Towards Naturalistic Voice Conversion: NaturalVoices Dataset with an Automatic Processing Pipeline

Ali N. Salman, Zongyang Du, Shreeram Suresh Chandra, İsmail Rasim Ülgen, Carlos Busso, Berrak Sisman

Department of Electrical and Computer Engineering, The University of Texas at Dallas

{ali.salman, zxd220002, shreeramsuresh.chandra, ismailrasim.ulgen, busso}@utdallas.edu, berraksisman@u.nus.edu

Abstract

Voice conversion (VC) research traditionally depends on scripted or acted speech, which lacks the natural spontaneity of real-life conversations. While natural speech data is limited for VC, our study focuses on filling in this gap. We introduce a novel data-sourcing pipeline that makes the release of a natural speech dataset for VC, named NaturalVoices. The pipeline extracts rich information in speech such as emotion and signal-to-noise ratio (SNR) from raw podcast data, utilizing recent deep learning methods and providing flexibility and ease of use. NaturalVoices marks a large-scale, spontaneous, expressive, and emotional speech dataset, comprising over 3,800 hours speech sourced from the original podcasts in the MSP-Podcast dataset. Objective and subjective evaluations demonstrate the effectiveness of using our pipeline for providing natural and expressive data for VC, suggesting the potential of NaturalVoices for broader speech generation tasks.

Index Terms: Data pipeline, automatic data sourcing, voice conversion

1. Introduction

VC aims to convert one speaker's voice to sound like that of target speaker while preserving linguistic content [1]. It has various applications such as movie dubbing, intelligent dialogue systems, real-time voice cloning, voice assistants, and conversational agents.

Most VC frameworks [1, 2] are typically trained using scripted or acted corpus [3, 4], resulting in the generation of high-quality speech predominantly with a reading or acting style. However, the process of collecting acted data for VC is labor-intensive and time-consuming, often burdened with inefficiencies. Real-life speech, in contrast, is spontaneous and encompasses various speaking styles [5], emotional expressions [4], nonverbal cues [6] such as laughter, and lip smacks, as well as dysfluencies like repetitions, hesitations or interruptions [7]. Therefore, the progress of VC models requires obtaining a more diverse dataset that reflects the richness, complexity, and expressiveness of spontaneous human speech.

In this paper, we introduce an automated method for sourcing data from podcasts for VC. Podcasts offer distinct advantages over other in-the-wild data sources such as YouTube data. Unlike the diverse and often noisy background of other data sources, podcasts generally have higher-quality audio recordings with clearer speech, making them particularly suitable for VC tasks. Additionally, podcasts provide a diverse range of speakers, ensuring that there's enough speech data for each speaker to effectively model their identity. This abundance of varied speakers and ample speech data for each speaker enables the successful transformation of speaker identity for VC task.

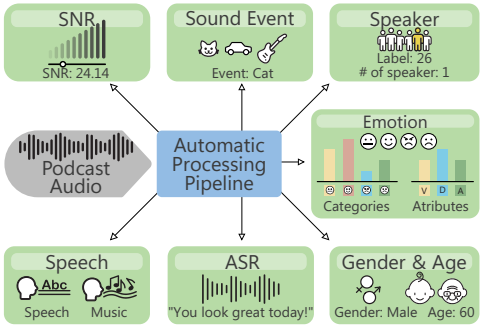


Figure 1: *An illustration of our data sourcing pipeline with various modules.*

The components of our pipeline are depicted in Figure 1. Leveraging state-of-the-art deep learning techniques from various speech tasks including diarization, automatic speech recognition, speaker recognition, and speech emotion recognition, our automated data sourcing pipeline presents an innovative solution to the challenges of gathering spontaneous and naturalistic data for VC. Applying this pipeline to the original podcasts in the MSP-Podcast dataset [9], we release a subset optimized for VC, known as NaturalVoices. Below, we outline several key advantages of our approach and the NaturalVoices dataset:

Richness and Expressiveness: NaturalVoices comprises 3,846 hours of speech data from numerous speakers, capturing natural emotional expressions, diverse communication styles, nonverbal vocal cues, and a variety of background sounds in authentic recording conditions. Our pipeline provides transcripts, speaker details (such as label and gender), signal-to-noise ratio (SNR), emotion attributes (arousal, dominance, valence) and emotion categories (neutral, angry, happy, sad), and sound event categories like laughter and animal sounds (e.g., rooster crowing). With such rich and expressive data provided by our pipeline, NaturalVoices can be utilized across a range of tasks in speech synthesis.

