Cross-Corpus Speech Emotion Recognition with HuBERT Self-Supervised Representation.

Miguel A. Pastor, Dayana Ribas, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
{mapastor,dribas,ortega,amiguel,lleida}@unizar.es

Abstract

Speech Emotion Recognition (SER) is a task related to many applications in the framework of human-machine interaction. However, the lack of suitable speech emotional datasets compromises the performance of the SER systems. A lot of labeled data are required to accomplish successful training, especially for current Deep Neural Network (DNN)-based solutions. Previous works have explored different strategies for extending the training set using some emotion speech corpora available. In this paper, we evaluate the impact on the performance of cross-corpus as a data augmentation strategy for spectral representations and the recent Self-Supervised (SS) representation of HuBERT in an SER system. Experimental results show improvements in the accuracy of SER in the IEMOCAP dataset when extending the training set with two other datasets, EmoDB in German and RAVDESS in English.

Index Terms: speech emotion recognition, cross-corpus, data augmentation, HuBERT, self-supervised representation.

1. Introduction

Human-machine interaction interfaces are more common day by day and many industries are incorporating related technologies, such as call centers, social communities, metaverse, customer identifications, and online/offline meetings, just to mention a few [1]. The possibility of understanding the behavior and feelings of the person involved in a human-machine conversation is a key point for ensuring a fluid interaction. This information translates into valuable business insights for further decision-making. Speech Emotion Recognition (SER) systems are among the main technologies required for this kind of application. However, currently, the performance of SER systems is still low in accuracy due to the huge challenge of recognizing the emotion from a single speech segment [2, 3]. Anyway, there are many research efforts focused on developing robust and accurate SER systems [4, 5], because multiple industries could benefit from the management of that kind of non-verbal communication. For example, call centers could get a complete quality analysis of their services, instead of using the current method, consisting of listening to a random selection of calls. Online gaming and other services, which include oral communication among users, could also use this kind of solution to detect incorrect behaviors of their users, reducing human intervention in the process.

One of the main problems to develop SER systems is the limited amount of emotional speech available. Namely, speech data with realistic (natural, spontaneous) emotions instead of acted ones, and therefore, emotion labels without subjectivity. The nature of human emotions makes this data collection a very difficult task in practice, and the fact is that the available speech data for developing SER systems usually include simulated emotions, with very subjective labeling, low duration, and mostly in the English language. Consequently, the generalization ability of the SER systems is compromised. The strategy of joining many speech emotion data available for enlarging the training is known as cross-corpus or joint training. This has been previously used in SER systems [6, 7] attempting to get more generalization and compensate for the small size of most of the SER databases. So far, most of these works have employed spectral and cepstral parameter sets as the entry of the classifier, such as Mel-Frequency Cepstral Coefficients (MFCC) or handcrafted spectral parameter sets such as those developed for the Compare Challenge[1].

In this paper, we explore the cross-corpus as a strategy for extending the training set. We combine two speech emotion datasets, EmoDb in German and RAVDESS in English, to augment the training set of the IEMOCAP dataset which is in the English language. First, we evaluate the cross-corpus performance by implementing a baseline system for SER based on spectral-based representations and a statistical classifier based on Support Vector Machine (SVM) [8]. Then, we update the system to a Deep Neural Network (DNN) solution employing the HuBERT Self-Supervised (SS) model for representation and classifiers based on DNN. In this case, we take advantage of the SUPERB (Speech processing Universal PERformance Benchmark) framework [9] for developing the SER system. In conclusion, the main contribution and novelty of this paper consist in evaluating the impact on the performance of cross-corpus for the SS representation HuBERT as a training extension strategy in an SER system.

In the following, section 2 presents the background and the previous works developed for studying the performance of cross-corpus in SER systems. The experimental setup in section 3 explains the organization and configuration of the experiments carried out to evaluate the hypothesis. Section 4 presents the system proposed divided into two parts the speech representation and the classifier, both based on DNN. Finally, section 5 discusses on obtained results and section 6 concludes the paper.

