Learning Rich Speech Representations with Acoustic-Semantic Factorization

Minxue Niu[1,2] Najmeh Sadoughi[1] Abhishek Yanamandra[1] Pichao Wang[1]
Zhu Liu[1] Vimal Bhat[1] S. Elizabeth Norred[1]
Amazon Prime Video[1] *University of Michigan*[2]
sandymn@umich.edu, {nnnourab, ayyanama, wpichao, zhuzliu, vimalb, liznor}@amazon.com

***Abstract*—Self-supervised pretraining has transformed speech representation learning, enabling models to generalize across various downstream tasks. However, empirical studies have highlighted two notable gaps. First, different speech tasks require varying levels of acoustic and semantic information, which are encoded at different layers within the model. This adds the extra complexity of layer selection on downstream tasks to reach optimal performance. Second, the entanglement of acoustic and semantic information can undermine model robustness, particularly in varied acoustic environments. To address these issues, we propose a two-branch multitask finetuning strategy that integrates Automatic Speech Recognition and transcript-aligned audio reconstruction, designed to preserve and disentangle semantic and acoustic information in a final layer of a pretrained model. Experiments with the pretrained Wav2Vec 2.0 model demonstrate that our approach surpasses ASR-only finetuning across multiple downstream tasks, and it significantly improves ASR robustness in acoustically varied (emotional) speech.**

***Index Terms*—Large Speech Models, Speech Representation, Factorization**

I. Introduction

Advancements in Large Speech Models through self-supervised pretraining have significantly improved the efficacy and generalizability of learned speech representations [1]. These pretrained speech models can produce general speech representations, which can be applied to a wide array of speech tasks, often approaching or even outperforming small models trained on individual tasks [2]. Recently, representations from prominent models such as Wav2Vec 2.0 [3] and HuBERT [4] have frequently been used as baselines when exploring new speech applications. However, Large Speech Models vary in their performance across tasks, and creating robust, re-usable representations that can generalize across many tasks remains an ongoing challenge [2, 5].

Various self-supervised tasks have been used for learning speech representations, including reconstructing the original audio signal (VQ-VAE [6]), contrastive next-token prediction (CPC [7], Wav2Vec [8]), masked token prediction (Wav2Vec 2.0 [3], HuBERT [4], WavLM [9]), among others. Though these pretraining tasks rely solely on the raw audio signal to construct the prediction target, the contextualized nature of these tasks enables models to capture higher-level semantic information from low-level acoustic patterns. This enables downstream adaptation for semantic tasks like Automatic Speech Recognition (ASR) [3]. To more directly inject semantic information into the model, subsequent finetuning on ASR is often deployed as a second phase of the pretraining (we refer to this phase as "ASR finetuning"), significantly improving the performance on semantic tasks [3, 4].

Speech understanding is inherently complex, as different tasks require different amounts of lexical and para-linguistic information. For instance, speaker recognition depends heavily on local acoustic features, enabling models to identify voices from brief audio snippets, while tasks like ASR necessitate higher-level abstractions focused on the semantic content. Previous work investigating the localization of acoustic and semantic information within the layers of speech models has shown that early layers in pretrained models typically capture more general acoustic characteristics, while semantic information emerges in the middle and later layers [10, 11]. The information encoded in the last layer is dominated by the last seen task. For instance, ASR finetuning encourages a better preservation of semantic information in the Wav2Vec 2.0 model, but it causes a significant loss of acoustic information in the final layer: finetuning Wav2Vec 2.0 with 100 hours and 960 hours of ASR data reduces acoustic information preservation by about 50% and 75%, respectively [11]. As a result, downstream applications often have to handle the extra complexity of model and layer selection for optimal performance [12], where the optimal layer is specific to individual model, task and dataset [10]. It remains unclear whether this acoustic-semantic trade-off is inherent to the tasks themselves, or can be mitigated through novel training techniques.

Moreover, semantic and acoustic information are entangled in the representations at each layer, which reduces model robustness when acoustic variations or text domain shift occur. For example, emotional speech often contains acoustic variations (e.g. change of pitch or tone) that can cause extra challenges for ASR systems [13]. Factorization techniques have been helpful in many speech tasks, where learning benefits from disentangling one or more information components. For instance, i-vectors and x-vectors are well-known methods for factorizing speaker representations [14, 15]. More applications include target speaker extraction [16], speech recognition [17], voice conversion [18], and even learning version-invariant music features [19], beyond speech. However, factorization has not been sufficiently explored in learning general speech representations.

