

INVESTIGATING SELF-SUPERVISED FEATURES FOR EXPRESSIVE, MULTILINGUAL VOICE CONVERSION

Álvaro Martín-Cortinas¹ Daniel Sáez-Trigueros¹ Grzegorz Beringer¹ Iván Vallés-Pérez¹
Roberto Barra-Chicote¹ Biel Tura-Vecino¹ Adam Gabrys¹ Thomas Merritt¹
Piotr Biliński^{1,2} Jaime Lorenzo-Trueba¹

¹Amazon AGI
²University of Warsaw

ABSTRACT

Voice conversion (VC) systems are widely used for several applications, from speaker anonymisation to personalised speech synthesis. Supervised approaches learn a mapping between different speakers using parallel data, which is expensive to produce. Unsupervised approaches are typically trained to reconstruct the input signal, which is composed of the content and the speaker information. Disentangling these components is a challenge and often leads to speaker leakage or prosodic information removal. In this paper, we explore voice conversion by leveraging the potential of self-supervised learning (SSL). A combination of the latent representations of SSL models, concatenated with speaker embeddings, is fed to a vocoder which is trained to reconstruct the input. Zero-shot voice conversion results show that this approach allows to keep the prosody and content of the source speaker while matching the speaker similarity of a VC system based on phonetic posteriorgrams (PPGs).

Index Terms— voice conversion, self-supervised learning (SSL), self-supervised speech representations (S3R)

1. INTRODUCTION

The goal of many-to-many voice conversion (VC) systems is to change the identity of a source speaker to that of a target speaker. Depending on the learning paradigm applied and the data needed for the training process, these systems can be divided in two [1]: supervised (trained with parallel data, expensive to produce) and unsupervised (trained with non-parallel data).

Typically, non-parallel VC systems make use of carefully designed bottlenecks [2], variational autoencoders [3], generative adversarial networks (GANs) [4, 5], normalizing flows [6] or intermediate representations such as text or PPGs extracted from automatic speech recognition (ASR) models [1, 7]. Nevertheless, GANs are known to be difficult to train [2], normalizing flows are text-conditioned or have a lower performance when the flow prior is trained jointly with its weights [6], bottlenecks usually drop desired information or suffer from speaker leakage even when carefully designed [2], and PPGs require external prosody information to control the prosody of the converted speech [8].

More recently, non-parallel VC models based on self-supervised speech representations (S3Rs) as intermediate features [9, 10] have been proposed. One important advantage of using S3Rs instead of PPGs as intermediate representations is that self-supervised learning (SSL) models are trained in a completely unsupervised way, whereas ASR models need curated datasets of audio and its corresponding

transcriptions. However, when using S3Rs the content and speaker information must be disentangled, in contrast with PPGs which are assumed to be speaker-independent.

Disentanglement of S3R features is usually performed by applying a bottleneck [9] or discretizing them [10]. Nevertheless, these techniques have several drawbacks. First, in [9] the bottleneck forces the removal of information from the source speech, which may not be only speaker information. Second, in [10] the discretization techniques resulted in poor intelligibility even with large codebooks. Finally, in both approaches only the last layer of all SSL models is used, even though the different layers of SSL models have been proved to contain different information [11, 12, 13, 14]. In particular, for prosody-intensive tasks, such as expressive VC, a weighted-sum of the layers of SSL models has proven to give good results [15].

Thus, this paper describes the following contributions:

- We propose a methodology for extracting content features for voice conversion as an average of carefully selected layers of WavLM [11] and HuBERT [16].
- We propose *Chameleon*, a method to automatically extract content representations from a learnable weighted average of the hidden states of an SSL model.
- We show that S3Rs trained only with English data achieve a higher intelligibility and preserve the prosody better in zero-shot multilingual VC than PPGs, even with PPGs extracted from a multilingual ASR.

