What Does Whisper Learn About Speech Emotion Recognition? Insights from Text-Independent Evaluation and Emotionally Incongruent Speech

Ryoga Kurosawa[1], *Bagus Tris Atmaja*[1], *Akira Sasou*[2], *Sakriani Sakti*[1]

[1] Nara Institute of Science and Technology, Japan
[2] National Institute of Advanced Industrial Science and Technology, Japan

kurosawa.ryoga.ku7@naist.ac.jp, bagus.tris@naist.ac.jp, a-sasou@aist.go.jp, ssakti@is.naist.jp

Abstract

Speech emotion recognition (SER) relies on both acoustic and linguistic information to infer a speaker's emotional state. Although Whisper has recently achieved promising performance for SER, it remains unclear how it exploits these complementary cues and generalizes to unseen utterances. In this paper, we investigate a Whisper-based SER model with a frozen Whisper encoder and a lightweight CNN classifier. We evaluate the model under speaker-independent (SI) and text-independent (TI) protocols, analyze representations across encoder layers, and examine the roles of lexical and acoustic information using emotionally incongruent synthesized speech. Experimental results show strong performance under the challenging TI setting, with the final Whisper encoder layer providing the most effective representation. The synthesized speech experiments further show that the model relies strongly on lexical information when lexical and acoustic emotional cues differ. These findings clarify the strengths and limitations of Whisper-based SER.

Index Terms: speech emotion recognition, Whisper, layer evaluation, linguistic bias

1. Introduction

In recent years, speech emotion recognition (SER) has increasingly relied on pretrained speech foundation models, particularly self-supervised learning (SSL) models such as HuBERT [1] and wav2vec2.0 [2]. These models learn general-purpose speech representations from large-scale unlabeled speech corpora, enabling effective transfer to downstream tasks despite the limited availability of emotion-labeled data. Previous studies have demonstrated that pretrained speech representations substantially improve SER performance [3, 4, 5]. Their effectiveness has also been shown to generalize across languages, making pretrained speech models attractive foundation models for multilingual SER [6]. More recently, Whisper [7], which is pretrained on large-scale paired speech and text data, has achieved competitive SER performance [8, 9, 10]. SER relies on both linguistic and paralinguistic information to infer a speaker's emotional state [11, 12]. Consequently, pretrained speech models capable of representing both types of information have become increasingly attractive for SER.

Despite its strong performance, however, the behavior of Whisper-based SER models remains insufficiently understood. In particular, it remains unclear how well Whisper generalizes to utterance texts that are not observed during training, which encoder layers provide the most informative representations for emotion recognition, and to what extent its predictions depend on linguistic rather than acoustic information. These questions are especially important because Whisper is optimized for automatic speech recognition (ASR) and may therefore encode lexical information more explicitly than conventional SSL speech models.

Previous studies have investigated these issues separately. Atmaja and Sasou [13] and Pepino et al. [14] showed that SER performance degrades under the text-independent (TI) protocol, in which utterance texts are separated across the training and test sets. Layer-wise studies have examined how emotion-related information is distributed across pretrained speech encoders and have reported that the final Whisper encoder layer can be particularly effective for SER [15, 9]. Other studies have evaluated pretrained SER models under acoustic-semantic conflicts and demonstrated that performance decreases when the emotion conveyed by the speaking style conflicts with the lexical meaning of the utterance [16]. However, these aspects have largely been examined independently. Consequently, no previous study has systematically connected Whisper's generalization to unseen utterances, its layer-wise representation characteristics, and its reliance on acoustic and linguistic emotion cues to provide a unified understanding of how Whisper performs SER.

To address these gaps, we investigate Whisper-based SER from three complementary perspectives. First, we evaluate a model consisting of a frozen Whisper encoder and a lightweight CNN classifier under both speaker-independent (SI) and TI protocols to examine its generalization to unseen speakers and utterance texts. Second, we conduct a layer-wise analysis of the Whisper encoder using both CNN- and MLP-based classifiers to identify which encoder representations are most informative for SER and whether the observed characteristics are consistent across evaluation protocols and classifier architectures. Third, we evaluate the model using synthesized speech with systematically manipulated lexical and acoustic emotional cues to analyze their relative influence on model predictions.