To address the challenge of preserving both acoustic and

TABLE I: Summary of downstream evaluation tasks.

task	dataset	metric	#train/#test	lr	#steps
SID	Voxceleb1	ACC	149k/5k	0.01	50k
SER	IEMOCAP	CCC	6k/2k	0.01	5k
E-ASR	IEMOCAP	WER	/10k	/	/
KS	Speech Commands	ACC	51k/3k	0.001	50k

semantic information in speech representations, we propose a factorized (two-branch) multitask finetuning framework. This framework is built on top of the Wav2Vec 2.0 model but can be easily adapted for other speech representation models. Our approach simultaneously performs ASR and audio reconstruction, utilizing a novel design that factorizes semantic and acoustic information into two separate representations at the last layer of the speech model. Specifically, the semantic branch follows the traditional ASR finetuning approach, while the acoustic branch is tailored to capture the rich, non-semantic features of speech by reconstructing the audio signal based on the output of the semantic head, which we call "transcript-aligned audio reconstruction". By leveraging this two-branch architecture, we ensure that each branch specializes in different aspects of the speech signal, encouraging the separation and preservation of both types of information. Experiments demonstrate that our method benefits both semantic and acoustic tasks, surpassing both the base and ASR-finetuned model on three out of four downstream tasks. This indicates a good preservation of both semantic and acoustic information as well as enhanced robustness.

Together, our work highlights two key contributions: 1) we show the feasibility of preserving both semantic and acoustic information in the same layer of a speech model through multi-task training, and 2) we show the promise of factorization for more robust general speech representation learning, benefiting various downstream tasks.

II. Tasks and Datasets

A. Upstream Finetuning

Automatic Speech Recognition (ASR) The Librispeech dataset consists of read English speech from audiobooks, along with rich labels including transcripts and speaker information. The full version contains 960 hours of speech (and has been used in Wav2Vec 2.0 pretraining). We use its 100 hour train-clean split for finetuning our upstream model.

B. Downstream Evaluation

We evaluate our learned semantic and acoustic representations on several downstream tasks. Details of the downstream datasets and finetuning setups are summarized in Table I.

Speaker Identification (SID) SID aims to classify a given audio into one of many previously-seen speakers, and we use this task to evaluate the preservation of acoustic information in our representations. We use the VoxCeleb1 dataset [20], which contains over 100,000 utterances from 1,251 celebrities extracted from Youtube Videos. We follow the train-validation-test split provided in its official release. The performance is measured by Accuracy (ACC).

Speech Emotion Recognition (SER) Emotional Activation is a measure of emotion, ranging from calm to excited. It is highly linked to acoustic signals in speech [21]. Therefore, we use the Activation regression task to measure the preservation of acoustic information. We use the IEMOCAP dataset [22], which contains approximately 12 hours of speech data and emotion activation labels annotated on a 5-point scale. We use Concordance Correlation Coefficient (CCC) as the evaluation metric, which assesses the agreement between predicted and true levels. The dataset contains five recording sessions, and we follow previous work [23] to run cross validation and report CCC mean and standard deviation across sessions.

ASR for Emotion Speech (E-ASR) To evaluate the robustness of semantic representations in acoustically challenging situations, we compare the ASR performance on emotional speech using speech transcripts from IEMOCAP, using Word Error Rate (WER) as the metric. Note that the models are trained to perform ASR, so no extra downstream training is involved.

Keyword Spotting (KS) KS is a semantic-dominant speech task that involves detecting specific keywords or phrases within an audio. We use the Speech Commands dataset [24] v0.01, which has speech utterances containing 31 words such as "bird", "house". ACC is used as the evaluation metric.