2. METHODS

All VC systems in this paper are built with a similar architecture: a content encoder whose hidden layer’s outputs are combined, a speaker encoder and a decoder. At training, the models reconstruct the input signal from the content features concatenated with the speaker embedding, with the assumption that the content features do not contain any speaker information. At inference, the model performs voice conversion by taking the speaker information from an utterance of the target speaker, and the content information from an utterance of the source speaker.

For the decoder, the universal vocoder BigVGAN [17] has been selected for its capabilities to generalize to unseen speakers and languages. BigVGAN is trained with the loss computed for the generator and the discriminator as described in [17].

As for the speaker encoder, two pre-trained and frozen models have been tested. The first is a speaker verification model based on the Generalized End-to-End Loss (GE2E-Loss embeddings) for its decreased training time and speaker verification Equal Error

Rate [18], which was trained with proprietary data and Common Voice [19]. For the second speaker encoder, a pre-trained version of WavLM with an X-vector head [20] has been used, which allows to extract both the content information from WavLM hidden states and the speaker information from the X-vector head.

Content features are extracted using one of three pre-trained and frozen models: Whisper [21], WavLM [11] and HuBERT [16]. Whisper’s intermediate features (PPGs) can be used to extract precise phonetic content without speaker information, i.e. requiring no additional disentanglement. Regarding WavLM and HuBERT, although WavLM is the state-of-the-art in most downstream tasks according to the SUPERB benchmark [22], HuBERT has also been included in this study to test the architecture proposed with more than one SSL model.

In the following subsections, the different methodologies proposed to generate the content features from the hidden representations of the content encoders are described in detail.

2.1. Whisper

Various VC systems based on phonetic posteriorgrams (PPGs) extracted from ASR systems can be found in the literature [23, 24, 25]. In this paper we train baseline VC models for English and multilingual setups using PPGs extracted using Whisper as ASR. Whisper is selected to obtain these intermediate representations as it is a multilingual ASR that has proven to be robust across various datasets and languages.

Whisper is built upon a transformer architecture [26], where a latent representation of the spectrogram is derived by the transformer encoder. Thus, in the models trained with Whisper as content encoder, the hidden states of the last layer of its transformer encoder are used as content features, which would be equivalent to PPGs or bottleneck features (BNFs) [10]. These features are not expected to contain speaker information, as it is not needed for the ASR task, hence no further disentanglement is needed.

For the monolingual case, only Whisper base (74M parameters) is considered. For the multilingual case, both Whisper-base and Whisper large v2 (1550M parameters) are considered to check if there are any significant differences in performance for languages where less training data was available.

2.2. WavLM and HuBERT with a fixed average

The authors of [11, 12] analyse and report the importance of each layer in different SSL models for different downstream tasks. Intuitively, the results show that higher layers are more related to abstract concepts, such as words, whereas lower layers are more related to local signal properties and low-level speech characteristics, such as speaker identity.

Consequently, useful content features should be extracted from hidden states of the last layers, since first layers most likely contain the most speaker information. For that reason, a fixed average of layers 8 to 12 in WavLM base+, and 7 to 12 in HuBERT base, have been used as a first approach to obtain the content features. These particular layers have been selected according to Figure 2 in [11].

2.3. WavLM with a learned weighted average (Chameleon)

Carefully selecting layers in self-supervised learning models is a costly method, as it requires to evaluate the importance of each layer in different downstream tasks and trying different combinations of them to find the best configuration. Thus, learning which hidden

states are the most important for the task at hand is a more scalable and general approach.

We propose a novel model, Chameleon, that generates the content features from the self-supervised hidden states by learning, per dimension, the linear combination of the layers of the SSL model that minimizes the decoder loss, and maximizes the disentanglement with the speaker embeddings.

To enforce the disentanglement, a L2 distance adversarial loss with gradient reversal [27] is added between the pre-trained speaker embedding and a predicted speaker embedding conditioned with the content features. Intuitively, to maximize the L2 distance the model will learn a weighted average of hidden states that cannot be used to predict the pre-trained speaker embedding. Figure 1 illustrates the architecture of Chameleon during training. At inference, the L2 distance is no longer needed.