The main contributions of this paper are as follows:

- We demonstrate that a Whisper-based SER model substantially mitigates the performance degradation under the TI evaluation.
- We conduct a layer-wise analysis of the Whisper encoder and show that the final encoder layer provides the most effective representation for SER.
- We evaluate the model using synthesized speech with systematically manipulated lexical and acoustic emotional cues to examine their relative influence on model predictions.

2. Related Work

2.1. Text-independent SER

Atmaja and Sasou [13] first introduced the TI evaluation protocol for speech emotion recognition (SER), demonstrating that

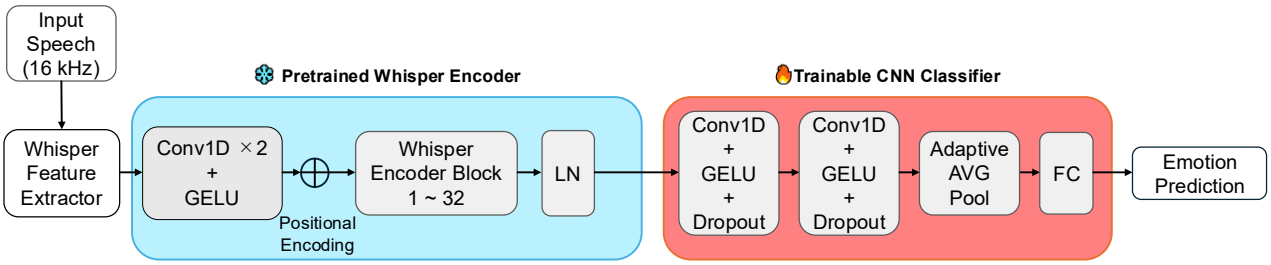


Figure 1: *Whisper-based SER model architecture. LN denotes layer normalization, and FC denotes a fully connected layer.*

performance degrades substantially when the training and test sets contain different utterance texts. The conventional SI protocol evaluates a model's ability to generalize to unseen speakers while allowing the same utterance texts to appear in both the training and test sets. In contrast, the TI protocol ensures that no utterance text appears in both the training and test sets, preventing models from exploiting sentence-specific lexical information and providing a more rigorous evaluation of generalization to unseen linguistic content. This evaluation is particularly important for comparing SER models that rely on different sources of information, as it better reflects their robustness to unseen utterances.

In a subsequent study, Atmaja and Sasou [17] evaluated self-supervised speech models, including HuBERT and WavLM [18], under the TI protocol. Although these pretrained models substantially improved SER performance over conventional approaches, a considerable performance gap between the SI and TI protocols remained, indicating that robust generalization to unseen utterance texts is still a challenging problem. However, while Whisper has recently emerged as a promising pretrained model for SER, its ability to generalize under the TI protocol has not yet been systematically investigated.

2.2. Whisper-based SER models

Although SSL speech representations implicitly encode linguistic information, numerous studies have shown that incorporating explicit textual representations further improves SER performance [19, 20, 21]. These findings suggest that lexical information is not fully captured by SSL speech embeddings alone.

With the recent success of Whisper, several studies have applied it to SER. Fukuda et al. [8] leveraged the Whisper decoder to incorporate linguistic information into SER through a trainable cross-attention module. Shendabadi et al. [9] conducted a layer-wise analysis of the Whisper Tiny and Small encoders to identify which encoder layers are most effective for SER. Feng and Narayanan [10] proposed a parameter-efficient fine-tuning strategy for both SSL- and Whisper-based SER models, demonstrating that high SER performance can be achieved while updating only a small subset of model parameters.

However, most Whisper-based SER studies have focused on improving recognition accuracy rather than understanding the characteristics of Whisper representations. Consequently, Whisper's ability to generalize under the TI evaluation protocol remains unexplored.

2.3. Layer-wise Analysis for SER

Recent studies have shown that different layers of pretrained speech models encode different types of information and contribute differently to downstream tasks such as SER. Li et al. [22] conducted a layer-wise analysis of wav2vec 2.0 by training an SER classifier on the output of each encoder layer. Their results showed that intermediate-layer representations achieved higher SER accuracy than those from the other layers. Furthermore, Pasad et al. [23] demonstrated that different layers of SSL speech models encode distinct types of information, with lower layers primarily capturing acoustic properties and higher layers representing more linguistic information.