2. Previous Work

There have been some previous works related to the cross-corpus strategy applied to emotion recognition. Schuller et al. [6] compared the results of two strategies for training the system: voting versus pooling. The voting system consists in training several systems with different datasets and fusion the final scores, while pooling consists of training one single system with data from several datasets altogether. They found that pooling databases for extending the training achieved bigger improvements than training with a single dataset and make a fusion of results. Then, Liu et al. [10] attempted to create a common subspace shared by all the databases used, and then train and

[1] http://www.compare.openaudio.eu/

10.21437/IberSPEECH.2022-16

test with different ones. More recently, Braunschweiler et al. [7] studied the cross-corpus strategy in an SER system based on DNN and combined it with classical data augmentation, by adding random noise and volume perturbations. The main problem related to cross-corpus data augment is the domain mismatch among train and test generated by the difference among recording conditions and the method used to get the emotional speech. To solve this problem, there are some proposals on the domain adaptation strategy. Milner[11] trained a domain classifier in parallel, considering that knowing the dataset where the audio comes from, results would be better. On the other hand, Hongchao et al. [12] and Lian et al. [13] used DANN (Domain Adversarial Neural Networks) in combination with cross-corpus. The goal of this kind of network is to get invariant features among datasets, which allows to improve performance and also, to use a more independent model of the database. Another form of domain mismatch is the different languages of the datasets. The viability of combining datasets of multiple and unrelated languages was studied by Zehra et al. [14]. They obtained significant accuracy improvement in assembling datasets in English, German, Italian, and Urdu. That last language is particularly important, due to the great differences from the rest, even belonging to the Indo-European family.

So far, cross-corpus studies have been mainly performed with spectral [15] and cepstral [7] representations. Also, hand-crafted parameter sets have been used, such as in Schuller et al. [6]. In fact, for the SER task, there are just a few works where the SS representations are used. This is the case of Pepino et al. [16] that evaluated the performance of Wav2vec2.0[17] for SER systems, but they did not use a cross-corpus strategy. In this paper, we use the recent SS representation HuBERT and evaluate its performance in a pooling strategy for cross-corpus. HuBERT is a general-purpose speech representation that has shown very promising performance for speech tasks[2].

3. Experimental setup

In all setups, we made a 5-fold multiple classification, with four different emotions. 10% of the training partitions are randomly reserved as development set in each epoch. We run 30000 epoch. The final result will be the test result in the epoch that has obtained the best development result. The test set consists of one-fifth of the speakers, gender-balanced. In IEMOCAP, the test set is one recording session.

3.1. Databases

For this experiment, we have chosen the following datasets: EmoDb [18], RAVDESS [19], and IEMOCAP [20]. Table 1 shows the main properties of databases. All of them are freely available. EmoDb and RAVDESS are standard acted datasets, which means that groups of professional actors are recorded when reading a text and pretending the emotion requested. The text read is the same for all actors and emotions, which means that the speech content is not relevant for emotion recognition. On the other hand, IEMOCAP is also an acted dataset but presents some differences. The actors improvise a dialogue under the guidance of the researchers, and then, the audio samples are labeled by other people, oblivious to the rest of the process. That is supposed to create a more realistic dataset, in which the speech is related to the acted emotion.

To unify and balance the emotions for all the datasets, they have been reduced to four: neutral, happiness, anger, and sadness. This is a common problem in emotion datasets, in which some emotions, such as surprise or disgust, are scarce, while neutral is much more abundant. All the audio records have one label. All audio samples are clean, mono-channel and their sample rate is 16 kHz. RAVDESS has a sample rate of 44.1 kHz, so we downsampled the dataset before combining it with the rest.

