Towards Improving Speech Emotion Recognition Using Synthetic Data Augmentation from Emotion Conversion

Karim M Ibrahim, Antony Perzo, Simon Leglaive

TOWARDS IMPROVING SPEECH EMOTION RECOGNITION USING SYNTHETIC DATA AUGMENTATION FROM EMOTION CONVERSION

Karim M. Ibrahim[1] *Antony Perzo*[1] *Simon Leglaive*[2]

[1]Emobot, France
[2]CentraleSupélec, IETR (UMR CNRS 6164), France
k.ibrahim@emobot.fr

ABSTRACT

One of the main challenges in speech emotion recognition is the lack of large labelled datasets. The progress in speech synthesis allows us to generate reliable and realistic expressive speech. In this work, we propose using a state-of-the-art end-to-end speech emotion conversion model to generate new synthetic data for training speech emotion recognition models. We first evaluate the quality of the converted speech on new unseen datasets, which proves to be on par with the training data. Then, we study the effect of using the synthesized speech as data augmentation. We show that this approach improves the overall performance of emotion recognition models on two different datasets, IEMOCAP and RAVDESS, both in the cases of speaker dependent and independent emotion recognition using a fine-tuned wav2vec 2.0.

Index Terms— speech emotion recognition, synthetic data, data augmentation, speech generation

1. INTRODUCTION

Speech Emotion Recognition (SER) has been gaining increasing attention due to its importance in multiple fields, such as health care, customer service, education, and human computer interactions [1]. Deep learning approaches have significantly improved the performance and accuracy of predicting emotions from speech. However, they require large amounts of labelled data which is currently lacking. Current datasets are small due to the complexity and time consuming effort in collecting and manually labelling the data by multiple annotators [2], particularly in less spoken languages, which results in models that overfit and poorly generalize. Approaches relying on using synthetic data from speech generative models propose an alternative to overcome this challenge in various speech classification tasks, including SER.

Synthetic data is artificially generated data, which can be used to replace or augment real data in training deep learning models. Such approach has multiple advantages in terms of data privacy and security [3], balancing skewed datasets [4, 5], as well as overcoming the lack of large datasets, as the case with SER [6]. The quality and realism of synthetic data is critical for its effectiveness in deep learning applications. The recent advances in generative models have significantly improved the quality of synthetic data.

Speech synthesis has witnessed substantial progress in recent years, thanks to the progress in deep learning and neural network architectures [7]. Emotion-conditioned speech generation, a subfield of speech synthesis, focuses on generating speech that conveys specific emotional characteristics. Traditional speech synthesis approaches rely on text-to-speech to vocalize the lexical content [8]. However, they struggle with non-verbal vocalization, e.g. laughter and cries, or with emphasis and rhythm, which is referred to as prosody [9]. This is an essential part in expressing and recognizing emotions. Recent approaches relying on textless speech synthesis achieve high-quality speech that can be conditioned on prosody and emotions [10]. In this work, we explore using such model for synthetic data augmentation in the task of SER. Our experiments show promising results of applying data augmentation using synthetic raw audio to improve the performance of SER models.

2. RELATED WORK

Traditional SER systems utilize hand-crafted features to train a classification model [2], while recent approaches rely on using features extracted from large pre-trained model, such as wav2vec 2.0 [11], which is fine-tuned for the downstream task. Data augmentation in traditional systems relied on synthesizing the hand-crafted features to be used during training as proposed by Sahu et al. [12], which uses Generative Adversarial Networks (GANs) [13]. Similarly, Bao et al. [14] proposed leveraging unlabeled speech datasets for data augmentation by using Cycle-GAN [15] to transfer the emotion style of the feature vectors. Both approaches showed promising results.

Although hand-crafted features are easier to model due to their lower dimensionality, they might be incompatible with state-of-the-art recognition models, such as wav2vec 2.0. Generating spectrograms or raw waveforms provides more flexibility by allowing us to train models directly on the raw data. Chatziagapi et al. [4] and Wang et al. [5] proposed generating mel spectrograms using GANs to tackle data imbalance by augmenting the minority classes. Similarly Eskimez et al. [16] used an improved version of GANs with higher generation quality to apply SER data augmentation using spectrograms.

Few approaches have explored synthetic data augmentation for SER in the waveform domain. Rizos et al. [17] proposed using speech emotion conversion to generate synthetic data using a StarGAN model [18] and the WORLD vocoder [19]. He et al. [20] improved the previous model by separating emotional features from emotion-independent features during the training process. Both approaches showed promising results. However, they rely on generating the input parameters to the vocoder, i.e. spectral envelope, fundamental frequency, and aperiodicity parameters. Recently, there has been a shift towards using end-to-end speech synthesis for its several advantages, e.g. alleviating the need for extensive feature engineering, as well as allowing rich conditioning on various attributes, such as speaker and emotion [21].