More recently, D'Asaro et al. [15] investigated Whisper as a feature extractor for multilingual SER and reported that representations from the final Transformer layer consistently yielded the best performance. Similarly, Shendabadi et al. [9] analyzed the layer-wise behavior of Whisper for SER and found that intermediate and lower encoder layers required more training epochs to converge than the final layer. However, it remains unclear whether these layer-wise characteristics are consistent under more challenging evaluation settings, particularly the TI protocol. To address this question, we conduct a layer-wise analysis of Whisper under both SI and TI evaluation protocols.

2.4. Emotionally Incongruent Speech

Acoustic-semantic conflicts are common in everyday communication, where the emotional information conveyed by speech does not necessarily match the lexical meaning of the utterance. Huang et al. [16] evaluated several pretrained SER models, including WavLM, HuBERT, wav2vec 2.0, and Whisper, using emotionally incongruent speech and reported substantial performance degradation under such conditions. Their findings suggest that pretrained SER models are vulnerable when acoustic and linguistic emotional cues conflict. However, their analysis primarily focused on classification performance, leaving the underlying behavior of Whisper largely unexplored.

In particular, it remains unclear how Whisper processes conflicting acoustic and linguistic information, whether this behavior is related to its ability to generalize under the TI protocol, and how representations from different encoder layers contribute to its predictions in such scenarios. A systematic understanding that connects Whisper's robustness to acoustic-semantic conflicts with its layer-wise representations and generalization ability is still lacking. To address these gaps, this work provides a unified analysis of Whisper through the TI evaluation, layer-wise representation analysis, and emotionally incongruent speech experiments.

3. Datasets and Model Description

3.1. Datasets

We used two Japanese emotional speech datasets to evaluate the proposed model. Japanese Twitter-based Emotional Speech (JTES) [24] is a large-scale read-speech corpus containing 20,000 utterances, totaling approximately 23.5 hours, recorded by 100 speakers (50 female and 50 male). It covers four emotion categories: anger, joy, sadness, and neutral. For each emotion, 50 emotion-dependent sentences were selected from texts collected from Twitter with consideration of phonetic and prosodic balance, and all speakers read the same set of 200 sentences.

Japanese emotional speech corpus with verbal content and nonverbal expressions (JVNV) [25] contains 1,615 utterances, totaling 3.94 hours, recorded by four speakers (two female and two male) across six emotion categories: anger, disgust, fear, happiness, sadness, and surprise. Throughout this paper, the *happiness* label in JVNV is treated as equivalent to the *joy* label in JTES. JVNV contains both verbal speech and non-verbal vocal expressions, such as laughter and sobbing. Each utterance includes at least one non-verbal expression, and the corpus provides both predefined and speaker-selected non-verbal expression conditions. Table 1 summarizes the statistics of the datasets.

Table 1: *Statistics of the emotional speech datasets. Emotions denotes the number of emotion classes.*

Dataset	Speakers	Utterances	Emotions	Duration
JTES [24]	100	20,000	4	23.5 h
JVNV [25]	4	1,615	6	3.94 h

3.2. Model Architecture

Figure 1 illustrates the proposed Whisper-based SER architecture. We used the encoder of Kotoba-Whisper-v2.0 [26] as a frozen feature extractor and optimized only the CNN classifier. Kotoba-Whisper-v2.0 is a distilled version of Whisper large-v3 [7, 27] trained on the large-scale Japanese speech corpus ReazonSpeech [28]. The encoder was kept frozen throughout training so that the experiments evaluate the intrinsic quality of the pretrained Whisper representations rather than representations adapted through task-specific fine-tuning. This also enables a fair comparison across encoder layers in the layer-wise analysis.

The input waveform was resampled to 16 kHz and converted by the Whisper feature extractor into a 128-channel log-Mel spectrogram. The spectrogram was processed by the Whisper encoder, which consists of two convolutional layers followed by 32 Transformer encoder blocks, to produce frame-level speech representations.

The encoder representations were fed into a trainable two-layer one-dimensional CNN classifier. Previous studies on pretrained speech representations have predominantly employed linear projection or pooling layers as downstream classifiers, while the design of the downstream network itself has received relatively little attention [29]. We therefore adopted a lightweight one-dimensional CNN, which has also been used as a downstream classifier for SER, to capture local temporal patterns with only a small number of trainable parameters. The CNN outputs were aggregated by adaptive average pooling and projected to the emotion classes using a fully connected layer. This lightweight architecture allows the SER performance to primarily reflect the quality of the frozen Whisper representations, making it suitable for the layer-wise analysis conducted in this study.

