturing the overall pitch contour, leading it to rely more on localized pitch variations, which may serve as perceptual anchors when global pitch patterns are unfamiliar. On the other hand, in a familiar linguistic environment, where global pitch structures are more reliably extracted, linear embeddings that summarize the pitch contour holistically may have been more beneficial, aligning with the way native speakers process intonation. These findings highlight the need to consider both language matching and pitch embedding in SER. Future work could further explore pitch perception in cross-linguistic emotion recognition.

6. Conclusion

This study proposed the Pitch Contour Model (PCM), which integrates pitch features with Transformer-based speech representations through Cross-Attention. PCM effectively models pitch contour patterns, improving dimensional SER. Experimental results showed that PCM outperformed the Base Model (BM) and prior studies, emphasizing the importance of pitch contour modeling. PCM-le-noNorm achieved the highest $\mathrm{CCC_{avg}}$ on MSP-Podcast, while PCM-le-norm performed best in IEMOCAP and with a Korean-pretrained model for MVD. The effectiveness of PCM-le-noNorm in lower pitch variability conditions suggests normalization may suppress relevant emotional cues.

Pretraining-finetuning language matching influenced the optimal pitch embedding strategy. CNN-based embedding was more effective with an English-pretrained model, whereas linear embedding performed best with a Korean-pretrained model, reflecting potential differences in pitch perception between native and non-native speakers. Despite these findings, this study is limited to English and Korean datasets. Future research should explore PCM's applicability in multilingual settings to better understand pitch-based representations across languages.

7. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No.2022-0-00608, RS-2022-II220608] and IITP grant funded by the MSIT [NO.2021-0-01343]

8. References

[1] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977. [Online]. Available: https://www.sciencedirect.com/science/article/pii/009265667790037X

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[4] W. de Vries, A. van Cranenburgh, and M. Nissim, "What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4339–4350. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.389

[5] M. Yang, R. C. M. C. Shekar, O. Kang, and J. H. L. Hansen, "What can an accent identifier learn? probing phonetic and prosodic information in a wav2vec2-based accent identification model," in *Interspeech 2023*, 2023, pp. 1923–1927.

[6] D. de Oliveira, N. R. Prabhu, and T. Gerkmann, "Leveraging semantic information for efficient self-supervised emotion recognition with audio-textual distilled models," in *Interspeech*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258967953

[7] E. Rodero, "Intonation and emotion: influence of pitch levels and contour type on creating emotions," *Journal of voice*, vol. 25, no. 1, pp. e25–e34, 2011.

[8] P. V. Thanh, N. T. T. Huyen, P. N. Quan, and N. T. T. Trang, "A robust pitch-fusion model for speech emotion recognition in tonal languages," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 386–12 390.

[9] 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, pp. 335–359, 2008.

[10] 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.

[11] "Multimodal video datasets," https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=58.

[12] 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)*, 2020, pp. 6474–6478.

[13] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5200–5204.

[14] S. Kwon *et al.*, "Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach," *Expert Systems with Applications*, vol. 167, p. 114177, 2021.

[15] S. Latif, R. K. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Interspeech*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:102350542

[16] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.

[17] Y. Gao, C. Chu, and T. Kawahara, "Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with asr and gender pretraining," in *Proc. Interspeech*, 2023, pp. 3637–3641.

[18] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6922–6926.

[19] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Interspeech 2021*, 2021, pp. 3400–3404.

[20] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning." in *Interspeech*, vol. 2021. Brno, 2021, pp. 4508–4512.

[21] W.-N. Hsu, A. Sriram *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233004449

[22] Kresnik, "wav2vec2-large-xlsr-korean [pretrained model]," 2025, accessed: 2025-02-05. [Online]. Available: https://huggingface.co/kresnik/wav2vec2-large-xlsr-korean

[23] L. Goncalves, A. N. Salman, A. R. Naini, L. M. Velazquez, T. Thebaud, L. P. Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey 2024-speech emotion recognition challenge: Dataset, baseline framework, and results," *Development*, vol. 10, no. 9,290, pp. 4–54, 2024.

[24] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[25] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[26] B. T. Atmaja, A. Sasou, and M. Akagi, "Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion," *Speech Communication*, vol. 140, pp. 11–28, 2022.

[27] S. Srinivasan, Z. Huang, and K. 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)*. IEEE, 2022, pp. 6442–6446.

[28] A. Lengeris, *Prosody and Second Language Teaching: Lessons from L2 Speech Perception and Production Research*, 02 2012, vol. 15, pp. 25–40.