SPEAKER ANONYMIZATION USING NEURAL AUDIO CODEC LANGUAGE MODELS

Michele Panariello,[1] Francesco Nespoli,[2] Massimiliano Todisco,[1] Nicholas Evans[1]

[1]EURECOM, France   [2]Microsoft UK

ABSTRACT

The vast majority of approaches to speaker anonymization involve the extraction of fundamental frequency estimates, linguistic features and a speaker embedding which is perturbed to obfuscate the speaker identity before an anonymized speech waveform is resynthesized using a vocoder. Recent work has shown that x-vector transformations are difficult to control consistently: other sources of speaker information contained within fundamental frequency and linguistic features are re-entangled upon vocoding, meaning that anonymized speech signals still contain speaker information. We propose an approach based upon neural audio codecs (NACs), which are known to generate high-quality synthetic speech when combined with language models. NACs use quantized codes, which are known to effectively bottleneck speaker-related information: we demonstrate the potential of speaker anonymization systems based on NAC language modeling by applying the evaluation framework of the Voice Privacy Challenge 2022.

Index Terms— Speaker anonymization, neural audio codec, language modeling

1. INTRODUCTION

Speaker anonymization involves the task of processing a speech signal to conceal the identity of the speaker while retaining the spoken content and other para-linguistic attributes such as intonation and prosody. As defined by the Voice Privacy Challenge [1], a speaker anonymization system should provide a certain trade-off between *privacy protection* and *utility preservation*. The former is measured by the difficulty of an attacker to recover the identity of the original speaker from an anonymized signal via automatic speaker verification (ASV). The latter is assessed primarily by the reliability of an automatic speech recognition (ASR) system to transcribe the anonymized waveform, among other, secondary metrics such as pitch correlation and gain of voice distinctiveness [2].

Most speaker anonymization systems are based on often-incremental changes to original work in [3], which operates upon three distinct components extracted from an input waveform: an F0 curve encoding prosodic information; some form of linguistic features encoding the spoken content; an x-vector [4] encoding the speaker identity. The x-vector is perturbed to conceal the identity of the speaker and then fed to a vocoder with the other two components in order to synthesize an anonymized waveform. This approach assumes that speaker information is contained entirely within the x-vector, even if this is known not to be the case [5, 6]. Residual speaker information captured in linguistic features and the F0 curve is re-entangled with the anonymized x-vector upon vocoding. An x-vector extracted afresh from the anonymized utterance then still contains speaker information which can be used by an adversary to reverse the anonymization and re-identify the speaker [5, 7]. Other researchers [5] have found that the level of speaker information contained in linguistic features can be reduced through their quantization.

Motivated by their successful application to numerous audio synthesis tasks [8, 9], we have sought to exploit the potential of neural audio codec (NAC) language modeling to design a speaker anonymization system that better suppresses speaker information and hence provides an improved trade-off between anonymization and utility. Such an approach is appealing since linguistic features are not used directly for waveform synthesis, as with previous approaches: instead, they are used to infer a set of NAC acoustic tokens with a language model. These features are quantized and therefore have the potential to bottleneck speaker information and improve anonymization. The final waveform is synthesized by decoding the acoustic tokens with a decoder neural network. We hope that the representational power of NACs should help to preserve speech quality and utility.

2. RELATED WORK

In the following we describe some related research which provided motivation for the work presented in this paper.