In this work, we focus on generating synthetic raw audio using textless end-to-end speech emotion conversion [10] to address SER

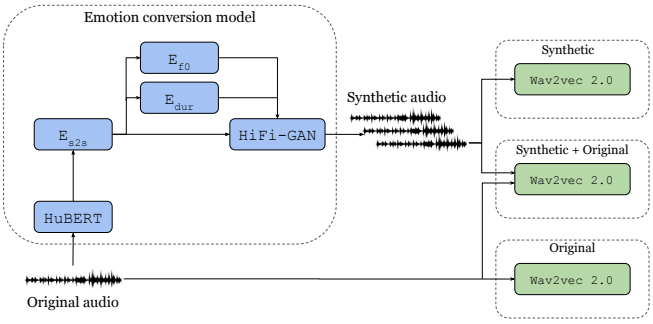

Fig. 1. An illustration of the proposed approach with different experimental setups of fine-tuning the wav2vec 2.0 model using either the emotionally converted synthetic data, original data, or both simultaneously.

data augmentation, which is suitable to use with the current state-of-the-art models applied to raw audio. Experimental results show that the proposed approach outperforms more traditional data augmentation techniques.

3. METHOD

To investigate the reliability of synthetic data in SER, we rely on two different models:

a) a generative model to synthesize speech (speech-to-speech emotion conversion);

b) an emotion classification model from the raw audio waveform (fine-tuned wav2vec 2.0).

An illustration of the pipeline is presented in Figure 1. In the following, we describe the architecture and training procedure for each model.

3.1. Speech-to-speech emotion conversion

To synthesize expressive speech, we make use of the current state-of-the-art in emotion conversion [10]. The approach is based on using phonetic-content representation of speech. Such representation allows the problem to be treated as a spoken language translation problem, where the objective is to learn to map this discrete speech representation between different emotions. Furthermore, to emphasize the prosody of different emotions, the approach uses two additional modules to predict the duration and the fundamental frequency (F0) of each phonetic representation. Finally, a variation of the HiFi-GAN neural vocoder [22] was used to synthesize the speech from the converted speech phonetic-content units. In the following, we briefly describe the details of each module. We refer to [10] for the full details of the model implementation.

3.1.1. Phonetic-content representation

To produce a low-level representation of the phonetic content of speech, we use the pre-trained representations of the large self-supervised model HuBERT [23]. This allows us to, not only represent the phonetic content, but also the non-verbal content such as laughter or cry. The input to the HuBERT model is the audio waveform $\mathbf{x} \in \mathbb{R}^T$ and the output is the embedded representation $\mathbf{z}' \in \mathbb{R}^L$, where T is the number of samples in the input waveform and L is the number of produced phonetic-content units. The representations are further discretized using K-means clustering with $K = 200$ to produce the final representation $\mathbf{z} \in \{1, ..., K\}^L$. Similar to [24, 10], we remove repeated units (e.g., 0,0,1,1,2 $\rightarrow$ 0,1,2), as the duration prediction module will predict the repetition for each phonetic-content unit based on the target emotion.

3.1.2. Unit translation module

To convert from one emotion to another, we use a sequence-to-sequence (E_{S2S}) model to translate the phonetic-content unit representations.

The input to the model is composed of (i) the source phonetic-content units representation $\mathbf{z}_{src} \in \{1, ..., K\}^L$ corresponding to the input signal associated with the original emotion; and (ii) the target emotion label y_{tgt}, which is represented as a one hot vector. This allows the model to add or remove non-verbal vocalizations which are appropriate to the target emotion. The model outputs the target phonetic-content unit representation $\hat{\mathbf{z}}_{tgt} \in \{1, ..., K\}^L$, which encodes the speech signal modified to match the target emotion:

$$\hat{\mathbf{z}}_{tgt} = E_{s2s}(\mathbf{z}_{src}, y_{tgt}). \quad (1)$$

The model is trained to minimize the cross-entropy loss between the predicted phonetic-content units $\hat{\mathbf{z}}_{tgt}$ and the ground-truth ones $\mathbf{z}_{tgt}$ using a dataset of parallel emotional utterances.

