INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Towards Paralinguistic-Only Speech Representations for End-to-End Speech Emotion Recognition

Georgios Ioannides[1,2], *Michael Owen*[2], *Andrew Fletcher*[2], *Viktor Rozgic*[2], *Chao Wang*[2]

[1]Carnegie Mellon University, USA
[2]Amazon Alexa, USA

gioannid@alumni.cmu.edu, {mpowen,afletchz,rozgicv,wngcha}@amazon.com

Abstract

We propose a methodology for information aggregation from the various transformer layer outputs of a generic speech Encoder (e.g. WavLM, HuBERT) for the downstream task of Speech Emotion Recognition (SER). The proposed methodology significantly reduces the dependency of model predictions on linguistic content, while leading to competitive performance without requiring costly Encoder re-training. The proposed paradigm is evaluated via Accuracy, Positive Pointwise Mutual Information, and visualization of the learned attention weights. This methodology generalizes well to a multi-language SER setting in addition to single-language SER, suggesting existing cultural commonalities in the paralinguistic domain between different languages. Experimental results demonstrate this ability by testing our model on unseen languages in a zero-shot fashion, suggesting our proposed method is inclusive in the context of speech and language, thus, making it applicable to a wide audience of speakers.

Index Terms: Deep Learning, Speech Emotion Recognition, Paralinguistics

1. Introduction

In general, SER focuses on recognition of the emotional state of the speakers conveyed through speech signals. The three most important factors for SER are understanding of: (a) "What was said?", i.e. the linguistic content of speech (or lexical features), (b) "How was it said?", i.e. paralinguistic content of speech (or acoustic features), and (c) "What was the context?", i.e. the extralinguistic content of speech, including the speaker identity, habitual aspects of the speaker's voice quality, pitch range and loudness. More specifically, evaluating human emotions using Deep Learning and Signal Processing techniques have led to marked performance improvements as researchers work to understand the underlying information channels used to convey emotions through text, speech and vision [1] [2] [3]. One limitation is that while many models have been proposed, they often do not make an explicit attempt to restrict or measure the incorporation of linguistic information in their model predictions and focus explicitly on paralinguistic information [4].

The motivation behind this work is that, in general, there are various use cases (when it comes to downstream tasks) that are associated with three distinct categories of speech representation extraction during the preprocessing stage: (a) paralinguistic-only speech representations, (b) linguistic-only speech representations, and (c) speech representations mixing both linguistic and paralinguistic information. The ultimate goal of this work is to eventually develop a methodology that would extract latent representations that contain only paralinguistic information, but do not contain lexical information, in order to address specific use cases that would benefit from it.

An example of such a use case is that paralinguistic-only speech representations can potentially be extracted by training a downstream task (e.g., SER) in one natural language (e.g., English) and then, the same model (without any further re-training) can be used during inference for the exact same downstream task, but in another language (e.g., Spanish), while speech representation extraction uses paralinguistic information only. More generally, this use case of transfer learning can be extended to domains outside language. For instance, it can extend to *formality style* [5] (i.e. formal vs informal tone) or type of *Conversational Information Seeking* [6] (i.e. Conversational Search vs Conversational Recommendation vs Conversational Question Answering). However, extracting paralinguistic-only speech representations, while preserving high predictive performance, while also transferring from one domain (e.g., language) to another is a particularly challenging task.

Traditional SER methodologies typically do not make use of end-to-end downstream tasks, but this leads to several limitations [7], including: (a) *error propagation*, i.e. when models are not trained end-to-end, errors made by one component of the system can propagate to subsequent components, leading to compounding errors that can result in poor performance, and (b) *suboptimal solutions*, i.e. when models are trained separately, each component may be optimized to perform well on its individual task, but the overall system may not be optimized for the end goal. This can result in suboptimal solutions that do not perform well in real-world scenarios. Therefore, extraction of informative paralinguistic-only speech representations for SER is especially important since these representations: (a) negate the need to employ either automatic speech recognition (ASR) [8] [9] [10] or any other intermediate step as part of the (pipeline) downstream task, and (b) can generalize better across domains (e.g., formality style, languages, etc.), while also being inclusive the context of speech and language.

