Targeted Representation with Information Disentanglement Encoding Networks in Tasks

Takumi NAGAWAKI* Keisuke IKEDA* Kohei CHIKE† Hiroyuki NAGANO† Masaki NOSE† Satoshi TAMURA*
* Gifu University, Gifu, Japan
E-mail: {nagawaki,ikeda,tamura}@asr.info.gifu-u.ac.jp
† RICOH, Kanagawa, Japan
E-mail: {kohei.chike,hiroyuki.nagano,masaki.nose}@jp.ricoh.com

***Abstract*—This paper proposes a novel model TRIDENT to obtain better embeddings for speaker, emotion and speech recognition. Conventional features including HuBERT are often chosen to represent speech features for a task, however, such the features still contain the other information irrelated to the task. Our hypothesis is that, if we can disentangle them, better embeddings and higher performance are expected. Our model is designed to obtain speaker-, emotion-, and context-related features exclusively, by enhancing HuBERT features. In addition, these embeddings can be used to reconstruct the original features. We carried out speaker verification, emotion recognition and speech recognition experiments using task-specific embeddings, respectively. The results demonstrated that our scheme could appropriately extract features for various speech processing tasks simultaneously, achieving enough accuracy.**

I. Introduction

Speech has a lot of information such as contents, individuality and emotion. In order to detect theae elements, speech technology has been investigated for decades. Automatic Speech Recognition (ASR) transcribes speech contents. Speaker Recognition (SR) recognizes who spoke given speech waveforms. Speech Emotion Recognition (SER) detects emotion information such as anger, sad and happy, from speech signals. To accomplish these techniques, several kinds of features and models have been proposed. For example, mel-frequency cepstral coefficients and hidden Markov models were used for ASR. Mel-spectral features were used in SR and SER, while machine learning methods such as support vector machine were adopted. Nowadays it is common to employ deep-learning models for the classification and recognition tasks. Deep learning is also utilized to extract more effective features for each purpose.

One of the state-of-the-art deep-learning techniques, Self-Supervised Learning (SSL), has attracted attentions of speech-processing researchers. SSL builds a model on a task, utilizing unlabelled data efficiently. Recent advances in SSL have shown great potential in capturing the diverse attributes from speech [1]. Among such the works, we focus on Hidden Unit BERT (HuBERT) [2]. HuBERT converts speech signals into time-series vectors as useful embeddings. HuBERT embeddings are information-rich, encapsulating phonetic, speaker, and prosodic features, making it a powerful tool for speech processing applications. HuBERT is often employed in many applications for example ASR or SER, however, it is not guaranteed that HuBERT embeddings are designated to have only the information on the target task; sometimes the embeddings encapsulate various attributes together. Therefore, extracting only task-specified features such as speaker identity, emotional, and contextual information is still challenging.

To address this issue, we propose a novel approach TRIDENT. Our model is designed to calculate three kinds of embeddings, each exclusively having speaker-, emotion-, and context-related information. We thus expect that we can obtain better representations and results in ASR, SR and SER respectively, by disentangling HuBERT outputs. In addition, the model has an ability to reconstruct the original HuBERT embeddings after once separating the features, so that we can improve the clarity and consistency of the extracted features much more. We consider that this approach not only enhances speech understanding but also has the potential to significantly improve applications in personalized speech synthesis, emotion recognition, and context-aware systems. In this paper, we introduce our TRIDENT model, and demonstrate its ability to disentangle and re-synthesize speech embeddings. Finally, we highlight its potential contributions to the field of speech processing.

The rest of this paper is organized as follows: Section II briefly introduces related works to our work. The methodology appears in Section III. To evaluate our scheme, experiments were conducted and details are described in Section IV. Finally Section V concludes this paper.