III. Model

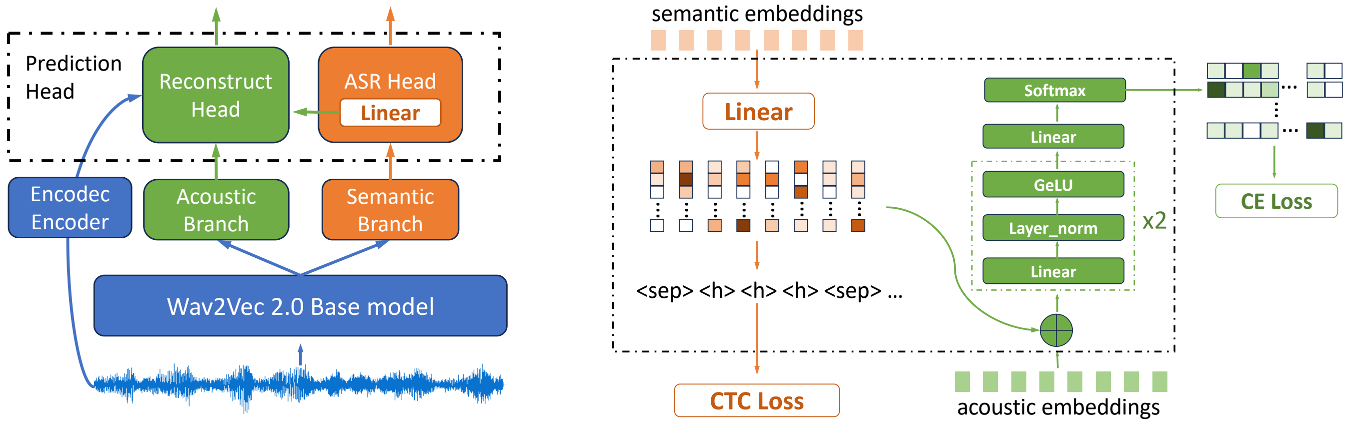
We propose a two-branch multitask framework to learn factorized acoustic and semantic representations of speech, illustrated in Figure 1a.

Semantic Branch and ASR Supervision The orange blocks show our semantic branch, which is also the baseline ASR-only finetuning approach as used in the Wav2Vec 2.0 model [3]. The output frame-level representations from the base model are fed into a linear layer with layer normalization to obtain the semantic representations. Then, another linear classification head is applied as a simple decoder to predict the logits. The ASR head is trained with a Connectionist Temporal Classification (CTC) Loss [25]. Given a batch of target sequences y and predicted sequences of logits x,

$$\mathcal{L}_{ASR} = \frac{1}{N}\sum_{i=1}^{N}\left(-\log\sum_{s\in\mathcal{S}(y_i)} p(s \mid x_i)\right) \tag{1}$$

$\mathcal{S}(y_i)$ is the set of all possible alignments of y_i with x_i. Note that pretrained Language Model decoders have been applied and can achieve better results on ASR [3]. However, since our goal here is to preserve more information in the semantic embedding rather than the decoder, we use a simple linear layer as the decoder.

Acoustic Branch and Reconstruction Supervision Similarly, we obtain the acoustic representations with a linear layer from the base model output. To supervise the learning of acoustic information, individual tasks may pay more attention



(a) Full model structure and modules.

(b) Details of the multitask factorization head.

Fig. 1: Model Structure.

to certain aspects like voice characteristics (Speaker Identification) or prosody (Emotion Recognition), but non-semantic audio components themselves are multi-faceted and difficult to define. In order to give a comprehensive representation of the acoustic information, we design the task of transcript-aligned audio reconstruction. As shown in Figure 1a, the acoustic embeddings (output of the acoustic branch) are concatenated with the output character logits from the ASR head, and are together used to reconstruct the audio signal through a reconstruction decoder. The idea is that the acoustic representations need to embed all information about the original speech other than the semantic information, which can be learned from the character logits. For the acoustic decoder, we use a stack of two linear layers with normalization and GeLU activation [26].

Reconstruction Target Another design choice of the reconstruction head is what to use as the reconstruct target. One straightforward approach is to reconstruct the original waveform or spectrogram [27]. However, doing this usually involves a much more complicated decoder to handle the sampling rate mismatch between the features and original signals, which adds to the model complexity and the difficulty of learning. Another natural approach is to continue the masked prediction pretraining task to recover local acoustic features, but that has the drawback that it is model-dependent: for example, Wav2Vec 2.0 is pretrained to reconstruct low-level local acoustic features learned by its CNN feature extractor [3], while HuBERT [4] uses more intermediate codebooks learned by two-phase clustering. If we simply keep the pretrain task as another head, the learned representations may not be consistent in terms of abstraction levels across models. Therefore, we chose Encodec [28], an off-the-shelf model-agnostic audio signal representation as our reconstruction target. Encodec is a neural audio codec model that compresses audio into sequences of discrete class tokens. We use its 6kbps model, which learns to quantize and represent each frame with tokens in eight codebooks, each of size 1024. Since Encodec provides trained encoder and decoder and has good reconstruction performance, being able to predict the Encodec tokens indicates sufficient information for reconstructing the audio.