Traditional disentanglement techniques in the literature often involve a specifically-designed loss function in order to learn latent representations (in speech [11] or other modalities). This explicitly enforces an arbitrarily-chosen disentanglement metric applied to some or all Encoder layers. Instead, the approach of this work is to implicitly train learnable parameters (of an additional module) that: (a) are included in the preprocessing stage, and (b) are intentionally guided by an appropriate downstream task (e.g., SER), thus, avoiding the limitations of (i) finding the most appropriate metric and (ii) the costly step of training the Encoder layers.

To address the above limitations in SER and speech representation extraction, we propose a novel approach based on efficient utilization of information obtained from the (intermediate) layers of a generic speech Encoder, as an attempt towards

10.21437/Interspeech.2023-497

bridging the gap of paralinguistic-only speech representations. Briefly, our results lead to the following main **contributions**:

- We extract the input embeddings from the intermediate layers of the Encoder, which was pre-trained using Self-Supervised Learning (SSL) and its model weights are kept frozen.
- We introduce a multi-head self-attention [12] [13] mechanism as a module inside the preprocessing stage (Figure 1), which converts the input embeddings into highly informative paralinguistic-only speech representations using learnable parameters in order to be used as input to the downstream SER task.
- We train the Decoder Classifier on phonetically different data (i.e. linguistically different) to identify and aggregate layer feature maps and paralinguistic features in these maps that are most informative for the downstream SER task.

2. Related work

Encoders of pre-trained models (PTMs) trained using SSL approaches (e.g., Wav2vec [14], HuBERT [15], and WavLM [16]) can generate speech representations (mixing both linguistic and paralinguistic information) informative enough to be used as input to a wide range of downstream tasks (listed in SUPERB benchmark [17]) including SER. Typically, the Decoder of the downstream task is a relatively simple network that performs either: (a) mean temporal pooling on the representations of the last Encoder layer [18], or (b) mean temporal pooling on the representations of all Encoder layers and weighted averaging of the representations of each layer [17], followed by fully-connected layers. While the model weights of the Encoder are kept frozen, such low-complexity Decoder networks can be independently trained (for a variety of downstream tasks) in a supervised way using a dataset which is typically much smaller than the one used for training the Encoder.

The authors in [19] draw the conclusion that SSL-based Encoders implicitly capture linguistic information from audio only, implying direct usage of intermediate transformer layer features will have linguistic information, and so they find that attempts in the literature for SER are influenced by linguistic information, which is a limitation. The author in [20] proposes a multi-lingual, multi-task SER approach, but with a similar limitation since, unlike this work, it does not introduce an additional module (or similar) inside the preprocessing stage in order to achieve extracting paralinguistic-only speech representations.

3. Methods

3.1. Encoder and Decoder models

We apply the proposed method on the SSL-based Encoder of a single PTM using different depths and complexities. PTMs used are HuBERT [15] and WavLM [16]. More specifically, the model weights of three particular PTM implementations are used as Encoders: (i) HuBERT-Large, (ii) HuBERT-Base and (iii) WavLM-Large. The reader can use HuggingFace or use the instructions in the original papers of HuBERT and WavLM to implement and use the Encoder models used in our work. If further clarifications are needed, readers can contact the authors. Regarding large-scale datasets, Libri-light [21] has been used for pre-training (i) and (iii). Librispeech [22] has been used for pre-training (ii). GigaSpeech [23] and English parts of VoxPopuli [24] have been used for pre-training (iii). Each PTM implementation is then used as input to the Decoder Classifier (described below), as depicted in Figure 1, which shows the proposed SER system architecture during inference (i.e. after training has already been performed). It should be noted that in this work, PTMs are used as they were pre-trained, i.e. without any further re-training during the preprocessing stage.