II. Related Work

A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units [2]

Hidden Unit BERT (HuBERT) is a self-supervised learning framework designed to capture diverse speech representations by learning hidden units from unlabeled speech data. HuBERT can learn from unlabeled audio data, which means that a large amount of raw audio data can be used effectively. Therefore, the model expected to understand fundamental structures and patterns of speech, making it applicable to various audio processing tasks. HuBERT clusters audio data to generate

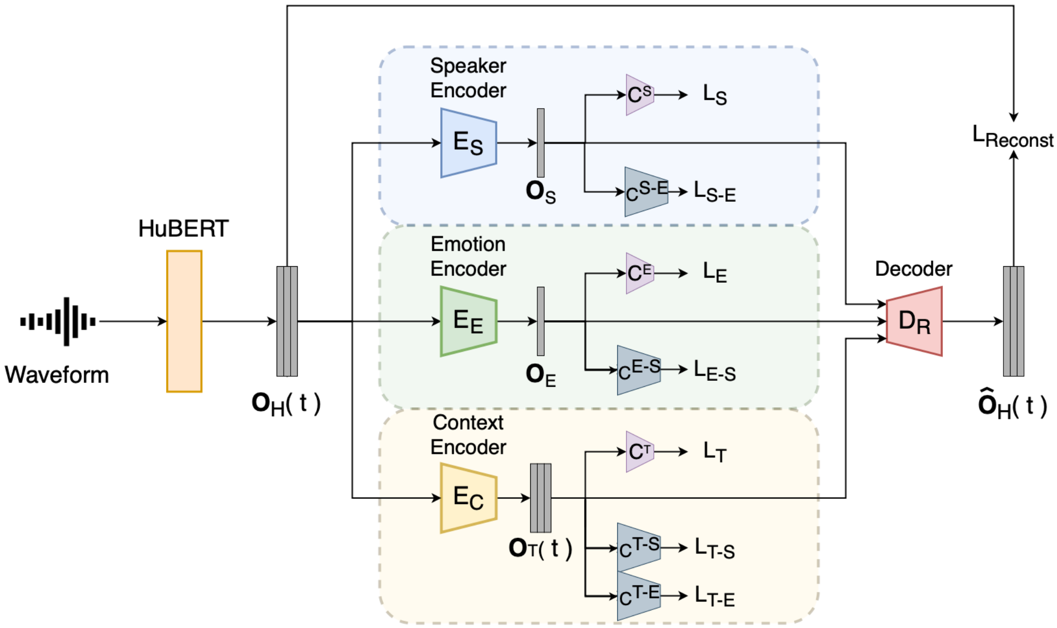

Fig. 1. Our model overview: The proposed method involves extracting embeddings using HuBERT, followed by three encoders: speaker encoder, emotion encoder, and context encoder. Each encoder is connected to classification and suppression models. These models help refine the embeddings by focusing only on the target task and suppressing irrelevant ones. The refined embeddings are then reconstructed by a decoder.

hidden units and sets up a task to predict these units, capturing the temporal features of speech data. Based on BERT, one of the transformer models, HuBERT utilizes bidirectional contextual information to deepen the understanding of audio data, resulting in high performance in many tasks e.g. speech recognition and generation.

B. Non-Parallel Sequence-to-Sequence Voice Conversion with Disentangled Linguistic and Speaker Representations [3]

Non-parallel sequence-to-sequence (seq2seq) voice conversion aims to transform the voice of one speaker into another one without requiring parallel training data. This approach involves disentangling linguistic content from speaker characteristics to achieve high-quality voice conversion. The proposed model consists of a text encoder, a recognition encoder, a speaker encoder, and a seq2seq decoder. The recognition encoder aligns acoustic features with phonetic sequences to produce linguistic representations of audio signals, while the speaker encoder generates embeddings that capture speaker-specific information. These components are integrated by the seq2seq decoder to reconstruct the acoustic features. Our method is inspired by this model structure.

III. Method

This section presents our proposed method: how to extract embeddings from raw speech waveforms leveraging the HuBERT model, how to obtain representations from three different encoders for speaker, emotion and context recognition tasks, and how to incorporate the vectors to reconstruct the original embeddings. Our proposed method, as illustrated in Fig.1, involves two main stages: feature-specific encoding using dedicated encoders, and embedding reconstruction using a decoder.

A. Feature Encoders

Initially, embeddings are extracted from the raw speech waveform using the HuBERT model. Let us denote the embeddings by $\mathbf{O}_H(t)$, where t is a time index. In this paper, the pretrained HuBERT model is expected. The HuBERT embeddings are then given to three independent encoders to obtain speaker, emotion, and context embeddings, described below.

*1) **Speaker Encoder**:* The speaker encoder E_S is designed to extract a feature vector $\mathbf{O}_S$ related to speaker identity:

$$\mathbf{O}_S = E_S\big(\mathbf{O}_H(t)\big) \tag{1}$$

The encoder is composed as follows: a self-attention layer, two fully connected layers, a Global Average Pooling (GAP) layer, and finally another fully connected layer. (Fig.2) Note that the representation $\mathbf{O}_S$ is obtained for one speech sequence.