X-vector–based speaker anonimization - The original x-vector–based pipeline introduced in [3] is the basis of much of the related work reported recently. An approach to dispense with the intermediate acoustic model was proposed in [2]. More refined x-vector anonymization functions were proposed in later work [10, 11], with some [12] achieving notable improvements to privacy protection levels, albeit under the assumption that the attacker does not have full knowledge of the anonymization system. Whatever the approach, x-vector perturbation does not prevent speaker-related information contained in the F0 curve and linguistic features

from *leaking* into the anonymized waveform upon vocoding. As a result, the x-vector which can be re-extracted from the anonymized waveform by a privacy adversary who wishes to reidentify the speaker tends to *drift* away from that at the vocoder input [13]. While the drift can be beneficial to anonymization, it hinders the design of more effective anonymization functions and can also be inverted by an adversary to undo the privacy protection [14]. Attempts to sanitize speaker information from linguistic features have also been explored, e.g. [7] based on the concept of differential privacy, which reports improvements to privacy at the cost of reduced utility and pitch correlation. The same issue was tackled in [5] by means of feature quantization, which is shown to be effective as a speaker information bottleneck, though with a degradation to utility. In this paper, we follow a similar approach, but propose a completely new synthesis pipeline in which we avoid the leakage of information between different speech components by design.

NAC language modeling - NACs were proposed recently for audio compression [15, 16]. They consist in convolutional autoencoders that compress audio to low-bitrate, tokenized representations which support high-fidelity decoding. Due to their discretized nature, encoded representations can be modeled using techniques normally used for language-related tasks, such as transformers. This idea was first introduced in [8] for a variety of audio generation tasks, and tailored to text-to-speech (TTS) in [9]. A transformer is used to convert the input (be it audio or text) to a set of high-level *semantic tokens*. These are then fed to another transformer which converts them into *NAC acoustic tokens* which can be decoded to resynthesize an audio signal.

The same technique can also be applied to voice conversion [8, 17], and is hence ideally suited to speaker anonymization. NAC language models operate on quantized codes, which are known to be beneficial to privacy protection [5]. Moreover, such models appear to naturally disentangle linguistic information into semantic tokens, while encoding speaker information and recording conditions mostly in acoustic tokens [8]. Hence, we propose a speaker anonymization system whereby an input utterance is re-synthesized by means of a NAC language model. Semantic tokens are kept unchanged, while acoustic tokens are substituted with those of a different speaker, the goal being to preserve the linguistic content of the signal while suppressing information related to the original speaker.

3. PROPOSED APPROACH

3.1. Neural audio codec language modeling

A diagram of the proposed system is shown in Figure 1. Following [9], it is comprised of a semantic encoder, a NAC (encoder and decoder), a pair of transformers and a pool of speaker prompts. They are described in the following.
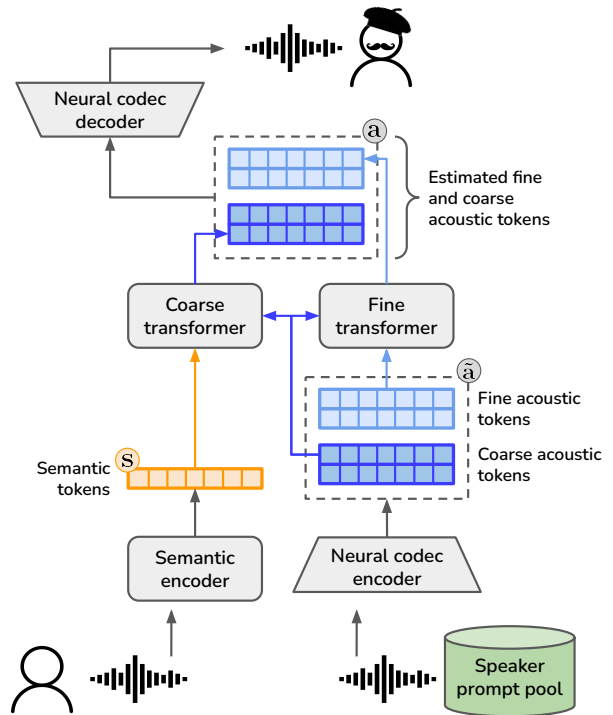

Fig. 1. Diagram of the proposed anonymization system.

The semantic encoder produces high-level semantic representations of the input signal using a codebook of N_S quantized embeddings. The output is a sequence of integers $\mathbf{s} \in \{1, \ldots, N_S\}^{T_S}$, where T_S is the number of frames and where each integer is a codeword index.