3.2. Evaluation metrics

To evaluate the classification performance we use Unweighted Average Recall (UAR) (eq:1). Values in the confusion matrix are employed for computing the score, i.e., true and false positive and negative rates (TP, FP, TN, FN). UAR is generally used in emotion classification because it considers each class by itself, so it is suitable to deal with the usual imbalance in the number of audios for each label.

$$UAR = 0.5 \cdot \frac{TP}{TP+FN} + 0.5 \cdot \frac{TN}{FP+TN} \tag{1}$$

4. Speech Recognition Systems Under Evaluation

In this work, the systems under evaluation are divided into two parts: the speech emotion representation and the classifier. We use two types of representations: spectral-based and DNN-based. Spectral representations are combined with the statistical SVM classifier as in previous works [8]. The parameter sets with spectral features employed in this work are GeMAPS [21], eGeMAPS [22] and ComParE [22]. They have 62, 88, and 6373 parameters respectively. On the other hand, the DNN-based representation used is HuBERT [23], which is a Self-Supervised Speech representation based on Transformers. In this case, we used DNN-based classifiers. First, we used a pooling of the sequence of HuBERT vectors followed by a linear layer, to get a basic accuracy using the most simple classifier. Then, a Convolutional Neural Network (CNN) with Self Attention and a CT-Transformer was employed as classifiers too.

4.1. Feature Extraction

4.1.1. Spectral based Features

We use three parameter sets to train the SVM classifier: ComParE, GeMAPS, and its extension, eGeMAPS. Low-Level Descriptors (LLD) of these parameters are presented in Table 2. That parameters are extracted in from overlapping variable length windows. Several statistical functions are applied to each LLD to obtain better representation, including mean, variance, kurtosis, and skewness. These features are spectral hand-crafted representations that describe different levels of speech information, namely phonetic, prosodic, etc. We use these parameter sets because they have been widely used for SER systems[7, 8].

4.1.2. DNN based representation: HuBERT

HuBERT [23] (Hidden Unit BERT) is a modification of BERT [24] (Bidirectional Encoder Representations from Transformers) language processing model, designed for speech processing. The model has three parts. First, a 7-layer CNN encoder, followed by a Transformer block, results in a variable-size projection. That size depends on the Transformer size, which gives us three different models. In this work, we always use the base size due to GPU-memory limitations.

[2] https://superbbenchmark.org/leaderboard

Database	**Duration**	**Language**	**Speakers**	**# Neutral**	**# Happy**	**# Angry**	**# Sad**	**Text**	**Quality**	**Type**
EmoDb	16 min	German	10	79	71	127	62	Read	Studio	Acted
RAVDESS	42 min	English	24	96	192	192	192	Read	Studio	Acted
IEMOCAP	7 hours	English	10	1708	1636	1103	1084	Improv.	Studio	Acted

Table 1: *Databases information. "# Emotion" indicates the number of labels of each emotion*

FEATURES	I	II	III
Energy parameters			
Shimmer, Loudness, HNR	X	X	X
Prob. of voicing			X
RASTA spectrum (energy)			X
RMS Energy, Zero-Crossing Rate			X
Frecuency parameters			
Pitch, Jitter	X	X	X
Formants 1-3. Frequency, energy	X	X	
BW Formants	1	1-3	
MFCC		1-4	1-14
Alfa Ratio, Hammarberg Index			X
Spectral Slope			X
Harmonic difference H1-H2, A3			X
RASTA spectrum (frecuency 0-8 kHz)			X
Spectral energy 250-650 Hz, 1-4 kHz			X
Spectral Flux, Centroid, Entropy, Slope			X
Psychoacoustic Sharpness			X
Psychoacoustic Harmonicity			X
Spectral Variance, Skewness, Kurtosis			X

Table 2: *Spectral features: (I) GeMAPS, (II) eGeMAPS, (III) ComParE*

The main advantage of this model over previous ones is the bidirectional training system. The previous models, such as Wav2vec[25], only had into account the previous audio samples. This is achieved by randomly erasing a random 15 % of the samples in text, instead of predicting the following sample.