Training Objective: To allow the encoded representations to have speaker information exclusively, an SR (identification) classifier C_S is used for model training. In addition, because the encoder is expected to exclude the emotion information, an SER suppression model C_{S-E} is also utilized; the suppression model is trained so that predicted results should be based on uniform distributions of which number corresponds to the number of emotions. This makes the speaker embeddings useless for SER. Consequently, we design a loss function of the speaker encoder according to the above discussion, defined as:

$$L_{\text{Speaker}} = L_S + \lambda \cdot L_{S-E} \tag{2}$$

where both the classifier C_S and the suppressor C_{S-E} consist of a single Linear layer,L_S is a cross-entropy loss for speaker identification, L_{S-E} is a cross-entropy loss for emotion suppression, and λ is a weighting factor.

2) ***Emotion Encoder****:* The emotion encoder E_E captures the emotional state of the speech. Its structure is the same as the speaker encoder. Similar to the former encoder, the emotion encoder also outputs a vector $\mathbf{O}_E$ as follows:

$$\mathbf{O}_E = E_E\big(\mathbf{O}_H(t)\big) \tag{3}$$

Training Objective: The basic idea of training loss function is also the same as the speaker encoder as:

$$L_{\text{Emotion}} = L_E + \lambda \cdot L_{E-S} \tag{4}$$

where both the classifier C_E and the suppressor C_{E-S} consist of a single Linear layer, L_E is a cross-entropy loss for emotion classification, L_{E-S} is a cross-entropy loss for speaker suppression C_{E-S}, and λ is a weighting factor.

3) ***Context Encoder****:* The context encoder E_T extracts contextual information from the given speech. The encoder is composed of a self-attention layer, followed by two fully connected layers. (Fig.2) The representation $\mathbf{O}_E(t)$ is obtained as:

$$\mathbf{O}_T(t) = E_T\big(\mathbf{O}_H(t)\big) \tag{5}$$

In this case, the vectors are generated frame by frame.

Training Objective: An ASR model C_T is adopted to observe a loss function L_T. We also leverage a speaker suppression classifier C_{T-S} and an emotion suppression classifier C_{T-E} to minimize irrelevant information. The training objective for the context encoder is then defined as:

$$L_{\text{Text}} = L_T + \lambda\ (L_{T-S} + L_{T-E}) \tag{6}$$

where both the classifier C_T and the suppressor C_{T-S}, C_{T-E} consist of a single Linear layer, L_T is a CTC loss with text labels, L_{T-S} is a cross-entropy loss for speaker suppression, and L_{T-E} is a cross-entropy loss for emotion suppression, and λ is a scaling factor.

B. Reconstruction Decoder

All the extracted features, which are speaker, emotion, and context embeddings $\mathbf{O}_S$, $\mathbf{O}_E$ and context $\mathbf{O}_T(t)$ respectively, are combined and fed into a reconstruction decoder D_R. The decoder reconstructs the original HuBERT embeddings as:

$$\hat{\mathbf{O}}_H(t) = D_R\big(\mathbf{O}_S, \mathbf{O}_E, \mathbf{O}_T(t)\big) \tag{7}$$

The purpose of this process is to preserve the integrity of the original speech attributes, while enhancing task-specific features. The simple reconstruction loss based on the mean squared error L_{Reconst} is chosen for training.

C. Training Strategy

The training strategy consists of two phases in each iteration. These two phases are learning opposing each other to effectively extract only the desired features.

Phase 1: In the first phase, the encoders (E_S, E_E, and E_T), decoder (D_R) and feature-based classifiers (C_S, C_E and C_T) are trained together, while the suppression classifiers are frozen. The total loss used in this phase includes all the loss values that we have already explained as:

$$L = L_{\text{Speaker}} + L_{\text{Emotion}} + L_{\text{Text}} + L_{\text{Reconst}} \tag{8}$$

Phase 2: Phase 2: In the second phase, only the suppression classifiers (C_{S-E}, C_{E-S}, C_{T-S} and C_{T-E}) are trained to minimize the influence of non-target features, for example, speaker information in emotion features. Specifically, C_{S-E} and C_{E-S} are trained using cross-entropy loss with the correct labels, while C_{T-S}, C_{T-E} employs CTC loss with the correct labels. This means that these models play an adversarial role. This adversarial training helps the encoders to learn more task-specific features by forcing the suppression classifiers to identify unwanted modalities.