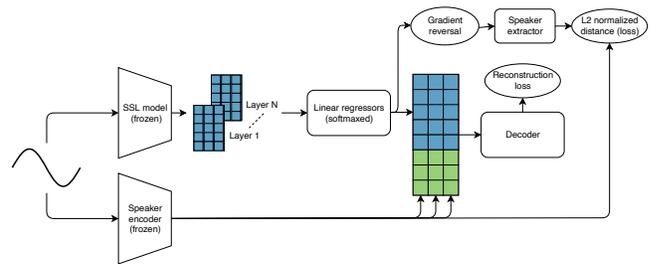


Fig. 1: Chameleon architecture during training.

Mathematically, the loss of the generator [17] is modified to be:

$$\mathcal{L}_G = \mathcal{L}_{BigVGAN} + \lambda_{L2}(w) \left\| \vec{s} - \vec{\hat{s}} \right\|^2, \quad (1)$$

where \vec{s} is the pre-trained speaker embedding, $\vec{\hat{s}}$ is the predicted speaker embedding and $\lambda_{L2}(w)$ is the weight of the L2 distance in the generator loss. $\lambda_{L2}(w)$ is a function of the parameter w in consideration, as it is positive for the speaker extractor parameters, negative for the linear regressors (due to the gradient reversal) and zero for the decoder parameters (they do not contribute to the L2 distance).

The speaker extractor that predicts the pre-trained speaker embedding is based on a transformer encoder with a CLS embedding [28]. No positional embedding is added to prevent the transformer from learning the content. This approach assumes that the speaker identity is independent of the ordering of the frames. After the transformer encoder, the CLS is linearly projected to match the dimensionality of the speaker embedding.

3. EXPERIMENTS AND RESULTS

3.1. Naming convention

In this section, the trained models are presented with the following naming convention:

- VCWhisper base and VCWhisper large v2 refer to the VC models trained with Whisper base and large v2 as content encoder (see Section 2.1), respectively, and the speaker verification model with GE2E-Loss embeddings.
- VCWavLM base+ and VCWavLMX base+ refer to VC models trained both with WavLM base+ as content encoder with a fixed average of layers 8 to 12 (see Section 2.2), using as speaker embeddings GE2E-Loss embeddings for the former and x-vectors for the latter.

- VCHuBERT base refers to a VC model trained with HuBERT base as content encoder with a fixed average of layers 7 to 12 (see Section 2.2) and GE2E-Loss embeddings as speaker embeddings.
- Chameleon refers to the model described in Section 2.3. In these experiments, WavLM base+ is used as the SSL model, because it is better than HuBERT in SUPERB [22], and x-vectors are used as speaker embeddings for simplicity because they can be directly extracted from WavLM base+.

All models have been trained with 8 NVIDIA Tesla V100 SXM2 16GB GPUs with an average training time of 3 days.

3.2. Evaluation metrics

Models are evaluated with both objective and subjective metrics. Subjective evaluations are conducted by 100 testers and 10 submissions per tester to evaluate the speaker similarity and the naturalness of the synthesized speech. Paired t-tests are used to determine whether there are significant differences between the models, which require a corrected p-value using Holm method [29] below 0.05. Objective evaluations consist of two metrics: the word error rate (WER), as a proxy of the intelligibility, and the F0 correlation between the source and the converted utterances, as a proxy of the extent to which the prosody of the source utterance is kept.

3.3. Monolingual experiments

Experimental setup. English models have been trained with LibriTTS [30] train-other-500 subset, which is composed of 310 hours of speech and 1160 speakers. To test English models, the source utterances are extracted from LibriTTS test-other subset, which contains 6.69 hours of speech with 33 different speakers, and the target speakers are extracted from LibriTTS test-clean, which contains 8.56 hours of speech and 39 speakers.