3.1.3. Prosody prediction

To emphasize the emotional expression, prosodic features are then predicted corresponding to the target emotion. For each phonetic-content units in the output of the translation model, the duration and fundamental frequency are predicted, subject to the target emotion. Similar to [10, 8], we use a Convolutional Neural Network (CNN) to learn the mapping between phonetic-content units to durations, referred to as the duration prediction model (E_{dur}). We remove the repetitions from the groundtruth unit representations and train the model to predict these durations per emotion using the Mean Squared Error (MSE), using their original duration as training labels.

Next, we train another model E_{F0} for the F0 predictions . Similar to the duration prediction model, we use a CNN model followed by a linear layer to predict the F0 value. The original F0 in the groundtruth recordings are extracted using YAAPT [25], then used as the training targets for the CNN model.

3.1.4. Speech synthesis

For synthesizing the speech from the converted phonetic-content units, we use a variation of the HiFi-GAN neural vocoder [22]. HiFi-GAN is modified to take as input a sequence of phonetic-content units after repeating them according to the predicted durations, along with the predicted F0, the target speaker embeddings, and the target emotion. These features are concatenated and fed into a sequence of convolutional layers to predict the waveform of the speech signal. This signal would constitute the synthetic data to be used in training the SER model.

3.2. Fine-tuned Wav2vec 2.0

Wav2vec 2.0 [11] is a framework for self-supervised learning of representations from raw audio. The model is composed of three different stages. The first stage, the local encoder, contains multiple

convolutional layers that encodes the audio waveform into embeddings with a stride of 20 ms and receptive field of 25 ms. The second stage is a contextualized encoder, which takes the embeddings from the previous stage as input. Its architecture is made of several transformer encoder blocks [26]. Finally, a quantization module is used to quantize the embeddings into discrete units. The model has been pre-trained and released to the public as a foundation model to be used in several downstream tasks, one of which is SER.

Several approaches have been proposed to fine-tune the wav2vec 2.0 model on the SER task [27, 28], which is currently providing the state-of-the-art performance on multiple datasets. Similar to [27], we adapt the wav2vec 2.0 model for SER by adding a linear layer for the downstream tasks. Additionally, we freeze the local encoder layers, i.e. the convolutional layers, and we fine-tune the contextual encoder, i.e. the transformer layers, along with the linear layers.

4. DATASETS

We use different datasets for each task. For the task of training the emotion conversion model, we use EmoV [29], a dataset of speech utterances recorded with multiple emotions. For the task of SER, we use two standard datasets, IEMOCAP [30], and RAVDESS [31]. In the following, we briefly introduce each dataset.

EmoV: We use the Emotional Voices Database (EmoV) for training and evaluating the emotion conversion model. EmoV is made up of 7000 utterances, each one is recorded with multiple different emotions: neutral, amused, angry, sleepy, and disgusted, by four native speakers (two male and two female speakers). Hence, these utterances are used to create the parallel pairs, where the same lexical content is recorded in multiple different emotions, which is used in training the emotion conversion model. To tackle the small size of the dataset (around 9 hours), we match parallel pairs between the different speakers, i.e. using parallel recordings of the same transcript in different emotions and by different speakers, as proposed in [10]. The dataset is split into train/validation/test with a ratio of 90/5/5, such that there is no overlap of utterances between the sets.

IEMOCAP: The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [30] consists of approximately 12 hours of scripted and improvised dialogues by 10 different speakers. The dataset is composed of 5 sessions, each including speech from an actor and an actress. Similar to the standard practice, we used 4 emotional classes: anger, happiness, sadness and neutral, and following the work in [32], we relabeled excitement samples as happiness. Additionally, we used standard durations of 8 seconds, where shorter samples are padded and longer ones are trimmed. IEMOCAP is used to fine-tune wav2vec 2.0 for the SER task, using a 5-fold cross validation.

RAVDESS The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [31] is a dataset of emotional speech and songs. It is recorded by 24 different actors (12 males and 12 females) reciting 2 statements "Kids are talking by the door" and "Dogs are sitting by the door" with 8 different emotions. Similar to IEMOCAP, we use 4 emotional classes: happy, sad, angry, and neutral. We used only the speech recordings and discarded the sung ones. Following [27], we merged the neutral and calm emotions together. Finally, we used standard durations of 5 seconds for this dataset, where shorter samples are padded and longer samples are trimmed.

Table 1. MOS evaluation of the original and synthetic audio on the EmoV, IEMOCAP, and RAVDESS datasets, expressed as mean and 95 % confidence interval across raters. The speech emotion conversion model is trained on EmoV.