4. Experiments

4.1. Experimental Settings

We evaluated the proposed model under speaker-independent (SI) and text-independent (TI) protocols. In the SI protocol, speakers were mutually exclusive across the training, validation, and test sets, whereas sentence texts could overlap. In the TI protocol, sentence texts were mutually exclusive across the three splits, whereas speakers could appear in multiple splits.

For JTES, both SI and TI partitions followed a 3:1:1 train/validation/test ratio. Under the SI protocol, 60 speakers (30 female and 30 male) were used for training, 20 speakers (10 female and 10 male) for validation, and the remaining 20 speakers for testing. Under the TI protocol, the 50 sentence texts associated with each emotion category were divided into 30, 10, and 10 texts for training, validation, and testing, respectively. All utterances corresponding to a given sentence text were assigned to the same split, ensuring that no sentence appeared in multiple partitions.

For JVNV, the TI protocol was constructed separately for regular-session and free-session recordings. For the regular-session data, only utterance IDs shared by all speakers within each emotion category were used, and the split was performed at the (emotion, utterance ID) level so that the same sentence text never appeared in multiple partitions. For the free-session recordings, all utterances were included, and each speaker-emotion subset was divided into training, validation, and test sets with a 6:2:2 ratio.

For the JVNV SI protocol, speakers F1 and M1 were used for training. To maintain balanced gender distributions in the validation and test sets, utterances from speakers F2 and M2 were divided equally between the validation and test sets within each emotion category.

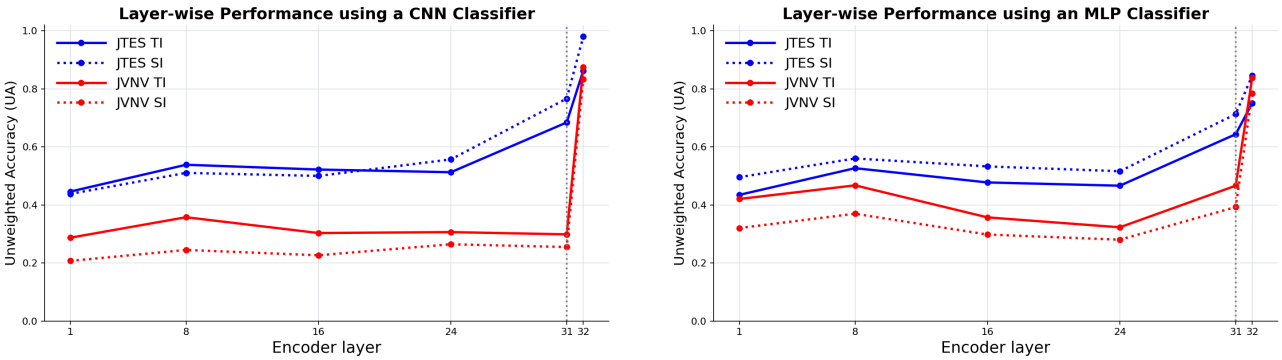
Table 2 summarizes the training hyperparameters. The hyperparameters were selected based on the validation performance in preliminary experiments. Because JTES and JVNV differ substantially in the number of speakers, dataset size, and acoustic characteristics, the hyperparameters were tuned independently for each dataset.

Table 2: *Training configurations for each dataset.*

Hyperparameter	JTES	JVNV
Effective batch size	16	16
Learning rate	1.0×10^{-4}	2.0×10^{-4}
Weight decay	0.01	0.01
Epochs	5	40
Warmup ratio	0.05	0.05
LR scheduler	Cosine	Cosine

4.2. Which Whisper Encoder Layers Best Represent Emotion?

To understand which encoder representations are most informative for speech emotion recognition, we conduct a layer-wise



(a) *CNN* (b) *MLP*

Figure 2: *Layer-wise UA obtained with (a) a two-layer CNN classifier and (b) a two-layer MLP classifier.*

analysis of Whisper representations. UA was used as the evaluation metric to enable direct comparison with previous studies. Due to computational constraints, we selected representative layers from the early, intermediate, and final stages of the encoder. Since the Kotoba-Whisper-v2.0 encoder contains 32 Transformer layers, we selected six representative layers: Layers 1, 8, 16, 24, 31, and 32. To ensure a fair comparison, all selected layers were evaluated using the same CNN classifier as the proposed model. Furthermore, we also evaluated a two-layer MLP classifier consisting of two fully connected layers with GELU activations followed by a linear output layer to examine whether the observed layer-wise characteristics are consistent across different classifier architectures. This classifier was trained using the same hyperparameters as the CNN classifier. All experiments were conducted on both the JTES and JVNV datasets under both SI and TI evaluation protocols.