For each of the aforementioned PTMs, the Encoder is kept frozen, while we train and evaluate the proposed Decoder on the downstream task of emotion classification (with the classes being angry, happy, sad, neutral) using focal loss. Evaluation is performed on a variety of other languages (e.g. German, Greek and Spanish). Training is performed in: (a) a single-language setting (e.g., using either a Small or a Large English dataset for training), and (b) a multi-language setting (e.g., using an English + Mandarin dataset for training). The specific datasets used in this work are described in detail in section 4.1.

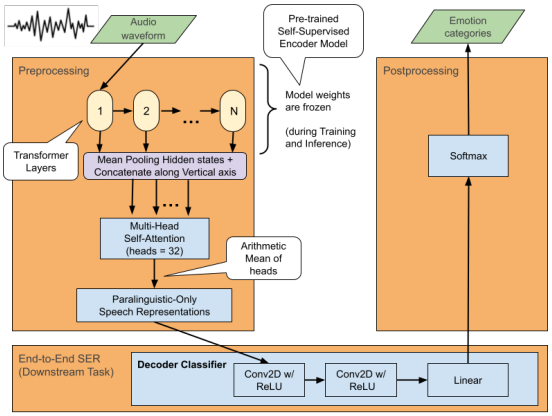

Figure 1: *Proposed SER system architecture*

Each output of the respective transformer layer (of the Encoder) is mean pooled, then all the mean pooled outputs are concatenated, and are fed as input embeddings into the multi-head self-attention module. These input embeddings can be represented mathematically as $X \in R^{N \times d}$ where each x_i is a d-dimensional vector and N is the total number of transformer layers of the Encoder (which is kept frozen). The self-attention mechanism in the module computes a new representation $C \in R^{N \times d}$ of the input sequence, where each c_i is a weighted sum of all input embeddings x_i. The weights for computing c_i are determined by the dot product of a query vector q_k, a key vector k_i and a value vector v_i, which are all learnable parameters of the module. Specifically, the weights for computing c_i are computed as follows:

$$\alpha_{ij} = \frac{exp(q_i^k \cdot k_j)}{\sum_{t=1}^{n} exp(q_i^k \cdot k_t)}$$

where i is the index of the query vector q_k, j is the index of the key vector k_i, and α_{ij} is the weight assigned to the value vector v_i. Once the weights are computed, the context vector c_i can be obtained as a weighted sum of the value vectors: $c_i = \sum_{j=1}^{n} \alpha_{ij} v_j$. The equation computes the context vector c_i for a single query vector q_k. As shown in Figure 1, the multi-head self-attention module computes several different context vectors (in our case 32) using multiple sets of learnable parameters for the query, key, and value vectors. These context vectors are then concatenated and passed through a linear transformation to obtain the final output of the module. Convolutional layers are subsequently used to learn local features (i.e. speech representations) across the generated context matrix from the self-attention module. The 2-dimensional convolutional kernel

achieves this over multiple context vector outputs of each transformer layer at each time step (whereby the time step is defined by the stride). This allows extraction of the latent representations that best explain the different emotion classes. As a result, the contribution (for the specific downstream task of SER) of each transformer layer can be associated with informative paralinguistic-only speech representations.

The model parameters of the convolutional layers in the Decoder classifiers associated with each of the three Encoders (i.e. HuBERT-Large, WavLM-Large and HuBERT-Base) can be found in Table 1. The embedding dimension of the self-attention module is 1024 for the two Large trained model variants and 768 for the Base variant. All Decoder Classifier models are trained for 35 epochs, with gradient accumulation every 2 epochs. All layer weights are initialized using Xavier initialization. The optimizer being used is Adam with both weight decay and an initial learning rate of 10^{-4}, coupled with the focal loss. The focal loss function is defined as $\texttt{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t)$, where $\gamma = 2.5$. A batch size of 8 is used and mixed precision training is utilized. During training and evaluation, audio files are split to a maximum of 5 second clips. If an audio file is longer than 5 seconds, then a new audio file is created with the remaining audio. Each audio file is passed through the trained Encoder model. The outputs of each transformer layer are extracted. Mean pooling is used along the row axis (i.e. each transformer layer output is aggregated using the arithmetic mean to a single row vector of size $1 \times D$) where $D \in \{768, 1024\}$. During evaluation, no weight updates are made to the Decoder, while the rest of the settings remain the same as during training.