Flexibility and Easy-to-Use : Our dataset and pipeline are publicly accessible[1]. Utilizing our pipeline, users can easily filter NaturalVoices speech data based on specific criteria such as SNR, emotion states and gender. This simplifies dataset utilization across diverse applications.

The rest of this paper is structured as follows: Section 2 provides an overview of related work. In Section 3, we outline the details of our proposed automatic data-sourcing pipeline. Section 4 introduces the NaturalVoices dataset and explores its potential applications in other speech tasks. Experimental results are presented in Section 5. Finally, Section 6 summarizes our findings and concludes the study.

[1]https://github.com/3loi/NaturalVoices

Table 1: *A comparison of the NaturalVoices dataset with other common VC datasets. The NaturalVoices dataset marks as the first large-scale, spontaneous, expressive, and emotional speech dataset for VC.*

	Speech Type	Total Hour	# of Speakers	SNR	Sound Event Categories	Emotion		Text
						Categories	Attributes	
VCTK [8]	Read	44	110	No	No	Neutral Speech		Transcript
ESD [4]	Read	15	10	No	No	Yes	No	Transcript
NaturalVoices	Spontaneous	3846	>2467	Yes	Yes	Yes	Yes	ASR

2. Related Work

2.1. Emotional Speech Datasets

Emotions play a vital role in human communication, yet understanding and synthesizing emotional speech remains challenging due to the nuanced and intricate expressions found in real-life conversations. Speech emotion recognition (SER) is an important task for emotion understanding. Existing emotional speech datasets for SER [10, 11] often rely on scripted recordings, resulting in overemphasized emotions that differ from authentic emotional nuances found in natural conversation. Alternatively, collecting other SER datasets [12, 13] through conversational improvisations for spontaneous speech is labor-intensive and costly [4]. However, as highlighted by Zhou et al. [4], these SER emotional datasets often lack lexical variability and may contain external noise and overlapping speech, making them unsuitable for VC.

Lotfian et al. [9] introduced the MSP-Podcast dataset for SER, consisting of retrieved speech segments with annotated emotion categories and attributes from podcast recordings. These podcasts contain natural conversations among diverse speakers discussing various topics. This diversity ensures large lexical and speaker variabilities, crucial for VC datasets. Inspired by this, we applied our pipeline to create NaturalVoices for VC task, without relying on annotations or labels from MSP-Podcast.

2.2. Datasets for Voice Conversion

VC frameworks require data from multiple speakers to model various aspects of speech, including speaker identity [1, 2], speaking style [14, 15], and emotional information [5]. VC datasets often consist of read speech, lacking natural spontaneity. For instance, the CSTR VCTK corpus [8] contains recordings from 110 English speakers, and similarly, other VC datasets [16–19] primarily provide only neutral speech. The ESD dataset [4] includes emotional speech from 10 speakers with various emotional states, and has been used widely in VC. However, ESD is limited in size and consists solely of acted speech. To address the need for more diverse datasets reflecting spontaneous human speech, this paper introduces NaturalVoices, a large, natural, and emotional dataset for VC applications.

3. Automatic Data Sourcing Pipeline

In this section, we introduce our automatic data-sourcing pipeline inspired by [20]. Initially, our pipeline divides each podcast from the MSP-Podcast dataset into segments of appropriate duration. Subsequently, it automatically annotates these segments with transcripts, speaker details, SNR, emotion attributes, and sound event categories using different modules, as illustrated in Figure 1. Our pipeline offers users the ability to easily and flexibly filter our dataset according to their preferences.

Diarization and ASR Module: We utilize the Faster Whisper model, an accelerated version of Whisper [21] combined with CTranslate2, to segment the audio of each episode into short segments (4-6 seconds) and provide accurate transcripts for each segment, as well as, language detection. However, minor misalignments persist in the model at segment start and end times. To resolve this, we adjust segment durations based on intervening silence: extending adjacent segments by 0.25 seconds each if silence exceeds 0.5 seconds, or merging segments for shorter silences. We then use the Montreal Forced Aligner (MFA) [22] to align audio samples with transcriptions, generating phone-level alignments using word-to-phoneme mappings.

Speech Detection Module: Some audio segments contain music instead of speech, which is usually undesired in many VC and speech synthesis applications, as the music can act as noise during training these models. We employ a temporal convolutional network [23] to distinguish between speech and music.