The NAC is an encoder-decoder architecture. The encoder maps input waveforms to a quantized, compressed representation from which the decoder reconstructs a high-fidelity waveform. Efficient compression is achieved with a set of Q hierarchical codebooks. Lower level codebooks capture coarser waveform characteristics, while finer details are captured by higher level codebooks. Following [8, 9], we refer to the first Q_C codebooks as 'coarse codebooks', and to the last $Q - Q_C$ codebooks as 'fine codebooks'. All have N_Q codewords so that the output of the encoder is $\tilde{\mathbf{a}} \in \{1, \ldots, N_Q\}^{Q \times T_A}$, where T_A is the number of frames into which the input is divided.

The coarse and fine transformers estimate a set of acoustic tokens $\mathbf{a}$ from a prompt of input semantic tokens $\mathbf{s}$ and acoustic tokens $\tilde{\mathbf{a}}$. Essentially, the transformers attempt to predict what semantic information should 'sound like' in the domain of quantized acoustic tokens. The coarse transformer autoregressively predicts coarse acoustic tokens, i.e. the codewords belonging to the coarse codebooks. More specifically, for frame t, the transformer predicts the probability distribution of token $\mathbf{a}_{q,t}$ conditioned on the following elements: the semantic prompt $\mathbf{s}$, the coarse tokens from the

acoustic prompt $\tilde{\mathbf{a}}_{<Q_C,:}$, and all previous predictions.[1] The modeled distribution is therefore

$$p\left(\mathbf{a}_{q,t}|\mathbf{s}, \tilde{\mathbf{a}}_{<Q_C,:}, \mathbf{a}_{<Q_C,<t}, \mathbf{a}_{<q,t}\right) \qquad (1)$$

for $q \in [1, Q_C]$. The fine transformer is instead non-autoregressive. It estimates the tokens of codebook q using all tokens belonging to codebooks $< q$ and all tokens of the acoustic prompt $\tilde{\mathbf{a}}$, thus modeling the distribution

$$p\left(\mathbf{a}_{q,:}|\tilde{\mathbf{a}}, \mathbf{a}_{<q,:}\right) \qquad (2)$$

for every $q \in [Q_C + 1, Q]$. Once the acoustic tokens $\mathbf{a}$ have been predicted for all codebooks $q \in [1, Q]$, they can be input into the decoder to synthesize an anonymized waveform.

The pool of speaker prompts is a set of acoustic tokens extracted by the NAC encoder from utterances belonging to a set of external speakers. Those speakers are referred to as *pseudo-speakers*, since they replace the original speaker in the anonymized utterance. As suggested in [8], acoustic tokens, especially the coarse tokens, can capture information related to the speaker identity. We use them to perform voice conversion, as detailed in the following.

3.2. Anonymization technique

A set of semantic tokens $\mathbf{s}$ is first extracted from the input utterance. They encode the high-level spoken content. Their quantization helps to suppress speaker-related information.

A pseudo-speaker is chosen by randomly selecting an acoustic prompt $\tilde{\mathbf{a}}$ from the speaker prompt pool. Anonymization can be performed at either speaker or utterance levels. At the *speaker level*, anonymization is performed using the same speaker prompt for each utterance corresponding to any one speaker. In contrast, for *utterance level* anonymization, a speaker prompt is selected at random for each utterance. While several anonymization systems include techniques to synthesize fictitious voices [5, 10, 11, 12], here we use real voices as pseudo-speakers to focus our analysis on the intrinsic anonymization capability of the NAC language model.

Prompted with $\mathbf{s}$ and $\tilde{\mathbf{a}}$, the coarse and fine transformers generate a set of acoustic tokens $\mathbf{a}$ which reflect the semantic information of the original utterance, but the acoustic characteristics of the pseudo-speaker. Acoustic tokens $\mathbf{a}$ are fed to the decoder which synthesizes the anonymized output waveform.