IV. Experiment

A. Experimental Conditions

The primary objective of this experiment is to evaluate the performance of the proposed TRIDENT model and disentangled speaker, emotion, and context embeddings from HuBERT representations. The goals include assessing the accuracy of the learned embeddings and verifying the effectiveness of the reconstructed embeddings. Our hypothesis is that the TRIDENT model can effectively separate these attributes, resulting in improved performance in Automatic Speaker Verification (ASV), SER, and ASR tasks.

The model training utilized two datasets. First, the *LibriSpeech train-clean-360* dataset [4], which was a subset of the LibriSpeech dataset, consisting of 360 hours of ”clean” English speech, was used to provide a robust training base. Note that, since LibriSpeech is a dataset with no emotion labels assigned, we froze the emotion encoder when training models with LibriSpeech. Second, the *RAVDESS* (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset [5] was employed to enable the model to learn emotional expressions. To further enhance the training data, we randomly added noise to the speech data using the **MUSAN** dataset [6] for data augmentation, ensuring diverse and comprehensive training inputs.

The model in the upstream interface of S3PRL [7], [8] was used as the pre-trained HuBERT model. The model was trained using *LibriSpeech-960* and the model structure follows the HuBERT base model. The output from the final output layer is used as the HuBERT embedding. Regarding the hyperparameter, λ was set to 0.5.

B. Automatic Speaker Verification

ASV was carried out to evaluate the performance of speaker embeddings. ASV tries to verify the identity according to their voices. In this experiment, the *VoxCeleb1* [9] dataset was used for testing, which contains over 100,000 utterances from 1,251 celebrities, extracted from videos uploaded to YouTube. The

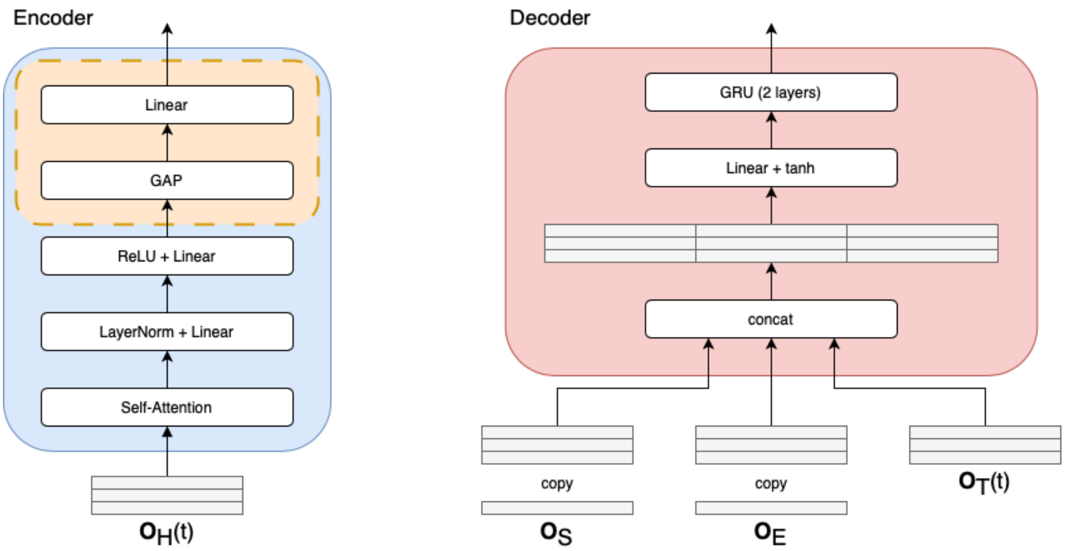

Fig. 2. The architecture of the encoder and decoder used in the proposed model. The encoders extract features from the HuBERT embeddings $\mathbf{O}_T(t)$. For the Speaker Encoder and Emotion Encoder, the components within the dotted region are utilized, including a GAP (Global Average Pooling) layer followed by a Linear layer. These components are omitted when using the Context Encoder. The extracted embeddings $\mathbf{O}_S$, $\mathbf{O}_E$, and $\mathbf{O}_T(t)$ are then passed to the decoder, where they are concatenated and processed through a Linear layer with a tanh activation, followed by a two-layer GRU (Gated Recurrent Unit), to reconstruct the original features

dataset is widely used for speaker verification tasks because of its diversity and real-world conditions.