Therefore, the reconstruction loss on each frame of sample i at frame t is a Cross Entropy Loss between the actual and predicted class, summed across C codebooks.

$$l_{i,t} = -\sum_{c=1}^{C}\sum_{k=1}^{K} y_{i,t,c,k} \log p(x_{i,t,c,k}) \tag{2}$$

The reconstruction loss weighs each frame equally and is averaged across sequence lengths T and batch size N:

$$\mathcal{L}_{Rec} = \frac{1}{N \cdot T}\sum_{i=1}^{N}\sum_{t=1}^{T} l_{i,t} \tag{3}$$

Together, we train the model with a weighted sum of the ASR loss and Reconstruction loss:

$$\mathcal{L} = \mathcal{L}_{ASR} + \lambda \mathcal{L}_{Rec} \tag{4}$$

IV. Experiments and Results

A. Experiment Setup

Training. We use the pretrained base Wav2Vec 2.0 checkpoint provided in the transformers library[1]. The acoustic and semantic branch each has hidden size of 768. We finetune the base model with our multitask factorization head with $\lambda = 1$ and an AdamW optimizer (learning rate = 1e-4) for 10,000 steps with a batch size of 128. We select the model with the lowest WER on the Librispeech dev-clean set.

Downstream evaluation. For the E-ASR task, we directly run model inference with the upstream model and decode the output logits from the semantic branch. We use the output from the acoustic branch for the SID and SER, and the output from the semantic branch for the semantic-focused task KS. For all utterance-level tasks (SID, SER, KS), we use a simple pooling layer followed by a linear classification head as the downstream mode, following previous approach [2]. The upstream model is frozen during evaluation. All representations

[1] https://huggingface.co/facebook/wav2vec2-base

TABLE II: Performance on downstream tasks. mtt.- multitask, fct. - factorization. The best performance across all models are highlighted in bold. Underscore indicates the better performance with or without factorization. The “#p_train” column shows the number of trainable parameters in upstream finetuning, while “#p_inference” indicates the number of parameters for inference, to obtain the representations (excluding task-specific head).

	#p_train	#p_inference	E-ASR (WER↓)	SER (CCC↑)	SID (ACC↑)	KS (ACC↑)
w2v2-base	/	94.4M	/	0.462±0.010	0.366	0.805
w2v2-100h	90.2M	94.4M	43.29%	0.406±0.009	0.134	**0.874**
ours, mtt.+fct.	114.0M	95.6M	**40.90%**	**0.500±0.011**	**0.387**	0.869
ours, mtt.-only	113.4M	95.0M	41.63%	0.485±0.006	0.335	0.852

TABLE III: Performance on upstream tasks. mtt. - multitask; fct. - factorization. '/' indicates that the model is not trained on and thus unable to perform the task.

	ASR (WER↓)		Reconstruction (ACC↑)	
	test-clean	test-other	test-clean	test-other
w2v2-100h	6.10%	13.3%	/	/
ours, mtt. + fct.	5.65%	13.90%	0.349	0.286
ours, asr-only	5.81%	14.33%	/	/
ours, rec-only	/	/	0.372	0.319
ours, mtt. only	5.68%	13.84%	0.348	0.285

we evaluate have the same dimension of 768. We tune the number of training steps and learning rate for each task to ensure convergence, and the optimal setups we report are detailed in Table I.

B. Downstream Results

We compare our model to the pretrain-only Wav2Vec2 model (w2v2-base) and the same model finetuned on 100 hour of speech (w2v2-100h) on downstream tasks (E-ASR, SER, SID and KS) in Table II. Note that the base model represents a stronger baseline for acoustic tasks while the 100h model is stronger at semantics [11]. As an ablation study, we also include a multitask-only model, where only one branch is trained with both heads.

First, the performance of baseline models (w2v2-base, w2v2-100h) verifies our assumption that **ASR-only finetuning does cause a loss of acoustic information**. Performance on acoustic tasks (SER and SID) both see a significant drop after finetuning with 100 hour of ASR data. On the other hand, the semantic task KS is improved by the finetuning, showing a trade-off between acoustic and semantic information.

Then, our model outperforms both baseline models on both acoustic tasks (SER ours 0.500 vs. base 0.462, SID ours 0.387 vs. base 0.366). On the downstream semantic task KS, our model approaches the better performance (ours 0.869, base 0.805, 100h 0.874). Those results indicate that **our model can successfully preserve both acoustic and semantic information** in the last layer of a speech model. What's more, **the factorization design provides an extra bump in the models performance**: it performs better across all four downstream tasks compared to multitask-only finetuning. Notably, on the challenging task of E-ASR, where semantic extraction can be biased by acoustic variations, using factorized representation further reduces WER from 41.63% to 40.90%, indicating enhanced robustness. Further, we note that our approach only adds a small computation overhead (1.3% more parameters) for downstream inference, compared to the baseline models.