4.3. Does Whisper Rely More on Lexical or Acoustic Information?

To investigate whether Whisper relies primarily on lexical or acoustic information, we evaluate the model using synthesized speech in which lexical and acoustic emotional cues are systematically manipulated. When lexical and acoustic emotional cues conflict, both SSL- and Whisper-based SER models have been reported to exhibit degraded performance [16]. Inspired by the experimental setting of Corrêa et al. [30], we constructed both congruent and incongruent synthesized speech conditions to examine the relative influence of lexical content and speaking style on model predictions. Because the JVNV dataset does not contain neutral speech, this experiment was conducted only on the JTES dataset.

To avoid lexical overlap with the training and evaluation data, all evaluation texts were newly generated using GPT-5.5 [31] and did not appear in the JTES corpus. Specifically, GPT-5.5 was prompted to generate 15 sentences for each of the four emotion categories (anger, joy, sadness, and neutral), resulting in 60 unique text prompts.

To verify that the observed behavior was not specific to a particular TTS model, the texts were synthesized using two state-of-the-art TTS models, CosyVoice3 [32] and Qwen3-TTS [33]. Speech was synthesized under three conditions: (1) emotional text with emotional speech, (2) emotional text with neutral speech, and (3) neutral text with emotional speech. Each condition also included 15 neutral-text/neutral-speech samples as a common baseline to enable consistent comparison across the three conditions. Therefore, none of the three conditions was entirely incongruent. In each condition, 15 of the 60 samples consisted of neutral text synthesized with neutral speech, while the remaining 45 samples followed the corresponding emotional text/speech configuration described above.

Emotional utterances excluded from the training sets under both the SI and TI protocols were selected from the JTES dataset as reference speech for TTS synthesis. Consequently, each condition consisted of 60 synthesized utterances for each TTS model, yielding a total of 360 evaluation samples.

For samples in which only one modality conveyed an emotional cue, the reference label was defined by that cue rather than by a fixed modality. Specifically, for emotional-text/neutral-speech samples, the text emotion was used as the reference label, whereas for neutral-text/emotional-speech samples, the speech emotion was used. This labeling scheme is used to quantify how the model responds when emotional information is conveyed primarily through either lexical content or speaking style, rather than to define one modality as the universally correct source of emotion.

5. Results and Discussion

5.1. SER Performance Under Different Data-Splitting Criteria

We report weighted accuracy (WA) and unweighted accuracy (UA) as evaluation metrics. Table 3 summarizes the SER performance under the speaker-independent (SI) and text-independent (TI) evaluation protocols.

On the JTES dataset, the proposed model achieved the highest performance under the SI protocol while substantially reducing the performance gap between the SI and TI protocols compared with previous SSL-based SER models. Although the proposed model did not achieve the highest accuracy under the TI protocol, the best TI performance reported by Nagase et al. [34] was obtained using a multimodal approach that integrates textual and acoustic information through a combination of early and late fusion. Nevertheless, the proposed model consistently outperformed previous SSL-based SER models under the TI protocol.

On JVNV, the proposed model achieved comparable performance under the SI and TI protocols, with WA differences of only 3.74% and UA differences of 4.44%. Unlike JTES, the TI protocol did not lead to a noticeable degradation in performance. This observation may be attributed to the characteris-

tics of the JVNV corpus rather than the proposed model itself. Specifically, JVNV contains only four speakers and includes diverse non-verbal vocal expressions, which may reduce the influence of lexical content and make emotion recognition rely more heavily on acoustic cues.

These results suggest that pretrained Whisper representations generalize well to unseen utterance texts. Such generalization is important for practical SER systems, which must process utterances that are not observed during training.