Dataset	Original	Synthetic
EmoV	4.40 ± 0.18	3.31 ± 0.18
IEMOCAP	4.48 ± 0.15	3.12 ± 0.20
RAVDESS	4.83 ± 0.08	3.38 ± 0.19

5. EXPERIMENTS

We perform two different evaluations, one on the perceived quality of the synthesized data, and one on the performance of the SER model. For the first case, we perform a subjective evaluation using Mean-Opinion-Score (MOS) through a blind listening test. Since the emotion conversion model was trained only on the EmoV dataset, we are particularly interested in the quality of the conversion on the IEMOCAP and RAVDESS datasets and how it compares to the quality on the dataset used for training.

We asked 9 participants to rate the perceived quality of the speech on a scale from 1 to 5. Table 1 shows the average MOS results of the test across the 9 raters on the three different datasets both for the original and synthetic speech. We observe a comparable quality of the synthesized data on all datasets, confirming the performance of the emotion conversion model can generalize to new datasets.

In order to test the influence of using synthetic data in training SER models, we defined the following scenarios for comparison:

1. **Original:** In the original scenario we train and test the wav2vec 2.0 model using the original dataset whether it is IEMOCAP or RAVDESS, without any augmentation, similar to [27].
2. **Synthetic:** We train using only the synthetic data, i.e. the generated speech from the emotion conversion model applied to each dataset. However, we test on the original dataset.
3. **Synthetic + Original:** In this case, we train with both the original dataset and the synthetic data as augmentation. We test the effect of augmenting the dataset by adding synthetic data with different ratios: 25%, 50%, 75%, and 100% of the size of the original dataset. The test is performed on the original recordings.
4. **Baseline:** In this case we augment the dataset using traditional audio augmentation [33]: adding noise, pitch shifting, and time stretching, applied on the input data. We train the model using both the original and augmented data and test on the original data.

Additionally, to test the effect of augmenting the dataset with new speakers from the generative model, we experiment with two setups:

a) *Speaker Dependent (SD):* In this case the test set and the training set have overlapping speakers.

b) *Speaker Independent (SI):* In this case we prevent overlaps between speakers in the training and test splits.

Since the synthetic data is composed of new speakers, it is a speaker independent setup by default. Figure 1 shows a summary of the different setups used in our experiments.

Table 2. Accuracy of the wav2vec 2.0 model fine-tuned on original data, synthetic data, or both, compared to traditional audio augmentation in the speaker dependent (SD) and speaker independent (SI) cases, expressed as mean and 95 % confidence interval across 5-fold cross validation on the IEMOCAP and RAVDESS datasets.

Dataset	Setup	Synthetic	Original [27]	Original + Synthetic	Baseline [33]
IEMOCAP	SD	-	74.16 ± 2.00	**76.19 ± 1.95**	74.17± 1.37
	SI	56.88 ± 3.64	63.97 ± 2.93	**66.06 ± 2.21**	65.32 ± 3.44
RAVDESS	SD	-	91.08 ± 2.93	**93.05 ± 2.12**	92.71 ± 2.20
	SI	47.42 ± 2.96	81.01 ± 3.19	81.29 ± 2.77	**82.29 ± 2.96**

Finally, we fine-tuned the wav2vec 2.0 on the two datasets, IEMOCAP and RAVDESS, using 5-fold cross validation for each of the previous scenarios. The training set is further split to training and validation sets with 0.9 and 0.1 ratios respectively. The model was trained on each training splits till convergence, using early stopping with a patience of 5 epochs based on the validation loss. We optimize the cross entropy loss using batches of 32 instances, Adam optimizer with a learning rate of 0.001 and linear weight decay. Regarding training the emotion conversion model, we used the same training parameters described in [10].

Table 2 shows the results of the experiments in the different setups. We observe a clear improvement in the performance of the SER model in the case of augmentation using the synthetic data compared to the use of the original dataset and the traditional audio augmentation. In the IEMOCAP dataset, we get 2.03% and 2.09% improvement when using synthetic data for augmentation, over training with the original dataset in the cases of speaker dependent and independent setups respectively. Similarly, in the case of the RAVDESS dataset, we get an improvement of 1.97% and 0.28% compared to the use of the original data alone. However, in the speaker independent setting on RAVDESS, the traditional audio augmentation outperforms the synthetic data augmentation. Furthermore, by training only using the synthetic data, we achieve an accuracy of 56.88% and 47.42% on the IEMOCAP and RAVDESS datasets respectively, which surpassed our expectations.

In Figure 2, we observe the effect of augmenting the dataset with different ratios of the synthetic data using 0.25, 0.5, 0.75, 1.0 relative size of synthetic data to the original. We find an incremental improvement in the performance of the model, specially on IEMOCAP dataset for the speaker dependent and independent setups. The improvement is relative to the dataset used, with 0.75 ratio being approximately the optimum ratio in majority of scenarios. This further validates the effect of using synthetic data in improving the performance of SER models.