4.2. Classification Methods

We use four different classification methods. First, a classical statistical-based classifier to establish the baseline: Support Vector Machines (SVM). Also, DNN-based classifiers are employed, starting from a basic pooling followed by a linear layer attempting to evaluate the performance of the raw SS representation. Then, more complex architectures are employed for classification: a CNN with Self-Attention and a Class Token (CT) Transformer.

4.2.1. Support Vector Machines

SVM has been widely used for speech processing including emotion recognition [6]. This method tries to find the optimal hyperplane to establish the boundary among the samples of the training dataset. The kernel used is RBF (Radial Basis Function). To choose the hyperparameters (C and γ) we made an exponential sweep inspired by the previous work of Kessing et al. [8]. The function of the C parameter is to control errors, therefore the higher C, the more errors are allowed in training. γ is a measure of the curvature of the boundary, so the higher the γ, the more curved is the border.

4.2.2. CNN Self Attention

Convolutional Neural Networks (CNN) with Self Attention mechanisms have been widely used in several speech processing tasks due to their parallelization properties and the great performance they have shown [26]. It consists of a regular DNN with a secondary branch. That branch is multiplied by the result of the primary one. Thus, the secondary branch indicates the primary one in which parts of the signal should the DNN pay attention [27].

In this work, we employ a CNN-Self Attention network based on a fully connected layer that adapts the size of the input to the size of the network. The architecture consists of a block repeated twice composed of dropout, five convolutional layers, and a non-linearity ReLU. Finally, an output block contains the following sequence of layers: dropout, five convolutional layers, fully connected, ReLU, and a final fully connected layer for output of the final classification.

4.2.3. CT Transformer

The CT-Transformer is inspired by the Vision Transformer (ViT) [28] developed for image processing, but in this case, we process the temporal sequence of embeddings from SS representations to perform classification. The Class token concept concentrates the class information in a single vector through several layers of self-attention mechanisms [24]. It encodes the temporal information in the training stage using the whole sequence of embeddings through a configurable number of heads and layers in the MSA block. In this case, we used two layers and six heads. The attention learns the weights to sum these embeddings for each layer and outputs a vector consisting of the concatenation of the attention heads. This way the Class token learns a global description of the utterance, where the multiple attention heads perform as slots of this final representation vector of the utterance. The multiple heads implied in the process can better capture the underlying information in the sequence than the pooling alternative we used in the previous section.

5. Results and Discussion

This section presents the experiments carried out for evaluating the performance of the cross-corpus strategy using two types of SER systems:

1. Spectral-based system: Spectral parameter sets as feature extractor (GeMAPS, eGeMAPS and ComParE) combined with SVM classifier.
2. DNN-based system: HuBERT SS representation as feature extractor with three DNN-based classifiers: Pooling + Linear Layer, CNN with Self-Attention, and CT-Transformer.

Experiments were executed in the SUPERB framework. Aiming to improve the performance, we carried out some auxiliary experiments by modifying the SER system architecture configuration. For instance, the projector dimension of the feature extractor was increased from 512 to 768, and the number of layers in the classifier was increased from 256 to 512. Obtained results showed a slight influence on the SER system performance, around $1 - 2\%$, without an increment in the training time.

5.1. Cross-corpus with Spectral-based system

Table 3 presents the results of the cross-corpus experiment using spectral representations and an SVM classifier. Three spectral representations (see Table 1) and two conditions were evaluated.

- *Matched Train Condition* is when the train and test sets belong to the same dataset.
- *Extended Train Condition* is when there are three datasets in the training set (EmoDb + RAVDESS + IEMOCAP).