Speech intelligibility. The transcriptions for the WER computation have been generated using Whisper base, with approximately 4,000 utterances. Table 1 shows that all SSL-based models achieve a lower WER than the baseline created with Whisper, and in particular the model based on WavLM with the x-vector head achieves the lowest WER.

Table 1: WER for English-only models with LibriTTS test-other.

Utterances	WER
Source audio (no VC)	11.5%
VCWhisper base	14.9% (+3.4%)
VCHuBERT base	13.6% (+2.1%)
VCWavLM base+	13% (+1.5%)
VCWavLMX base+	12.2% (+0.7%)
Chameleon	12.4% (+0.9%)

Prosody. Table 2 illustrates the F0 correlation with the source utterance for each model. The results show that SSL-based models keep the prosody information of the source utterance better than the model based on Whisper.

MUSHRA speaker similarity and naturalness. The tests are composed of 100 testcases, where 15% are male-to-female, 15% female-to-male, 30% are the utterances with the highest WER, and the rest are randomly chosen. Cross-gender examples are included because they are the most difficult for speaker similarity, whereas the samples with the highest WER are expected to be the most complicated

Table 2: F0 correlation to source utterance per monolingual model with 95% confidence intervals.

Model	F0 correlation
VCWhisper base	59.3 \pm 0.9
VCHuBERT base	67.3 \pm 0.9
VCWavLM base+	65.7 \pm 1.0
VCWavLMX base+	64.7 \pm 1.0
Chameleon	62.9 \pm 0.9

in terms of intelligibility. Table 3 illustrates that SSL-based models achieve the same naturalness and speaker similarity as the baseline Whisper.

Table 3: Naturalness and speaker similarity with 95% confidence intervals for English models trained with GE2E-Loss embeddings.

Utterances	Naturalness	Speaker similarity
Ground Truth (GT)	72.2 \pm 1.4	76.6 \pm 1.5
VCWhisper base	69.9 \pm 1.5	73.5 \pm 1.7
VCHuBERT base	70.0 \pm 1.4	72.4 \pm 1.7
VCWavLM base+	70.4 \pm 1.5	73.2 \pm 1.7

Similar tests were conducted to compare Chameleon, which learns the weighted average of WavLM hidden states, with VCWavLM base+ and VCWavLMX base+, which perform a fixed average. The results, which are not included due to space restrictions, show that the three models have no significant difference in terms of speaker similarity and naturalness. Finally, Figure 2 shows that Chameleon has learned to give more weight to the hidden states of layers 8 to 12, the same layers manually selected for the fixed average in VCWavLM base+ and VCWavLMX base+.

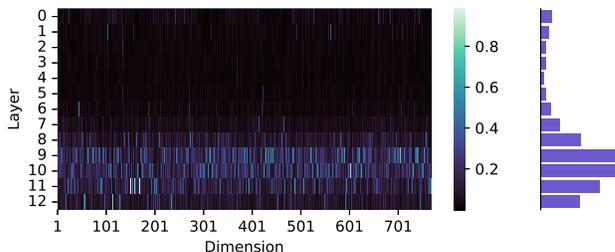


Fig. 2: Chameleon weights with the histogram per layer.

3.4. Multilingual experiments

Experimental setup. Multilingual models have been trained with a balanced subset of 200,000 utterances from the train split of Multilingual LibriSpeech (MLS) [31], where 25,000 were included per language: English, German, Dutch, Spanish, French, Italian, Portuguese and Polish. Validation and test subsets are created from MLS development and test splits respectively, with a total of 4,000 utterances each, 500 utterances per language. For testing, the target speakers are all English speakers from LibriTTS test-clean.

Speech intelligibility. The transcriptions for the WER computation have been generated using Whisper large v2, as it has a significantly

Table 4: WER per language for multilingual models with LibriTTS test-other.