In our evaluation, speaker embeddings were obtained and compared using the cosine similarity of speaker embeddings from pairs of audio samples. The Equal Error Rate (EER) was calculated to measure the accuracy of these comparisons. A HuBERT model from S3PRL, used as the base model, was compared by learning speaker classification and extracting embeddings with the same data as the proposed model. As a comparison of the accuracy of the reconstructed structure, the accuracy of the model excluding the reconstruction loss of the proposed method was also verified.

The results are summarized in Table I. The EER of our proposed TRIDENT model was 19.7%, slightly higher than the EER of 20.5% for the S3PRL HuBERT used as the base model and 20.1% for the model excluding reconstruction errors.

TABLE I
ASV PERFORMANCE COMPARISON

Model	**EER(%)**
HuBERT in S3PRL (Baseline)	20.5
Ours (Proposed)	19.7
Ours (w/o reconstruction)	20.1

C. Speech Emotion Recognition

SER was performed to evaluate the performance of emotion embeddings. In this experiment, the *CREMA-D* (Crowd-sourced Emotional Multimodal Actors Dataset) [10] was used, which contains 7,442 original clips from 91 actors. In these clips, actors between the ages of 20 and 74, representing a variety of races and ethnicities, express six different emotions: anger, disgust, fear, happy, neutral, and sad. Each sentence is presented with four different emotion levels: low, medium, high, and unspecified.

In our evaluation, emotion embeddings were obtained and used to train the emotion classifier. Weighted Average Recall (WAR) was a metric commonly used in SER tasks to evaluate the performance. Unlike the accuracy, which can be biased towards the majority class in imbalanced datasets, WAR calculates the average recall (sensitivity) across all classes, weighted by the class distribution. It is defined as the sum of the recall values for each class, weighted by the number of instances in each class, providing an evaluation metric that accounts for class imbalance. It takes values between 0 and 1, with higher values indicating more accurate performance.

The results are summarized in Table II. The WAR of our proposed TRIDENT model was 57.0%, which was lower than the WAR of 69.6% for the base model S3PRL HuBERT. The WAR of 59.7% was obtained in the model excluding reconstruction errors.

TABLE II
SER PERFORMANCE COMPARISON

Model	**WAR**(%)
HuBERT in S3PRL (Baseline)	69.6
Ours (Proposed)	57.0
Ours (w/o reconstruction)	59.7

D. Automatic Speech Recognition

Automatic Speech Recognition (ASR) was used to evaluate the performance of context representations in the TRIDENT model. In this experiment, the LibriSpeech datasets were chosen: the *LibriSpeech-train-clean-100* subset for training, and the *LibriSpeech-test-clean* subset for testing. The LibriSpeech corpus is a collection of approximately 1,000 hours of 16kHz read English speech derived from audiobooks, prepared by Vassil Panayotov with the assistance of Daniel Povey.

In the evaluation, the Word Error Rate (WER) was calculated to measure the accuracy of speech recognition. The proposed TRIDENT model was compared to the baseline S3PRL HuBERT model minus reconstruction errors.

The results are summarized in Table III. Our proposed TRIDENT model achieved a WER of 8.61%, exceeding the WER of 11.47% for HuBERT in the baseline S3PRL. Regarding the reconstruction errors. the proposed model with the error slightly outperformed the WER of the model excluding reconstruction errors.

TABLE III
ASR Performance Comparison

Model	WER (%)
HuBERT in S3PRL (Baseline)	11.47
Ours (Proposed)	8.61
Ours (w/o reconstruction)	8.93

E. Discussion

It is found that significant improvements were observed in ASV and ASR tasks, indicating that the TRIDENT model is particularly adept at separating and utilizing speaker-specific and contextual information. In the SER task, our proposed scheme achieved roughly 60% WAR; since the SER task is still a challenging task, the performance seems to be acceptable. In addition, the experimental results also suggest that the reconstructed structure helps to improve the consistency and accuracy of each embedding.