Speaker Recognition Module: A crucial aspect of training VC models is maintaining a unique and distinct speaker set to accurately represent speaker characteristics. We address this in two steps: Firstly, we utilize PyAnnote [24, 25] to identify 'local speakers' within individual audio files. Secondly, we consolidate 'local speakers' from different files into 'global speakers' through manually annotating a single segment per file. This process creates a unique 'global speaker' set for overlap-free applications and offers flexibility with 'local speakers'.

Gender Classification and Age Module: We employ a model [26] for age and gender prediction, providing additional information for the speaker in each segment.

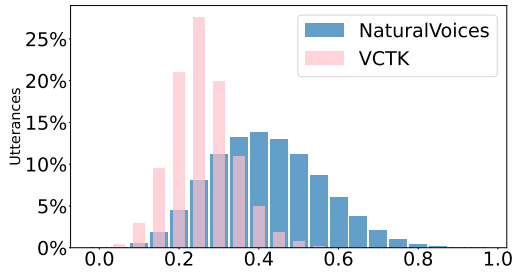
Emotion Attribute and Category Module: We employ two high-performing emotion detection models. The first, PEFT-SER [27], is a categorical model based on WavLM and LORA, trained to identify primary emotion classes within speech utterances, including neutral, happiness, sadness, and anger. The second is a regression-based WavLM model [28], designed to assess the emotional spectrum of valence (ranging from negative to positive), arousal (from calm to excited), and dominance (from weak to strong) in each audio segment, thus providing a broader emotional context.

SNR Module: SNR quantifies the level of desired signal relative to background noise in an audio recording. We estimate the SNR value for each audio segment using WADA-SNR [29].

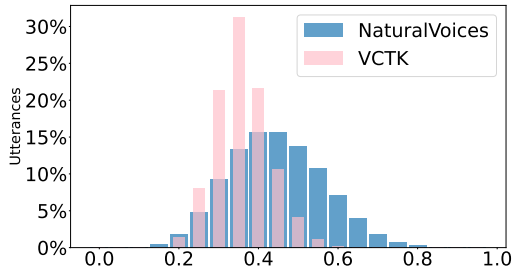
Sound Event Detection Module We use the AST model [30] to predict over 500 different sounds (e.g., honking, alarm, animal noises, etc) for each segment. By using this model we can filter the data to include or exclude any type of background or abnormal sounds.

4. NaturalVoices Dataset

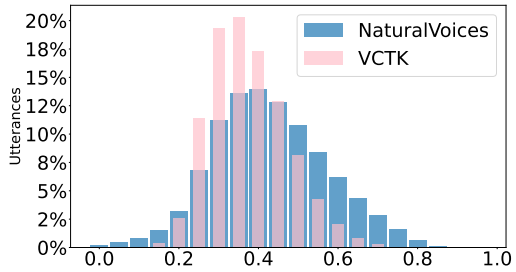
Our research introduces a novel pipeline and the creation of publicly available, large-scale speech dataset, named NaturalVoices. NaturalVoices includes 3.8k hours of speech with over 2 million utterances, averaging 6.67 seconds each. Within this, 2.6k hours contain single-speaker speech in English, with 1.3k hours labeled for 2,467 speakers. The dataset includes 1,115 female speakers and 1,338 male speakers. All data is down-sampled to 16kHz. For each utterance, we provide other automatically annotated information obtained from the pipeline



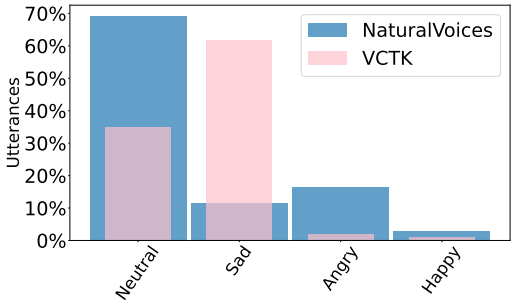
(a) *Arousal*



(b) *Dominance*



(c) *Valence*



(d) *Categorical*

Figure 2: *Emotion attribute (arousal, dominance, valence) and category (neutral, sad, angry, happy) distribution in NaturalVoices versus VCTK.*

that are 1) Signal-to-Noise Ratio, 2) emotion attributes (arousal, dominance, valence), 3) emotion categories (neutral, angry, happy, sad), 4) speech or music classification, and 5) sound event categories. With this dataset, we aim to facilitate advancements in speech processing, in particular in expressive speech synthesis and VC.