System	Matched Train	Extended Train
Evaluation dataset: EmoDb		
GeMAPS-SVM	78,21%	77,40%
eGeMAPS-SVM	78,94%	79,67%
ComParE-SVM	82,50%	**83,53%**
Evaluation dataset: RAVDESS		
GeMAPS-SVM	60,43%	64,06%
eGeMAPS-SVM	61,28%	63,91%
ComParE-SVM	65,60%	**67,81%**
Evaluation dataset: IEMOCAP		
GeMAPS-SVM	58,91%	58,52%
eGeMAPS-SVM	59,58%	59,79%
ComParE-SVM	62,79%	**63,54%**

Table 3: *UAR results of handcrafted spectral parameters with SVM classifier*

We see that the best results were obtained for the Extended train condition at combining all datasets in the training. This result is consistent for all databases evaluated. The best parameter set is ComParE, with improved results beyond 3% for all datasets/conditions evaluated. This is an expected result because ComParE set has a huge amount of parameters (6373) compared to GeMAPS (62) and eGeMAPS (88). In table 2 we can see that RASTA spectrum and energy are the main difference among ComParE and eGeMAPS/GeMAPS features, so they could have a notable responsibility for the improvement.

5.2. Cross-corpus with DNN-based system

Table 4 shows the results of the cross-corpus experiments using HuBERT and DNN-based classifiers. In this case, we did not compute individual results for EmoDb and RAVDESS, because IEMOCAP is almost exclusively used in systems that use DNN.

System	Matched Train	Extended Train
Evaluation dataset: IEMOCAP		
Hubert-LinearLayer	65,60%	65,86%
Hubert-CNNSelfAtt	65,53%	**67,58%**
Hubert-Transformer	64,48%	67,03%

Table 4: *UAR results of HuBERT with Neural classifier using Train = Test and Extended train = IEMOCAP + RAVDESS + EmoDb*

Comparing with previous results using the Spectral-based system we can see that the DNN-based system outperforms results in the previous table 3 for all cases. For instance, the most basic classifier, namely the linear layer, has 1.5% of absolute improvement. The results obtained with combined datasets are better than when training only with IEMOCAP. These results agree with the SVM classifier, showing that data augmentation via a combination of databases is a good way to increase the performance of emotion classification systems.

Results confirm the previous finding in [14] about the performance improvements when training with datasets that include different languages. This indicates that the DNN-based system can use the augmentation of training data for generalizing the model despite the differences. On the other hand, obtained results are in line with the findings in Pepino et al. [16], where the authors used a different SS representation (Wav2vec2.0[17]) to get the embeddings of the audio.

Note that despite the CT-Transformer being the most complex architecture, its results are similar to the other DNN-based classifiers. We think that the reason behind this may be that such a complex model needs more data to be properly trained. The biggest train set still has eight hours of audio. On the other side note that the results obtained with HuBERT representation are better than handcrafted parameter sets. This is consistent with other tasks of voice processing, in which Transformer representations have caused a significant improvement in the previous results.

6. Conclusions

In this paper, we have evaluated the performance of the cross-corpus strategy for data augmentation in SER systems. Obtained results demonstrate that the cross-corpus benefits the system performance in both, Spectral-based systems and DNN-based systems. For DNN-based systems, we have evaluated the HuBERT SS representation which outperforms the traditional Spectral-based system. By combining SS representations with cross-corpus we achieved significant improvement in IEMOCAP (UAR=67,58%), which is 3% better than SVM with ComParE parameter sets. Results show the utility of both the SS representations and the combination of the databases, even when those databases are in different languages. Further research is needed to study the suitable properties of databases to impact the system's performance. Also, better results could be obtained by using other SS representations which obtained better results than HuBERT in other tasks, such as WavLM[29]. This work invites further research in the cross-corpus for emotion classification to elucidate which properties of emotional datasets make them suitable to be combined.

7. Acknowledgements

This work was supported in part by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Grant 101007666; in part by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU" / PRTR under Grants PDC2021-120846-C41 PID2021-126061OB-C44, and in part by the Government of Aragon(Grant Group T36_20R).

8. References

[1] N. Gupta, *Human-Machine Interaction and IoT Applications for a Smarter World.* Milton: Taylor Francis Group, 2022.