4. EXPERIMENTAL SETUP

Our codebase is branched from Bark,[2] an open-source, NAC-based TTS system which is very similar to VALL-E [9].

The modules described in Section 3.1 are all taken from Bark. The semantic encoder has a HuBERT backbone [19] and a LSTM [20] back-end which predicts the semantic token associated to the HuBERT feature vector output at each frame. The semantic dictionary is of size $N_S = 10000$. The coarse and fine transformers are 12-layer GPT-like models [21], each having $Q = 8$ different codebooks with $N_Q = 1024$ codewords. The first $Q_C = 2$ codebooks are considered coarse. The NAC is EnCodec [16]. The difference between Bark and our system is that, being a TTS model, Bark estimates semantic tokens $\mathbf{s}$ corresponding to an input text using a further *semantic transformer*. In our implementation, we bypass the semantic transformer and use ground truth semantic tokens from the input waveform thereby providing for voice conversion instead of TTS. With this setup, we are able to use pre-trained Bark modules off-the-shelf, without the need for any training.

We adopt the Voice Privacy Challenge 2022 protocol [2] for evaluation. The test set comprises subsets of the *LibriSpeech* [22] and *VCTK* [23] databases. The pool of speaker prompts is taken from the Bark voice library. It consists of 130 utterances collected from speakers of different gender and nationality.[3] The threat model is the *semi-informed* attack described in [2]. Trial utterances are anonymized at the *speaker level*. The attacker is assumed to have access to the anonymization system. They anonymize a set of external data (librispeech-clean-360) at the *utterance level* and use it to train an ASV system (a TDNN with a PLDA back-end [24]). They also have access to original (non-protected) enrollment utterances which they anonymize at the *speaker level*. The attacker thus has enrollment and trial utterances both of which are anoymized and uses an ASV system to verify whether they correspond to the same speaker. The attacker has no knowledge of which pseudo-speaker was used for anonymization on the test utterance and will hence likely select a different pseudo-speaker to anonymize the enrollment utterance. The privacy metric is the resulting EER estimated from a large number of ASV trials. Utility is assessed by training an ASR system on the same anonymized version of librispeech-clean-360 and by estimating the word error rate (WER) from anonymized test data. Two additional metrics are defined in the VoicePrivacy Challenge evaluation plan [2]. The first is the F0 curve correlation ρ^{F0} between original and anonymized utterance which is used as a measure of prosody preservation. The second is the gain of voice distinctiveness G_{VD} which is used to estimate how well the anonymized voices of different speakers can be distinguished [2].

We adopt the B1b and T11 participant system from the Voice Privacy Challenge 2022, in addition to the system pro-

---

[1]In practice, the sequence upon which to perform regression is flattened to $\left(\mathbf{s}, \tilde{\mathbf{a}}, \mathbf{a}_{1,1}, \mathbf{a}_{2,1}, \ldots, \mathbf{a}_{Q_C,1}, \mathbf{a}_{1,2}, \mathbf{a}_{2,2}, \ldots, \mathbf{a}_{Q_C,2}, \ldots\right)$. See [8, 9] for more details.

[2]The original source code is available at `www.github.com/suno-ai/bark`, though we built our system from the port included in the

CoquiTTS library available at `www.github.com/coqui-ai/TTS`. Our source code, as well as audio samples, will be made available upon publication.

[3]`www.github.com/suno-ai/bark/tree/main/bark/assets/prompts/v2`

System	LibriSpeech				VCTK			
	EER (%)	WER (%)	G_{VD}	ρ^{F_0}	EER (%)	WER (%)	G_{VD}	ρ^{F_0}
Original	4.4	4.2	0	1	3.2	12.8	0	1
Original (eval. pipeline of [18])	1.5	2.5	n.a.		1.1	7.6	n.a.	
B1b [2]	8.6	4.4	-5.8	0.78	9.7	10.7	-7.1	0.81
T11 [10]	20.6	3.9	-19.0	0.68	39.7	7.9	-18.4	0.73
Champion et al. [5]	17.5	4.5	n.a.	0.67	28.0	10.0	n.a.	0.73
Champion et al. (noise on F0) [5]	23.4	4.6	n.a.	0.52	40.8	10.3	n.a.	0.60
Ours	28.5	7.5	-1.5	0.68	45.5	18.9	-2.1	0.74
Ours (eval. pipeline of [18])	34.1	4.6	n.a.		36.6	15.5	n.a.	