	HuBERT/WavLM Large Decoder Classifier (4,329,981 parameters)	HuBERT-Base Decoder Classifier (2,395,645 parameters)
1st CNN Block	• Kernel size: (16,16) • Stride: 1 • Filter size: 512	• Kernel size: (8,8) • Stride: 1 • Filter size: 512
2nd CNN Block	• Kernel size: (2,2) • Stride: 8 • Filter size: 1	• Kernel size: (1,1) • Stride: 8 • Filter size: 1

Table 1: *Decoder Classifier model architectures*

4. Results and Discussion

4.1. Datasets

This work uses the following datasets: (i) IEMOCAP [25] (5-fold cross-validation is employed across the 5 Sessions in which 4 Sessions are used in training and 1 Session is left out for validation), (ii) ESD (the data from the 10 native Mandarin Chinese speakers is used in training) [26], (iii) AESDD Greek data (only used as test data) [27], (iv) EMO-DB German data (only used as test data) [28], (v) MESD Spanish data (only used as test data) [29] and (vi) MSP-Podcast (35% of the data is randomly selected for use in training) [30]. Only the emotion categories of *neutral*, *happiness*, *anger* and *sadness* are considered across all the datasets. All datasets contain speech samples from various speakers and genders. Datasets (iii)-(v) are used to evaluate the non-linguistic dependency performance between the models trained on the following datasets: (1) only on *IEMOCAP* (i.e. Small English dataset), (2) on *IEMOCAP + MSP-Podcast* (i.e. Large English dataset) and (3) on *IEMOCAP + ESD* (i.e. English + Mandarin dataset). This results in approximately the same amount of training data for both datasets (2) and (3) in order to have a fair comparison. The frozen Encoder of each of the three PTM implementations (as described in section 3.1) is used to train and evaluate a SER Decoder on datasets (1) and (3) mentioned above. The best performing PTM in this setting is then used as described in section 4.4. The model weights of each Encoder are kept frozen during fine-tuning of the Decoder Classifier (i.e. its model weights are not modified in any way). The sampling rate of train and validation audio files is 16 kHz.

4.2. Experimental Setup

The evaluation metrics used in this work are the *accuracy (ACC)* and the *Positive Pointwise Mutual Information (PPMI)* [31]. ACC is defined as $\text{ACC} = \frac{total\ correct\ predictions}{total\ number\ of\ predictions}$. The PPMI is used to estimate the extent of lexical dependence of a specific utterance, x, on a specific emotion category, y. PPMI is defined as $\text{PPMI}(x;y) = \max\left(\log_2 \frac{p(x,y)}{p(x)p(y)}, 0\right)$. A pair of outcomes, x and y belong to the discrete random variables X and Y. These outcome pairs quantify the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions, assuming statistical independence. Positive PMI values imply the discrete variables being associated with each other are co-occurring more frequently than expected under a statistical independence assumption, whilst a PPMI value of 0, indicates perfect statistical independence (i.e. the lower the PPMI, the better). Eight A100 NVIDIA GPUs were used for a total of approximately 30 hours.

4.3. Current benchmarks

In Table 2, we present 5-fold cross-validation accuracy (ACC) and standard deviation (std) for models we tested on the IEMOCAP dataset (1) and previous state-of-the-art ACCs reported in the literature. The proposed *Attention-Guided-WavLM-Large-v1* and *v2* models trained on datasets (1) and (3) respectively (as described in section 4.4) achieve state-of-the-art ACCs of *0.7376* and *0.7432* respectively. Evaluation in the single-language setting demonstrates that all four proposed models achieve state-of-the-art performance on dataset (1), when compared to their corresponding methods in the literature.