C. Upstream Analysis

Although our goal is to learn a general representation that benefits various downstream tasks, we also analyze the upstream tasks' performance to better understand the model's behavior.

As shown in Table III, **our model can achieve reasonable reconstruction performance without hurting ASR performance**, compared to the model solely finetuned on ASR. On the clean speech test set (the same domain as training data), our model gets 5.65% WER– lower than the semantic-only baseline (5.81%) and the w2v2-100h model (6.10%). It has a slightly worse but comparable performance on the noisier test-other set, despite not being trained with data augmentation (ours 13.90%, wav2vec2 13.30%), outperforming the ASR-only model (14.33%). Comparing the multitask model with models trained solely with ASR or Reconstruction, we find that the two tasks have a small conflict with each other to achieve the best performance, but can both achieve reasonable performance through multitask training. Additionally, factorization doesn't make a significant difference on the upstream performance, compared to the multitask-only model. However, as we show in Section IV-B, this extra factorization constraint brings robustness to downstream tasks.

V. Conclusion

In this work, we study the challenge of learning factorized acoustic and semantic information in speech representations. We proposed a two-branch multitask finetuning framework that integrates ASR and transcript-aligned audio reconstruction to factorize and preserve these different types of information in the final layer of a pretrained model. Our results demonstrate superior performance and enhanced robustness across multiple downstream tasks. Furthermore, our analysis shows the feasibility of co-training ASR and reconstruction. We believe our experiments offer valuable insights for building more informative and robust speech models. While our work focuses on factorization at the final layer, earlier separation of the branches may offer further benefits. Future research will explore factorization across different layers.

References

[1] Abdelrahman Mohamed, Hung-Yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N Sainath, and Shinji Watanabe, "Self-supervised speech representation learning: A review," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1179–1210, Oct. 2022.

[2] Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-Tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-Yi Lee, "SUPERB: Speech processing universal PERformance benchmark," *arXiv [cs.CL]*, May 2021.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, vol. 33, pp. 12449–12460, 2020.

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

[5] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk, "HEAR: Holistic evaluation of audio representations," *arXiv [cs.SD]*, Mar. 2022.

[6] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv [cs.LG]*, July 2018.

[8] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv [cs.CL]*, Apr. 2019.

[9] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[10] Ankita Pasad, Bowen Shi, and Karen Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP*. IEEE, 2023, pp. 1–5.

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

[12] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, "Emotion recognition from speech using Wav2vec 2.0 embeddings," *arXiv [cs.SD]*, Apr. 2021.

[13] Yuanchao Li, Zeyu Zhao, Ondrej Klejch, Peter Bell, and Catherine Lai, "ASR and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition," *arXiv [eess.AS]*, May 2023.

[14] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[15] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*. IEEE, 2018, pp. 5329–5333.

[16] Zhaoxi Mu, Xinyu Yang, Sining Sun, and Qing Yang, "Self-supervised disentangled representation learning for robust target speech extraction," *AAAI*, vol. 38, no. 17, pp. 18815–18823, Mar. 2024.

[17] Yuying Xie, Thomas Arildsen, and Zheng-Hua Tan, "Disentangled speech representation learning based on factorized hierarchical variational autoencoder with self-supervised objective," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2021, pp. 1–6.

[18] SiCheng Yang, Methawee Tantrawenith, Haolin Zhuang, Zhiyong Wu, Aolan Sun, Jianzong Wang, Ning Cheng, Huaizhen Tang, Xintao Zhao, Jie Wang, et al., "Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion," *arXiv preprint arXiv:2208.08757*, 2022.

[19] Jiahao Xun, Shengyu Zhang, Yanting Yang, Jieming Zhu, Liqun Deng, Zhou Zhao, Zhenhua Dong, Ruiqi Li, Lichao Zhang, and Fei Wu, "Discover: Disentangled music representation learning for cover song identification," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 453–463.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[21] Jo-Anne Bachorowski and Michael J Owren, "Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context," *Psychological science*, vol. 6, no. 4, pp. 219–224, 1995.

[22] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[23] Sandeep Kumar Pandey, Hanumant Singh Shekhawat, and SR Mahadeva Prasanna, "Deep learning techniques for speech emotion recognition: A review," in *2019 29th international conference RADIOELEKTRONIKA (RADIOELEKTRONIKA)*. IEEE, 2019, pp. 1–6.

[24] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[25] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, June 2006.

[26] Dan Hendrycks and Kevin Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[27] Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek, "Pre-training audio representations with self-supervision," *IEEE Signal Processing Letters*, vol. 27, pp. 600–604, 2020.

[28] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *arXiv [eess.AS]*, Oct. 2022.