Table 1. Results of the analyzed systems on the Voice Privacy Challenge 2022 test subsets.

posed by Champion et al. [5] as baselines. System T11 is the non-TTS system that achieved the highest privacy level in the 2022 challenge.[4] The work in [5] was the first to propose the use of codebook-based feature quantization.

5. RESULTS

Results are shown separately for LibriSpeech and VCTK test sets in Table 1. Our system achieves the highest privacy levels: 28.5% EER for LibriSpeech; 45.5% EER for VCTK. The substantially lower EERs of 17.5% and 28.0% for the two test sets and the system of Champion et al. suggest that our quantization approach is more effective in removing speaker information than that proposed in [5]. In an effort to further suppress speaker information, Champion et al. also experimented with the addition of Gaussian noise to the input F0 curve. Improvements to privacy nonetheless result in a lower pitch correlation $\rho^{F_0} \approx 0.55$.[5] For our method, the pitch correlation is in the order of $\rho^{F_0} \approx 0.7$ on average, and compares favorably with that of other systems in the literature [2]. In terms of privacy protection, our model also comfortably outperforms the T11 system by 8% and 5% EER for LibriSpeech and VCTK test sets respectively. The gain in voice distinctiveness for the T11 system are also low. This is not surprising since the system maps all speakers to similar pseudo-speakers. In contrast, our system gives values of $G_{VD} \approx -2$, denoting substantially better speaker distinctiveness.

However, utility estimates for our model are lower than that of other systems. The WER increase from 4.2% (original data) to 7.5% for the LibriSpeech subset and from 12.8% to 18.9% for the VCTK subset. Similar issues have also been reported in the literature. The authors of [8] show that the NAC copy-synthesis of LibriSpeech test-clean dataset causes

an increases to the WER of its own ASR system from 2.5% to 6%, with similar results being reported in [9]. Nevertheless, informal listening tests on our data do not reveal any notable artifacts or degradation to intelligibility.

In an attempt to shed light on the cause for this phenomenon, we repeated similar experiments using a different ASR architecture, namely that reported in [18], which is retrained according to the same setup described in Section 4. The issue persists. The WER increases from 2.5% to 4.6% for the LibriSpeech subset and from 7.6% to 15.5% for the VCTK subset. These findings suggest that the degradation to utility is more dependend on the NAC language model than on the ASR system. As suggested in [8], this could be due to the quality of some pseudo-speaker prompts, since the extracted fine acoustic tokens tend also to capture aspects of the (potentially poor) *recording conditions*, the characteristics of which are then transferred to anonymized outputs. More thorough experimentation to help us better understand this phenomena is already underway.

6. CONCLUSIONS

We present a novel approach to speaker anonymization based on a neural audio codec (NAC) language model. Our system performs voice conversion by extracting a set of semantic tokens from an input signal and using them to estimate a set of acoustic tokens belonging to a different speaker, which in turn are used to synthesize an anonymized speech signal with a NAC decoder. The quantized nature of the semantic and acoustic tokens successfully bottlenecks speaker-related information delivering substantially improved anonymization performance without compromising prosody or speaker distinctiveness. While informal listening tests show that anonymized signals are of high quality and intelligibility, automatic transcription with a speech recognition system shows a modest degradation to utility. Future work should investigate strategies to better protect utility while retaining the benefits to privacy safeguard, such as using high-quality speaker prompts or fine-tuning parts of the system with utility preservation constraints.