Table 3: *SER performance under speaker-independent (SI) and text-independent (TI) data-splitting criteria. Note that results for TI from [34] are based on the fusion of acoustic and linguistic features.*

		SI		TI	
Dataset	Model	WA (%)	UA (%)	WA (%)	UA (%)
JTES	MLP [13]	88.85	–	65.04	–
	SVC + Aug [35]	98.05	98.05	76.75	76.75
	CNN [34]	68.47	71.31	**90.70**	**91.22**
	HuBERT [17]	–	–	76.75	76.75
	WavLM [17]	–	–	78.25	78.25
	Ours	**98.22**	**98.22**	85.70	85.70
JVNV	Ours	84.11	83.33	87.85	87.77

5.2. Layer-wise Analysis

The results are presented in Figure 2. In contrast to previous findings on SSL-based SER models [22], the Whisper encoder consistently achieved the highest performance when features from the final encoder layer were used, regardless of the classifier architecture. This observation is consistent with the findings of Shendabadi et al. [9], who also reported that the final Whisper encoder layer achieved the best SER performance. In both the CNN- and MLP-based classifiers, performance improved toward deeper layers and peaked at the final layer. These findings suggest that, unlike conventional SSL models in which intermediate-layer representations are often more effective for SER, the final-layer representations of Whisper retain highly discriminative information for emotion recognition. This trend was consistently observed across all experimental settings, although the underlying reason for this behavior remains unclear.

These results suggest that deeper Whisper encoder representations may increasingly emphasize linguistic information while preserving acoustic cues that remain useful for SER. One possible explanation is that Whisper is optimized for automatic speech recognition. Therefore, its deeper encoder layers may progressively transform the input representation into representations that facilitate transcription by the decoder. However, the present experiments do not directly analyze layer-wise representations. Therefore, this interpretation should be regarded as a hypothesis.

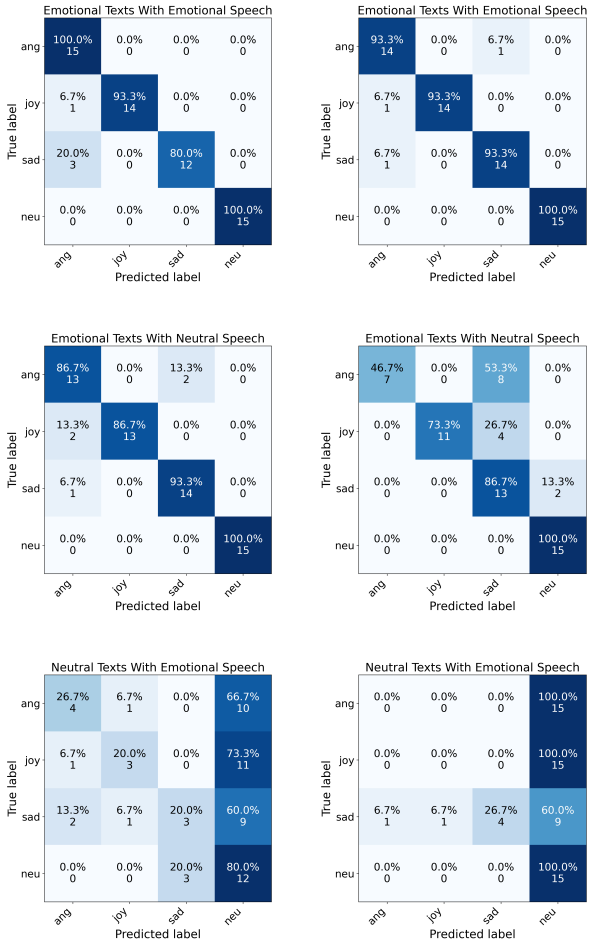
5.3. Emotionally Incongruent Speech Evaluation

The synthesized speech was evaluated using the proposed CNN-based SER models using the final Whisper encoder layer, trained on JTES under the SI and TI protocols. The evaluation results are summarized in Table 4.