4.1. Analysis

In our study, we conducted a comparative analysis between our dataset, NaturalVoices, and two widely used VC datasets: ESD [4] and VCTK [8], as summarized in Table 1. The results indicate that our dataset contains more natural and spontaneous speech, along with additional automatically annotated information. Compared to ESD, our dataset has a larger scale, including more speakers and greater lexical variability. Compared to VCTK, our dataset is more expressive, larger in scale, and includes more speakers. In the following sections, we provide a detailed comparison of VCTK from various perspectives.

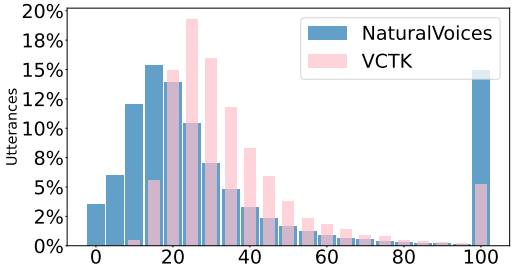


Figure 3: *SNR distribution in NaturalVoices and VCTK.*

Figures 2a, 2b, and 2c provide visual representations of the distribution of emotion attributes – arousal, dominance, and valence – within both the VCTK and NaturalVoices datasets. Arousal reflects the degree of excitement or calmness, dominance indicates the intensity of an emotion, and valence denotes the positivity or negativity of an emotion [31]. As illustrated in Figures 2a, 2b, and 2c, NaturalVoices exhibits a broad range across these three emotion attributes, indicating a higher level of expressiveness and the presence of stronger emotions. Conversely, VCTK demonstrates less variation, predominantly skewing towards negative attributes.

Moreover, categorical analysis highlights our dataset's wider spectrum of emotions in Figure 2d, including a higher prevalence of neutral, angry, and happy expressions, alongside a significantly larger dataset compared to VCTK. The SNR distribution in Figure 3 highlights the diversity of NaturalVoices in terms of background noise and varied environments, with segments below 40 SNR and at high SNR levels of 100. Compared to VCTK, our dataset contains more samples in the high SNR range (above 100) and a broader distribution across SNR levels, including more samples in the 0-20 SNR range. VCTK, on the other hand, predominantly focuses on the 20-40 dB range, with fewer samples in the extreme SNR values. This diversity underscores our dataset's wide range of recording scenarios, including instances with challenging background noise levels and those with optimal signal-to-noise ratios.

4.2. Possible Applications

Our dataset presents a wealth of opportunities for diverse applications in speech processing and machine learning. In this section, we explore its potential impact on speech synthesis and other speech tasks.

Expressive Speech Synthesis and VC: Our dataset clearly facilitates speech synthesis and VC research. Furthermore, it can be used for expressive VC [5] and emotional VC [4]. While expressive VC focuses on converting speaker identity for emotional speakers, emotional VC specifically targets the conversion of emotional states. By providing a rich collection of natural speech data with diverse emotional attributes, our dataset enables VC models to capture and model nuanced emotional expressions. Additionally, with additional automatically annotated text, researchers can leverage this dataset to enhance the expressiveness and emotional fidelity of synthesized speech by developing text-to-speech (TTS) models.

Noisy-to-Noisy VC and Noise Robust VC: Our dataset's diverse range of SNRs makes it ideal for exploring noisy-to-noisy VC, where the goal is to conduct identity conversion while preserving background sounds [32]. Researchers can use this dataset to develop robust VC models that are capable of handling noisy input conditions, leading to improved performance in real-world scenarios with varying levels of background noise.

Spontaneous Speech Modeling and Generation: Understanding, modeling and generating spontaneous speech pose ongoing challenges. Our dataset encompasses a diverse range of conversational styles, pauses, and speech disfluencies, capturing the natural variability present in spontaneous speech. Researchers can leverage this dataset to train models for spontaneous speech generation, as well as for tasks such as spontaneous speech recognition and understanding.

Weakly-Labeled Supervised Training: Given that all annotations in our dataset are generated by deep learning models, the presence of weakly-labeled data offers a unique opportunity for weakly-labeled supervised training approaches. In such cases, labels may be incomplete or noisy, yet researchers can effectively leverage this data using techniques like semi-supervised learning and self-training. By doing so, they can improve model performance and generalization on various tasks, including ASR, speaker diarization, and SER.