Model	**ACC (std)**
LightHuBERT-Small[32]	0.6412 (N/A)
HuBERT-Base[18]	0.6492 (N/A)
LightHuBERT-Stage-1[32]	0.6625 (N/A)
HuBERT-Large[18]	0.6762 (N/A)
WavLM-Base+[18]	0.6865 (N/A)
WavLM-Large[16]	0.7062 (N/A)
Attention-Guided-HuBERT-Base	0.6578 (0.021)
Attention-Guided-HuBERT-Large	0.7291 (0.022)
Attention-Guided-WavLM-Large-v1	0.7376 (0.019)
Attention-Guided-WavLM-Large-v2	**0.7432** (0.019)

Table 2: *Models' ACC on IEMOCAP dataset*

4.4. Towards paralinguistic-only representation extraction

The authors in [33], highlight that Mandarin Chinese and English are very different languages in the phonetic (i.e. linguistic) domain, making it difficult to transfer from one to another. Their numerous differences and relatively high variability (in the linguistic domain) make them ideal for training a single SER system, thereby enabling the learnable model parameters in the preprocessing stage (1) to be focused on autonomously learning/extracting common paralinguistic factors in both languages. The following general hypothesis can be stated.

Hypothesis 1. *Let $\mathcal{U}$ be one language, and let $\mathcal{V}$ be another language. Given that we have an objective, $\mathcal{Z}$, whereby (i) $\mathcal{Z}$ is the prediction of emotion category from the speech signals of $\mathcal{U}$ and $\mathcal{V}$, and (ii) $\mathcal{U}$ and $\mathcal{V}$ have limited or no common phonetic (i.e. linguistic) properties, then if the common properties of $\mathcal{U}$ and $\mathcal{V}$ in the paralinguistic domain are learned, a better performance for $\mathcal{Z}$ is achieved than if they were not learned.*

Test set	Happy	Anger	Sadness	Neutral	PPMI
IEMOCAP (Small English dataset)					
MESD	0.1904	1.0429	2.7425	0.6709	1.2812
IEMOCAP + 35% of MSP-Podcast (Large English dataset)					
MESD	0.5910	3.2978	1.3846	0.3564	1.4074
IEMOCAP + ESD (English + Mandarin dataset)					
MESD	0.9142	1.0429	0.5439	0.7539	**0.8138**

Table 3: *Average PPMI model performance*

Test set	Happy	Anger	Sadness	Neutral	ACC
IEMOCAP (Small English dataset)					
EMO-DB	0.9155	0.7953	0.1613	0.7595	0.6873
AESDD	0.8319	0.6777	0.1230	N/A	0.5414
MESD	0.8333	0.3007	0.0833	0.4406	0.4146
IEMOCAP + 35% of MSP-Podcast (Large English dataset)					
EMO-DB	0.7465	0.3858	0.2419	0.9114	0.5575
AESDD	0.5546	0.3802	0.0656	N/A	0.3315
MESD	0.6875	0.0769	0.2361	0.8671	0.4669
IEMOCAP + ESD (English + Mandarin dataset)					
EMO-DB	0.6761	0.9921	0.9677	0.9494	**0.9115**
AESDD	0.6723	0.9256	0.5657	N/A	**0.7210**
MESD	0.6250	0.3986	0.9028	0.7063	**0.6698**