Utterances	Dutch	English	French	German	Italian	Polish	Portuguese	Spanish
Source audio (no VC)	9.2%	7.7%	7.8%	6.3%	13.3%	5.6%	8.4%	5.4%
VCWhisper base	11.5%	9.4%	9.2%	8.6%	15.8%	7.9%	10.9%	6.0%
VCWhisper large v2	11.4%	9.7%	9.1%	8.5%	16.0%	7.6%	10.3%	6.7%
VCHuBERT base	10.1%	8.4%	8.6%	7.4%	14.8%	6.5%	9.8%	6.1%
VCWavLMX base+	10.0%	8.5%	8.4%	7.1%	14.6%	6.3%	10.1%	5.4%
Chameleon	9.6%	8.3%	8.6%	6.8%	14.4%	6.3%	9.3%	6.0%

better performance than Whisper base in languages different to English. Table 4 shows that, as in the English-only case, all SSL-based models achieve a lower WER in most locales than the baselines built with Whisper. The results obtained with Chameleon are particularly remarkable, as it is the model with the lowest WER in 6 out of 8 locales. These results illustrate that even though WavLM was trained with English data only, it can successfully generalise to other locales. **Prosody.** Table 5 illustrates the F0 correlation for each model considering all locales. As in the English-only case, the results show that SSL-based models keep the prosody information of the source utterance better than the model based on Whisper.

Table 5: F0 correlation to source utterance with 95% confidence intervals per multilingual model.

Model	F0 correlation
VCWhisper base	56.4 ± 0.8
VCWhisper large v2	58.0 ± 0.8
VCHuBERT base	68.4 ± 0.7
VCWavLMX base+	68.3 ± 0.8
Chameleon	72.8 ± 1.1

MUSHRA speaker similarity and naturalness. Similarly to the monolingual case, MUSHRA tests were conducted to evaluate the speaker similarity and naturalness of the converted utterances with the different models. In particular, Chameleon, VCWavLMX base+ and Whisper large v2 are evaluated in English (to compare with the monolingual case), Italian (high WER) and Spanish (low WER). These models are selected for the MUSHRA test because Chameleon and VCWavLMX have the lowest WER in different locales, and Whisper large v2 is significantly better than Whisper, in F0 correlation and WER in most locales.

As in the monolingual case, in English the results show that the proposed systems have no differences in terms of these two characteristics. Nevertheless, both in Spanish and Italian the SSL-based models are significantly better than VCWhisper large v2 in terms of naturalness. In terms of speaker similarity, Chameleon has a significantly lower speaker similarity in Italian than the other models because the gradient reversal weight in the total loss has to be carefully fine-tuned to reach an optimal value, which is not trivial in multilingual datasets. Table 6 shows the results for Italian, and they are omitted for Spanish and English due to space restrictions.

4. DISCUSSION

SSL models outperform ASR models as content encoders both in prosody and intelligibility, as well as naturalness for some languages such as Spanish and Italian. At the same time, they provide equal speaker similarity. As SSL models encode most of the information

Table 6: Naturalness and speaker similarity with 95% confidence intervals for Italian.

Utterances	Naturalness	Speaker similarity
GT	65.1 ± 1.7	75.9 ± 1.7
VCWhisper large v2	53.9 ± 1.7	35.4 ± 1.8
VCWavLMX base+	57.1 ± 1.7	35.7 ± 1.8
Chameleon	57.3 ± 1.7	31.8 ± 1.7

in the source speech, they contain prosodic information which makes the reconstruction task much simpler and improves the final naturalness and intelligibility of the system. In contrast, ASR models only encode the information relevant for the transcription task, so the prosodic information has to be inferred in detriment of intelligibility and naturalness.

Nevertheless, in SSL models content and speaker information must be disentangled using an appropriate combination of the model’s hidden states which discards those more related to speaker identity. If the disentanglement is not done correctly, the speaker information of the source utterance is used during training to reconstruct the input, and at inference that results in speaker leakage.

5. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we have proposed computing disentangled content features by carefully averaging the hidden states in different layers of SSL models, either with a fixed average or a learnable weighted average, i.e. Chameleon. The main drawback of performing a fixed average is to previously determine which layers of the SSL model are more related to content, e.g. with downstream tasks. Chameleon’s learning paradigm automatically determines those layers, but as a downside the gradient reversal has to be fine-tuned to avoid speaker leakage. In future work, speaker embeddings could also be learned from SSL features in an unsupervised framework similar to Chameleon, forcing the model to separate SSL features into content or speaker features.

6. REFERENCES

- [1] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li, “An overview of voice conversion and its challenges: From statistical modeling to deep learning,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.
- [2] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, “AutoVC: Zero-shot voice style transfer with only autoencoder loss,” in *Proceedings of ICML*, 2019.

- [3] Yuki Saito, Yusuke Ijima, Kyosuke Nishida, and Shinnosuke Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *ICASSP*, 2018.
- [4] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *ICASSP*, 2019.
- [5] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks," in *SLT*, 2018.
- [6] Tom Merritt, Abdelhamid Ezzer, Piotr Bilinski, Magdalena Proszewska, Kamil Pokora, Roberto Barra-Chicote, and Daniel Korzekwa, "Text-free non-parallel many-to-many voice conversion using normalising flows," in *ICASSP*, 2022.
- [7] Kadria Ezzine, Mondher Frikha, and Joseph Di Martino, "Non-parallel voice conversion system using an auto-regressive model," in *International Conference on Advanced Systems and Emergent Technologies (ICASET)*, 2022.
- [8] Zheng Lian, Rongxiu Zhong, Zhengqi Wen, Bin Liu, and Jianhua Tao, "Towards fine-grained prosody control for voice conversion," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021.
- [9] Yist Y. Lin, Chung-Ming Chien, Jheng-Hao Lin, Hung-yi Lee, and Lin-shan Lee, "Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," in *ICASSP*, 2021.
- [10] Wen-Chin Huang, Shu-Wen Yang, Tomoki Hayashi, and Tomoki Toda, "A comparative study of self-supervised speech representation based voice conversion," *IEEE Journal of Selected Topics in Signal Processing*, 2022.
- [11] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Michael Zeng, Xiangzhan Yu, and Furu Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.
- [12] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.
- [13] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," in *Interspeech*, 2021.
- [14] Yuanchao Li, Yumnah Mohamied, Peter Bell, and Catherine Lai, "Exploration of a self-supervised speech model: A study on emotional corpora," in *SLT*, 2022.
- [15] Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G. Ward, "On the utility of self-supervised models for prosody-related tasks," in *SLT*, 2022.
- [16] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.
- [17] Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon, "BigVGAN: A universal neural vocoder with large-scale training," in *International Conference on Learning Representations*, 2023.
- [18] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*, 2018.
- [19] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *International Conference on Language Resources and Evaluation*, 2019.
- [20] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018.
- [21] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *arXiv*, 2022.
- [22] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdel rahman Mohamed, and Hung yi Lee, "Superb: Speech processing universal performance benchmark," in *Interspeech*, 2021.
- [23] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *ICME*, 2016.
- [24] Luis Serrano, Sneha Raman, David Tavaréz, Eva Navas, and Inma Hernaez, "Parallel vs. Non-Parallel Voice Conversion for Esophageal Speech," in *Interspeech*, 2019.
- [25] Chengxin Chen, Wei-Zhong Zheng, Syu-Siang Wang, Yu Tsao, Pei-Chun Li, and Ying-Hui Lai, "Enhancing intelligibility of dysarthric speech using gated convolutional-based voice conversion system," in *Interspeech*, 2020.
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.
- [27] Bastian Schnell, Goeric Huybrechts, Bartek Perz, Thomas Drugman, and Jaime Lorenzo-Trueba, "EmoCat: Language-agnostic Emotional Voice Conversion," in *ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 72–77.
- [28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.
- [29] Sture Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 1979.
- [30] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Interspeech*, 2019.
- [31] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech*, 2020.