5. Experiments for VC

In this work, our primary focus is to explore the use of NaturalVoices in VC under various settings. We demonstrate the effectiveness of NaturalVoices across a range of VC scenarios.

5.1. Experimental Setup and Evaluation Metrics

We employ TriAANVC [33] as our chosen VC model. Employing an encoder-decoder architecture, it effectively disentangles content and speaker features. Through the Triple Adaptive Attention Normalization (TAAN) block, the model extracts detailed and global speaker representations via adaptive normalization, ensuring preservation of source content with siamese loss and time masking. TriAANVC showcases state-of-the-art performance in non-parallel any-to-any VC tasks, as evidenced by evaluations on the VCTK dataset.

We trained TriAANVC on our dataset, and compared the results with those obtained from the VCTK dataset. For training details, we utilized a batch size of 100, conducted training over 500 epochs, employed the Adam optimizer with a learning rate of 10^-5, and set model parameters to H = 256, C = 512, and L = 6, with CPC features. We used the ParallelWaveGAN [34] vocoder trained on our dataset to generate utterances. For the VCTK dataset, we employed the pre-trained model [2]. We randomly selected six speakers as seen speakers and six speakers as unseen speakers from their demo page.

For objective evaluations, we focus on two key metrics: speaker similarity and intelligibility, which are both crucial aspects in assessing the effectiveness of VC systems. Speaker similarity is measured using the acceptance rate from a Speaker Verification (SV) model, based on cosine similarity between embedding vectors of target and converted speech. The threshold is determined using the equal error rate from our NaturalVoices dataset. Word Error Rate (WER) and Character Error Rate (CER) assess script discrepancies, with the script for converted utterances obtained using a pre-trained Whisper model [21]. For subjective evaluation, we conducted a Mean Opinion Score (MOS) [5] listening test to evaluate speech quality and intelligibility. 12 subjects participated in all experiments.

5.2. Experimental Comparisons and Results

We utilized automatic annotations from our pipeline to filter training data from NaturalVoices, aiming for settings similar to VCTK: multiple speakers with neutral speech. This portion of our dataset was then employed to train TriAANVC. Since our dataset is out of distribution compared to VCTK, we also trained a vocoder with our dataset to enhance performance. Consequently, we obtained speech samples generated by two vocoders: one pre-trained on VCTK (denoted as NaturalVoices$_{VCTK}$), and the other trained on our dataset (referred to as NaturalVoices). The results presented in Table 2 demonstrate that NaturalVoices achieves better performance in terms of speaker similarity and comparable performance for CER and WER, underscoring the efficacy of our dataset for VC tasks. Furthermore, training the vocoder on our dataset yields performance improvements, a noteworthy consideration given the dataset's out-of-distribution nature relative to VCTK. Subjective evaluation results in Table 4 also affirm that our datasets are suitable for VC.

We examined the benefits of dataset size expansion by filtering a larger training dataset, denoted as NaturalVoices$_{-Large}$, to train TriAANVC. Table 2 reports that a larger dataset enhances speaker identity conversion in the VC model. Given the spontaneous nature of NaturalVoices, increasing the size of the training data presents a challenge for the model designed for acted speech, particularly in modeling linguistic information for spontaneous speech.

Additionally, we explored the effect of different SNR levels on the VC model. We filtered our data into two settings: Low SNR (0-20 dB) and high SNR (80-100 dB). As demonstrated in Table 3, we can observe that the performance of the VC model is affected by background noise. Our dataset contributes to the development of robust VC models by providing valuable training data.

Table 2: *Objective results for TriAANVC with VCTK and NaturalVoices in* seen-to-seen *(S2S) and* unseen-to-unseen *(U2U) settings.*

Training Data	SV(%)↑		CER(%)↓		WER(%)↓	
	S2S	U2U	S2S	U2U	S2S	U2U
VCTK [8]	71.10	80.30	16.42	12.38	25.74	19.82
NaturalVoices	93.65	89.16	17.01	18.55	27.20	30.43
NaturalVoices$_{vctk}$	80.75	82.76	19.31	19.26	30.70	31.02
NaturalVoices$_{-Large}$	96.78	73.80	19.68	22.26	30.37	33.31

Table 3: *Objective results for TriAANVC across varying SNR levels in a S2S setting.*