Table 4: *Average ACC model performance*

In multi-lingual setting experiments, we demonstrate that training on a mix of English and Mandarin data significantly improves SER accuracy on unseen languages when compared to English-only training. Evaluation on MESD Spanish data reveals that training on linguistically different data reduces the models' lexical dependency, as mutual-information between lexical content and predicted emotions is significantly lower than for English-only training. In other words, when the attention-based Decoder is trained in multi-lingual settings where languages are different in the linguistic domain, the attention mechanisms aid in the extraction of paralinguistic representations that generalize better across different/unseen languages with paralinguistic domain commonalities. In Tables 3 and 4, the average PPMI and ACC are reported for emotion category predictions of the *Attention-Guided-WavLM-Large-v2* model trained on the *IEMOCAP* dataset, *IEMOCAP + ESD* and *IEMOCAP + 35% of MSP-Podcast*. The EMO-DB and AESDD datasets are used to quantify how well the model learns multi-word paralinguistic properties because utterances in these span across multiple words, while MESD data only includes single word utterances. EMO-DB and AESDD include samples of all emotions for every lexical phrase. So evaluations of PPMI are not applicable (and hence, not included in Table 3), since PPMI should be 0 due to the fact that lexical content and emotional tones are explicitly independent for these datasets via PPMI.

Augmenting English data with Mandarin Chinese data during training consistently reduces the average PPMI indicating reduced dependency on linguistic information. For MESD, the average PPMI is reduced from *1.2812* to *0.8138* (Table 3), implying that multi-language training enables reduced lexical correlation on a language not used in the training. Across all datasets, the augmentation approach improved overall the average ACC indicating that the augmented training led to a more language-general ability to perform SER since it presumably picks up shared cultural properties between languages in the paralinguistic domain. In Table 4, the methodology proposed balances a trade-off in performance between the 'happy' and 'sadness' categories across different languages. This is in contrast to augmenting the Small English dataset (i.e. IEMOCAP) with additional English language data (i.e. 35% of MSP-Podcast), which detriments test language performance significantly, suggesting that a stronger linguistic dependence of the trained Decoder models exists.

4.5. Analysis of Encoder Layer Contribution

Different downstream tasks require information from different layers of the Encoder. Typically, higher level representations that resemble linguistic information are found in top layers while lower layers contain information relevant for e.g., speaker identification task [16]. We compute average normalized attention weights (mirroring the evaluation process in [34]) on the EMO-DB dataset in an attempt towards explainability and interpretability. The Normalized Attention weights of *Attention-Guided-WavLM-Large-v1* and *v2* are shown in Figure 2.

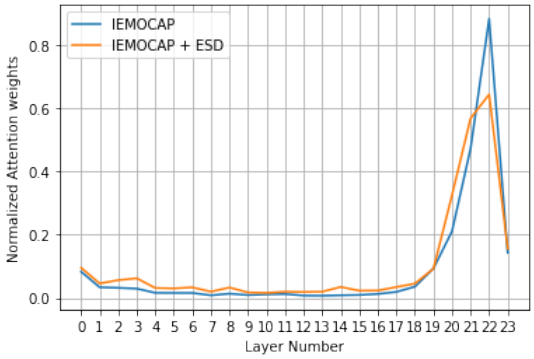

Figure 2: *Layer-wise attention weights*

It is evident that prior to augmenting the training data of the model with the Mandarin Chinese dataset, there is higher contribution of later layers. After training augmentation with Mandarin, the attention weights contribution shifts to lower layers *by approximately 30%*. This suggests that lower-level speech features with augmentation are utilized more than without augmentation. Hence, the model relies more on paralinguistic information rather than on linguistic information, as it is also supported by the results in Table 3.

5. Conclusion

In this work, a methodology for aggregation of information from different transformer layer outputs of a speech Encoder for SER is proposed. The method outperforms previous approaches in the literature by a significant margin and is inclusive in the context of speech and language, making it applicable to a wide audience of speakers. This methodology (a) promotes extraction of paralinguistic-only speech representations for SER by using training sets from two languages, which have limited (if any) common linguistic properties, and (b) evaluates paralinguistic and lexical decoupling in model predictions. Results show that the proposed methodology leads to significant improvements in both ACC and PPMI for various languages and thus, suggesting the existence of cultural commonalities in the paralinguistic domain.

6. References

[1] Y. B. Singh and S. Goel, "A systematic literature review of speech emotion recognition approaches," *Neurocomputing*, vol. 492, pp. 245–263, 2022.