On the other hand, our method could not achieve the performance of the baseline model. It might be due to distortions in the RAVDESS dataset for model training and the CREMA-D dataset. We believe such the performance degradation sometimes occurs when using several emotion datasets for training, due to inconsistency of emotion labels or unbalanced data as a whole. To confirm the above discussion, visualization of the data which are not used for training in the RAVDESS dataset is conducted. Fig 3 shows the UMAP-based visualization results. The figure shows that the emotions in the RAVDESS dataset are appropriately separated in general.

According to the results, we re-trained the whole model using the CREMA-D dataset when performing SER. The results tell us that the WAR drastically increased to 73.1%. This significant improvement indicates the mismatch between two datasets.

To verify whether each representation was appropriately separated, we conducted supplemental experiments; we evaluated the accuracy of speaker and emotion embedding on another task, respectively. As a result, the WAR using speaker embedding was 48.8%. Regarding ASV, in this case we used classification accuracy: 77.9% for speaker embeddings, and 59.4% for emtion embeddings. It turned out that there was a significant decrease in accuracy if task and embedding were different. This indicates that our TRIDENT model can compute embeddings exclusively for each task, even if not perfectly.

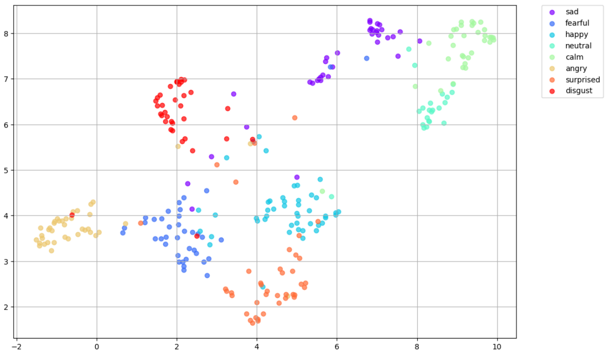

Fig. 3. Visualization of emotional embeddings in the RAVDESS dataset.

On the other hand, the simplicity of the neural network layers used in the three different encoders raises an important consideration. Although the training process utilized 360 hours of audio data, it remains uncertain whether such a compact model architecture can fully capture the complexity of the data distribution. While the current results are promising, it is essential to explore whether more sophisticated model architectures could better learn the intricate patterns present in the data. Addressing this limitation is a key direction for future work, where we plan to investigate the potential benefits of using more advanced models. This would ensure that the model's complexity is better aligned with the scale and diversity of the training data, potentially leading to further improvements in performance.

Finally, we would like to address the reconstruction. The reconstruction accuracy during testing achieved the MSE of 0.0068. This indicates that speaker-, emotion-, and contextual embeddings can still have significant information to regenerate corresponding HuBERT representations. We then carried out additional experiment. Assuming that a speaker vector is replaced into another one's, we checked the difference between the reconstructed vector and the true vector in the same emotion and context; The results were not good unfortunately, suggesting that our embeddings still need to be improved.

V. Conclusion

This paper presents our new model TRIDENT, which is designed to disentangle and re-synthesize speaker, emotion, and context information obtained from HuBERT embeddings. Experimental results demonstrated that TRIDENT improved performance in automatic speaker verification and automatic speech recognition tasks, and achieved acceptable results in emotion recognition. The reconstruction accuracy and classification performance in the tasks without the reconstruction function indicate the effectiveness and enhancement.

Our future works will focus on optimizing the model architecture and training procedures, as well as exploring different speech representations. The ability of TRIDENT to refine and separate speech attributes will offer promising directions for advancements in speech processing applications.

References

[1] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, *et al.*, *Superb: Speech processing universal performance benchmark*, 2021. arXiv: 2105.01051 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2105.01051.

[2] 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.

[3] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2019.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[5] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, e0196391, 2018.

[6] D. Snyder, G. Chen, and D. Povey, *MUSAN: A Music, Speech, and Noise Corpus*, arXiv:1510.08484v1, 2015. eprint: 1510.08484.

[7] A. T. Liu, S.-W. Li, and H.-y. Lee, *Tera: Self-supervised learning of transformer encoder representation for speech*, 2020. arXiv: 2007.06028 [eess.AS].

[8] S.-w. Yang, H.-J. Chang, Z. Huang, *et al.*, "A large-scale evaluation of speech foundation models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[9] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[10] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.