The models maintained high accuracy when emotional text was synthesized with either emotional or neutral speech. In contrast, performance decreased substantially when neutral text was paired with an emotional speaking style. Figure 3 presents the confusion matrices obtained using speech synthesized by

Table 4: *Performance of JTES-trained SER models on synthesized speech generated by different TTS models.*

TTS model	Synthesis condition		WA (%)	
	Text emotion	Speech emotion	SI	TI
CosyVoice3	Emotional	Emotional	93.33	95.00
CosyVoice3	Emotional	Neutral	91.67	90.00
CosyVoice3	Neutral	Emotional	**36.67**	**41.67**
Qwen3-TTS	Emotional	Emotional	95.00	96.67
Qwen3-TTS	Emotional	Neutral	76.67	78.33
Qwen3-TTS	Neutral	Emotional	**31.67**	**33.33**



(a) *CosyVoice3* (b) *Qwen3-TTS*

Figure 3: *Confusion matrices for speech synthesized by (a) CosyVoice3 and (b) Qwen3-TTS under the speaker-independent (SI) evaluation protocol. Top: emotional-text/emotional-speech condition; middle: emotional-text/neutral-speech condition; bottom: neutral-text/emotional-speech condition. Each condition includes 15 neutral-text/neutral-speech samples as a common baseline.*

CosyVoice3 and Qwen3-TTS under the SI protocol, respectively. Despite the different TTS models, both figures exhibit a consistent tendency: among the 45 samples in which neutral text was paired with an emotional speaking style, the majority were classified as neutral. This tendency was consistently observed under both the SI and TI evaluation protocols, indicating that the Whisper-based SER model was strongly influenced by lexical cues.

These findings suggest that the predictions of the proposed model are predominantly influenced by lexical emotion cues when lexical and acoustic emotional information differ. As discussed in Section 5.2, one possible explanation is that Whisper's deeper encoder representations increasingly emphasize linguistic information while preserving acoustic cues useful for downstream tasks. However, because the present study does not directly analyze the information encoded at each encoder layer, this interpretation should be regarded as a hypothesis rather than a definitive conclusion.

Although the proposed model achieved promising results, several limitations remain. First, the mechanism underlying the observations discussed in Section 5.2 remains unclear. Therefore, the interpretation presented in this study should be regarded as a hypothesis rather than a definitive explanation. Second, the evaluation using synthesized speech with manipulated lexical and acoustic emotional cues was conducted under limited experimental conditions. The synthesized speech used for evaluation was limited in both quantity and diversity, and the evaluation was performed only on the JTES dataset. Consequently, it remains unclear whether the observed findings generalize to more diverse datasets and practical SER scenarios.

Overall, these findings suggest that robust SER requires the effective integration of both acoustic and linguistic information, rather than predominant reliance on either source alone.

6. Conclusions

In this work, we investigated Whisper-based speech emotion recognition (SER) under the challenging text-independent (TI) evaluation protocol. On the JTES dataset, the proposed model achieved the best SI performance (98.22% WA/UA) and competitive TI performance (85.70% WA/UA), substantially reducing the performance gap between the two evaluation protocols compared with previous SSL-based SER models. Furthermore, the proposed model also achieved competitive performance on the JVNV dataset under both SI and TI evaluation settings. These results demonstrate that competitive SER performance can be achieved using only a frozen Whisper encoder and a lightweight CNN classifier. Layer-wise analysis further revealed that, unlike previous findings on SSL-based speech models, the final Whisper encoder layer consistently provided the most effective representation for SER regardless of the classifier architecture. In addition, experiments using synthesized speech with systematically manipulated lexical and acoustic emotional cues showed that the proposed model's predictions were strongly influenced by lexical information when the two cues differed.

Future work will further investigate the internal representations learned across Whisper encoder layers to better understand the layer-wise behavior discussed in Section 5.2 and to clarify why the final encoder layer is particularly effective for SER. Based on these findings, we will explore methods that effectively exploit both linguistic and acoustic information, rather than relying predominantly on lexical cues, to improve the robustness of SER under the TI conditions. We will also evaluate the proposed approach on larger and more diverse emotionally incongruent speech datasets, as well as additional emotional speech corpora, to examine whether the observed findings generalize to more realistic SER scenarios.

7. Acknowledgments

This paper is based on results obtained from a project JSPS KAKENHI, Grant Numbers JP25H01139 and 24K02967, JST NEXUS, Grant Number JPMJNX25C1, and J-PEAKS, Grant Number JPJS00420240021.

8. Generative AI Use Disclosure

Generative AI tools were used only for minor language editing, such as grammar checking and wording refinement to improve clarity. All manuscript content and scientific work—including problem formulation, methodology development, experimental design and execution, and data analysis were conducted entirely by the authors. The authors reviewed the final manuscript and take full responsibility for its accuracy and integrity.