6. CONCLUSION

In this work, we investigated speech-to-speech emotion conversion for generating synthetic data to be used for data augmentation in SER models. We validated the perceptive quality of the synthetic data when applying the emotion conversion on new unseen data through a subjective evaluation. Our experiments across two different datasets showed an improvement when using the proposed approach compared to more traditional data augmentation techniques. These results encourage us to further explore synthetic data for SER to overcome the challenges of collecting a reliable and large dataset. Furthermore, such approach is promising in the cases of less spoken languages, where there is no labelled dataset. In future work, we plan to explore the efficacy of our approach on multiple languages.

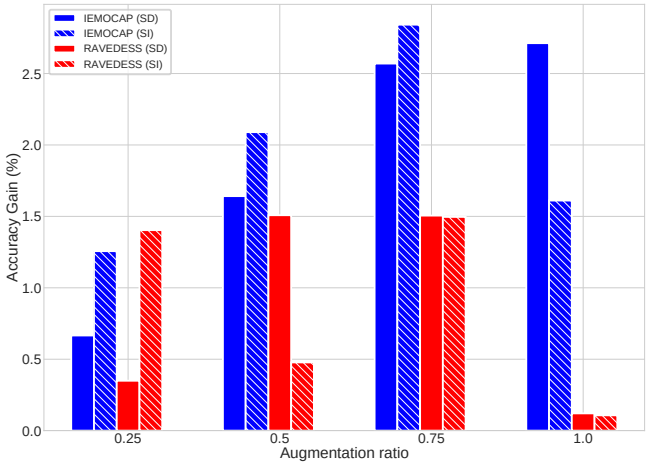

Fig. 2. Average accuracy gain across different synthetic augmentation ratios of the original dataset size for both IEMOCAP and RAVDESS in the speaker dependent (SD) and speaker independent (SI) cases.

7. REFERENCES

[1] 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*, 2021.

[2] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2021.

[3] S. I. Nikolenko, *Synthetic data for deep learning*, Springer, 2021.

[4] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, and A. Katsamanis, "Data augmentation using GANs for speech emotion recognition.," in *Proceedings of the International Speech Communication Association (INTERSPEECH)*, 2019.

[5] S. Wang, H. Hemati, J. Guonason, and D. Borth, "Generative data augmentation guided by triplet loss for speech emotion recognition," *arXiv preprint arXiv:2208.04994*, 2022.

[6] S. Latif, A. Shahid, and J. Qadir, "Generative emotional AI for speech emotion recognition: The case for synthetic emotional speech augmentation," *Applied Acoustics*, 2023.

[7] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review

of deep learning based speech synthesis," *Applied Sciences*, 2019.

[8] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proceedings of the International Conference on Learning Representations*, 2020.

[9] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Language and Cognitive Processes*, 2010.

[10] F. Kreuk, A. Polyak, J. Copet, E. Kharitonov, T. A. Nguyen, M. Rivière, W.-N. Hsu, A. Mohamed, E. Dupoux, and Y. Adi, "Textless speech emotion conversion using discrete & decomposed representations," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, 2020.

[12] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," *arXiv preprint arXiv:1806.06626*, 2018.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, 2014.

[14] F. Bao, M. Neumann, and N. T. Vu, "CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition.," in *Proceedings of the International Speech Communication Association (INTERSPEECH)*, 2019.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[16] S. E. Eskimez, D. Dimitriadis, R. Gmyr, and K. Kumanati, "GAN-based data generation for speech emotion recognition.," in *Proceedings of the International Speech Communication Association (INTERSPEECH)*, 2020.

[17] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "StarGAN for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[18] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[19] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[20] X. He, J. Chen, G. Rizos, and B. W. Schuller, "An Improved StarGAN for Emotional Voice Conversion: Enhancing Voice Quality and Data Augmentation," in *Proceedings of the International Speech Communication Association (INTERSPEECH)*, 2021.

[21] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. A. J. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of the International Speech Communication Association (INTERSPEECH)*, 2017.

[22] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, 2020.

[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*, 2021.

[24] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, et al., "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, 2021.

[25] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2002.

[26] 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.

[27] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, 2021.

[28] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[29] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.

[30] 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*, 2008.

[31] 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*, 2018.

[32] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, 2017.

[33] S. Wei, S. Zou, F. Liao, et al., "A comparison on data augmentation methods based on deep learning for audio classification," in *Journal of Physics: Conference Series*. IOP Publishing, 2020.