	SV(%) ↑	CER(%)↓	WER(%)↓
Low SNR	85.11	32.69	46.91
High SNR	93.65	17.01	27.20

Table 4: *MOS results for NaturalVoices with 95% confidence interval.*

	Quality↑	Intelligibility↑
Generated Speech	3.17±0.23	3.77±0.36
NaturalVoices	4.38±0.16	4.79±0.18

6. Conclusion

In our paper, we introduce NaturalVoices, a novel large-scale spontaneous speech dataset comprising over 3,800 hours of diverse and emotional speech sourced from podcast data, leveraging an innovative data-sourcing pipeline. The pipeline operates concurrently to predict multiple labels relevant to speech synthesis and is designed to accommodate expansion as new and improved models emerge. Experiment results show VC model can generate natural and intelligible speech by using NaturalVoices, indicating its potential for broader speech generation applications. We will explore other expressive speech synthesis tasks with NaturalVoices in future work.

[2]https://github.com/winddori2002/TriAAN-VC

7. References

[1] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[2] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100–2104.

[3] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[4] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[5] Z. Du, B. Sisman, K. Zhou, and H. Li, "Expressive voice conversion: A joint framework for speaker identity and emotional style transfer," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 594–601.

[6] T. A. Nguyen, W.-N. Hsu, A. d'Avirro, B. Shi, I. Gat, M. Fazel-Zarani, T. Remez, J. Copet, G. Synnaeve, M. Hassid *et al.*, "Expresso: A benchmark and analysis of discrete expressive speech resynthesis," in *INTERSPEECH 2023*. ISCA, 2023, pp. 4823–4827.

[7] J. Lian, C. Feng, N. Farooqi, S. Li, A. Kashyap, C. J. Cho, P. Wu, R. Netzorg, T. Li, and G. K. Anumanchipalli, "Unconstrained dysfluency modeling for dysfluent speech transcription and detection," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[8] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, vol. 6, p. 15, 2017.

[9] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

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

[11] 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.

[12] 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, pp. 335–359, 2008.

[13] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[14] Z. Wang, X. Wang, Q. Xie, T. Li, L. Xie, Q. Tian, and Y. Wang, "Msm-vc: High-fidelity source style transfer for non-parallel voice conversion by multi-scale style modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3883–3895, 2023.

[15] Z. Du, B. Sisman, K. Zhou, and H. Li, "Disentanglement of Emotional Style and Speaker Identity for Expressive Voice Conversion," in *Proc. Interspeech 2022*, 2022, pp. 2603–2607.

[16] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Speech Synthesis Workshop*, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:7750363

[17] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods ," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 195–202.

[18] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," *arXiv preprint arXiv:2008.12527*, 2020.

[19] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016." in *Interspeech*, 2016, pp. 1632–1636.

[20] S. Upadhyay, W.-S. Chien, B.-H. Su, L. Goncalves, Y.-T. Wu, A. Salman, C. Busso, and C.-C. Lee, "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, September 2023.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[22] E. Chodroff. Montreal Forced Aligner. [Online]. Available: https://lingmethodshub.github.io/content/tools/mfa/mfa-tutorial

[23] Q. Lemaire and A. Holzapfel, "Temporal convolutional networks for speech and music detection in radio broadcast," in *20th International Society for Music Information Retrieval Conference, ISMIR 2019, 4-8 November 2019*. International Society for Music Information Retrieval, 2019.

[24] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. INTERSPEECH 2023*, 2023.

[25] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. INTERSPEECH 2023*, 2023.

[26] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, and B. Schuller, "Speech-based age and gender prediction with transformers," 06 2023.

[27] T. Feng and S. Narayanan, "Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[28] L. Goncalves, A. N. Salman, A. Reddy Naini, L. Moro-Velazquez, T. Thebaud, L. Paola Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey2024 - speech emotion recognition challenge: Dataset, baseline framework, and results," in *Odyssey 2024: The Speaker and Language Recognition Workshop)*, vol. To appear, Quebec, Canada, June 2024.

[29] C. Kim and R. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Interspeech 2008*, Brisbane, Australia, September 2008, pp. 2598–2601.

[30] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[31] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1959–1972, 2022.

[32] C. Xie and T. Toda, "Noisy-to-noisy voice conversion under variations of noisy condition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3871–3882, 2023.

[33] H. J. Park, S. W. Yang, J. S. Kim, W. Shin, and S. W. Han, "Triaan-vc: Triple adaptive attention normalization for any-to-any voice conversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[34] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.