[2] G. Castellano, L. Kessous, and G. Caridakis, *Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech.* Berlin, Heidelberg: Springer Berlin Heidelberg,

2008, pp. 92–103. [Online]. Available: https://doi.org/10.1007/978-3-540-85099-1_8

[3] A. Thakur and S. Dhull, “Speech emotion recognition: A review,” in *Advances in Communication and Computational Technology*, G. S. Hura, A. K. Singh, and L. Siong Hoe, Eds. Singapore: Springer Singapore, 2021, pp. 815–827.

[4] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, “A comprehensive review of speech emotion recognition systems,” *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021.

[5] Y. B. Singh and S. Goel, “A systematic literature review of speech emotion recognition approaches,” *Neurocomputing*, vol. 492, pp. 245–263, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222003964

[6] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, “Using multiple databases for training in emotion recognition: To unite or to vote?” Proceedings of the Annual Conference of the International Speech Communication Association, 8 2011, pp. 1553–1556.

[7] N. Braunschweiler, R. Doddipatla, S. Keizer, and S. Stoyanchev, “A study on cross-corpus speech emotion recognition and data augmentation,” in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 24–30.

[8] A. Keesing, Y. Koh, and M. Witbrock, “Acoustic features and neural representations for categorical emotion recognition from speech.” Proceedings of the Annual Conference of the International Speech Communication Association, 08 2021, pp. 3415–3419.

[9] S. Yang, P. Chi, Y. Chuang, C. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G. Lin, T. Huang, W. Tseng, K. Lee, D. Liu, Z. Huang, S. Dong, S. Li, S. Watanabe, A. Mohamed, and H. Lee, “SUPERB: speech processing universal performance benchmark,” *CoRR*, vol. abs/2105.01051, 2021. [Online]. Available: https://arxiv.org/abs/2105.01051

[10] N. Liu, Y. Zong, B. Zhang, L. Liu, J. Chen, G. Zhao, and J. Zhu, “Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5144–5148.

[11] R. Milner, M. A. Jalal, R. W. M. Ng, and T. Hain, “A cross-corpus study on speech emotion recognition,” *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 304–311, 2019.

[12] H. Ma, C. Zhang, X. Zhou, J. Chen, and Q. Zhou, “Domain adversarial network for cross-domain emotion recognition in conversation,” *Applied Sciences*, vol. 12, no. 11, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/11/5436

[13] Z. Lian, J. Tao, B. Liu, and J. Huang, “Domain adversarial learning for emotion recognition.” Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 10 2019.

[14] W. Zehra, A. R. Javed, Z. Jalil, T. Gadekallu, and H. Kahn, “Cross corpus multi-lingual speech emotion recognition using ensemble learning,” *Complex Intelligent Systems*, vol. 7, 01 2021.

[15] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, “Cnn+lstm architecture for speech emotion recognition with data augmentation,” *Workshop on Speech, Music and Mind (SMM 2018)*, 2018.

[16] L. Pepino, P. E. Riera, and L. Ferrer, “Emotion recognition from speech using wav2vec 2.0 embeddings,” in *Interspeech*, 2021.

[17] A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *ArXiv*, vol. abs/2006.11477, 2020.

[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech.” in *INTERSPEECH*. ISCA, 2005, pp. 1517–1520. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2005.html#BurkhardtPRSW05

[19] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),” Apr. 2018, Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak. [Online]. Available: https://doi.org/10.5281/zenodo.1188976

[20] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[22] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, “The interspeech 2016 computational paralinguistics challenge: Deception, sincerity and native language,” 09 2016, pp. 2001–2005.

[23] 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.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT*, 2019, p. 4171–4186.

[25] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in *INTERSPEECH*, 2019.

[26] M. Sperber, J. Niehues, G. Neubig, S. Stüker, and A. H. Waibel, “Self-attentional acoustic models,” in *INTERSPEECH*, 2018.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.

[29] 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*, pp. 1–14, 2022.