9. References

[1] 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.

[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] 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.

[4] P. Jafarzadeh, A. M. Rostami, and P. Choobdar, "Speaker emotion recognition: Leveraging self-supervised models for feature extraction using wav2vec2 and hubert," *arXiv preprint arXiv:2411.02964*, 2024.

[5] 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.

[6] B. T. Atmaja and A. Sasou, "Evaluating multilingual self-supervised learning models for multilingual speech emotion recognition," *Journal of Signal Processing*, vol. 30, no. 5, 2026.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 28 492–28 518.

[8] R. Fukuda, T. Kano, A. Ando, and A. Ogawa, "Speech emotion recognition based on large-scale automatic speech recognizer," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[9] A. Shendabadi, P. Izadirad, M. Salehi, and M. Bijankhan, "Speech emotion recognition leveraging openai's whisper representations and attentive pooling methods," *arXiv preprint arXiv:2602.06000*, 2026.

[10] T. Feng and S. Narayanan, "Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[11] 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.

[12] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[13] Atmaja, Bagus Tris and Sasou, Akira, "Effect of different splitting criteria on the performance of speech emotion recognition," in *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*. IEEE, 2021, pp. 760–764.

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

[15] F. D'asaro, J. J. M. Villacís, G. Rizzo, and A. Bottino, "Using large speech models for feature extraction in cross-lingual speech emotion recognition," in *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, 2024, pp. 258–265.

[16] D. Huang, Y. Lv, R. Xiong, C. Jin, and X. Peng, "When tone and words disagree: Towards robust speech emotion recognition under acoustic-semantic conflict," *arXiv preprint arXiv:2601.04564*, 2026.

[17] B. T. Atmaja and A. Sasou, "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124 396–124 407, 2022.

[18] 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.

[19] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition," in *Interspeech 2020*, 2020, pp. 3755–3759.

[20] 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.

[21] S. Dutta and S. Ganapathy, "Leveraging content and acoustic representations for speech emotion recognition," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[22] Y. Li, Y. Mohamied, P. Bell, and C. Lai, "Exploration of a self-supervised speech model: A study on emotional corpora," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 868–875.

[23] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[24] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," in *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2016, pp. 16–21.

[25] D. Xin, J. Jiang, S. Takamichi, Y. Saito, A. Aizawa, and H. Saruwatari, "Jvnv: A corpus of japanese emotional speech with verbal content and nonverbal expressions," *IEEE Access*, pp. 19 752–19 764, 2024.

[26] Kotoba Technologies, "kotoba-whisper-v2.0," https://huggingface.co/kotoba-tech/kotoba-whisper-v2.0, 2024, accessed: Jul. 23, 2026.

[27] S. Gandhi, P. Von Platen, and A. M. Rush, "Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling," *arXiv preprint arXiv:2311.00430*, 2023.

[28] Y. Yin, "Reazonspeech: A free and massive corpus for japanese asr," 2023.

[29] S. Shen, F. Liu, H. Wang, Y. Wang, and A. Zhou, "Temporal shift module with pretrained representations for speech emotion recognition," *Intelligent Computing*, vol. 3, p. 0073, 2024.

[30] P. Corrêa, J. Lima, V. Moreno, L. Ueda, and P. Costa, "Evaluating emotion recognition in spoken language models on emotionally incongruent speech," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 17 952–17 956.

[31] OpenAI, "Introducing GPT-5.5," https://openai.com/index/introducing-gpt-5-5/, 2026, accessed: 2026-07-29.

[32] Z. Du, C. Gao, Y. Wang, F. Yu, T. Zhao, H. Wang, X. Lv, H. Wang, C. Ni, X. Shi *et al.*, "Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training," *arXiv preprint arXiv:2505.17589*, 2025.

[33] H. Hu, X. Zhu, T. He, D. Guo, B. Zhang, X. Wang, Z. Guo, Z. Jiang, H. Hao, Z. Guo *et al.*, "Qwen3-tts technical report," *arXiv preprint arXiv:2601.15621*, 2026.

[34] R. Nagase, T. Fukumori, and Y. Yamashita, "Speech emotion recognition with fusion of acoustic-and linguistic-feature-based decisions," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 725–730.

[35] B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," *Sensors*, vol. 22, no. 16, p. 5941, 2022.