[2] M. Ganganna, "Speech emotion recognition and implementation: A survey." *8th National Conference on Advancements in Information Technology*, 2022.

[3] J. Shor, A. Jansen, W. Han, D. Park, and Y. Zhang, "Universal paralinguistic speech representations using self-supervised conformers." *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[4] J. de Lope and M. Graña, "An ongoing review of speech emotion recognition," *Neurocomputing*, 2023.

[5] S. Rao and J. Tetreault, "Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 129–140. [Online]. Available: https://aclanthology.org/N18-1012

[6] J. Dalton, S. Fischer, P. Owoicho, F. Radlinski, F. Rossetto, J. R. Trippas, and H. Zamani, "Conversational information seeking: Theory and application," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3455–3458. [Online]. Available: https://doi.org/10.1145/3477495.3532678

[7] T. Caselli, P. Vossen, M. Erp, A. Fokkens, F. Ilievski, R. Beviá, M. Lê, R. Morante, and M. Postma, "When it's all piling up: Investigating error propagation in an nlp pipeline," *CEUR Workshop Proceedings*, vol. 1386, 01 2015.

[8] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, and K. Keutzer, "Squeezeformer: An efficient transformer for automatic speech recognition," 2022. [Online]. Available: https://arxiv.org/abs/2206.00888

[9] F.-J. Chang, J. Liu, M. Radfar, A. Mouchtaris, M. Omologo, A. Rastrow, and S. Kunzmann, "Context-aware transformer transducer for speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 503–510.

[10] L. Lu, C. Liu, J. Li, and Y. Gong, "Exploring transformers for large-scale speech recognition," in *Interspeech*, 2020.

[11] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," in *Interspeech*, 2021.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need." *Advances in Neural Information Processing Systems*, 2017.

[13] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "Multi-head attention: Collaborate instead of concatenate," 2020. [Online]. Available: https://arxiv.org/abs/2006.16362

[14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020.

[15] 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 Trans. Audio, Speech and Lang. Proc.*, 2021.

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[17] H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-w. Yang, S. Dong, A. Liu, C.-I. Lai, J. Shi, X. Chang, P. Hall, H.-J. Chen, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities." *Association for Computational Linguistics*, 2022.

[18] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, S. Dong, S.-W. Li, S. Watanabe, A. rahman Mohamed, and H. yi Lee et al., "Superb: Speech processing universal performance benchmark." *Interspeech*, 2021.

[19] 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 & Machine Intelligence*, 2023.

[20] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[21] J. K. et al., "Libri-light: A benchmark for asr with limited or no supervision," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[22] 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)*, 2015.

[23] G. C. et al., "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in *Interspeech*, 2021.

[24] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *International Joint Conference on Natural Language Processing*, 2021.

[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, 2008.

[26] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, 2022.

[27] N. Vryzas, R. Kotsakis, A. Liatsou, C. A. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," *Journal of AES*, 2018.

[28] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech." *9th European Conference on Speech Communication and Technology*, 2005.

[29] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "Mexican emotional speech database based on semantic, frequency, familiarity, concreteness, and cultural shaping of affective prosody," *Data*, 2021.

[30] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, 2019.

[31] E. H. Sanchez, M. Serrurier, and M. Ortner, "Learning disentangled representations via mutual information estimation," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Springer International Publishing, 2020.

[32] R. Wang, Q. Bai, J. Ao, L. Zhou, Z. Xiong, Z. Wei, Y. Zhang, T. Ko, and H. Li, "Lighthubert: Lightweight and configurable speech representation learning with once-for-all hidden-unit bert." *Interspeech*, 2022.

[33] C. Y. Suen, "Computational analysis of mandarin sounds with reference to the english language," in *Proceedings of the 9th Conference on Computational Linguistics - Volume 1*. Academia Praha, 1982.

[34] G. Brunner, Y. Liu, D. Pascual, O. Richter, M. Ciaramita, and R. Wattenhofer, "On identifiability in transformers." *International Conference on Learning Representations (ICLR)*, 2020.