---

[4]Results available at `www.voiceprivacychallenge.org/results-2022`. The overall highest privacy level was in fact achieved by a TTS-based system [11] that barely passed the prosody preservation requirement of scoring $\rho^{F_0} > 0.3$. In general, TTS-based systems are known to almost completely erase speaker information, at the cost of a severe loss of intonation and prosody. Therefore, we do not include [11] in our comparative analysis.

[5]This result is provided by courtesy of the main author of [5].

7. REFERENCES

[1] N. Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech 2020*, 2020, pp. 1693–1697.

[2] Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Hubert Nourtel, Pierre Champion, Massimiliano Todisco, Emmanuel Vincent, Nicholas Evans, Junichi Yamagishi, and Jean-François Bonastre, "The voiceprivacy 2022 challenge evaluation plan," 2022.

[3] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre, "Speaker Anonymization Using X-vector and Neural Waveform Models," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 155–160.

[4] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[5] Pierre Champion, Anthony Larcher, and Denis Jouvet, "Are disentangled representations all you need to build speaker anonymization systems?," in *Interspeech 2022*. Sept. 2022, pp. 2793–2797, ISCA.

[6] Pierre Champion, Denis Jouvet, and Anthony Larcher, "A Study of F0 Modification for X-Vector Based Speech Pseudonymization Across Gender," 2021.

[7] Ali Shahin Shamsabadi, Brij Mohan Lal Srivastava, Aurélien Bellet, Nathalie Vauquier, Emmanuel Vincent, Mohamed Maouche, Marc Tommasi, and Nicolas Papernot, "Differentially private speaker anonymization," *Proceedings on Privacy Enhancing Technologies*, vol. 1, pp. 98–114, 2023.

[8] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour, "AudioLM: a Language Modeling Approach to Audio Generation," July 2023, arXiv:2209.03143 [cs, eess].

[9] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," Jan. 2023, arXiv:2301.02111 [cs, eess].

[10] Jixun Yao, Qing Wang, Li Zhang, Pengcheng Guo, Yuhao Liang, and Lei Xie, "NWPU-ASLP System for the VoicePrivacy 2022 Challenge," in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022.

[11] Sarina Meyer, Pascal Tilli, Florian Lux, Pavel Denisov, Julia Koch, and Ngoc Thang Vu, "Cascade of phonetic speech recognition, speaker embeddings gan and multispeaker speech synthesis for the VoicePrivacy 2022 Challenge ," in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022.

[12] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia Tomashenko, "Language-independent speaker anonymization using orthogonal householder neural network," *arXiv preprint arXiv:2305.18823*, 2023.

[13] Michele Panariello, Massimiliano Todisco, and Nicholas Evans, "Vocoder drift compensation by x-vector alignment in speaker anonymisation," in *Proc. 3rd Symposium on Security and Privacy in Speech Communication*, 2023.

[14] Michele Panariello, Massimiliano Todisco, and Nicholas Evans, "Vocoder drift in x-vector–based speaker anonymization," in *Proc. INTERSPEECH 2023*, 2023, pp. 2863–2867.

[15] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 30, pp. 495–507, Nov. 2021.

[16] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High Fidelity Neural Audio Compression," Oct. 2022, arXiv:2210.13438 [cs, eess, stat].

[17] Zhichao Wang, Yuanzhe Chen, Lei Xie, Qiao Tian, and Yuping Wang, "Lm-vc: Zero-shot voice conversion via speech generation based on language models," 2023.

[18] Francesco Nespoli, Daniel Barreda, Jörg Bitzer, and Patrick A. Naylor, "Two-Stage Voice Anonymization for Enhanced Privacy," in *Proc. INTERSPEECH 2023*, 2023, pp. 3854–3858.

[19] 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 Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 3451–3460, oct 2021.

[20] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, nov 1997.

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[22] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[23] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," *University of Edinburgh. The Centre for Speech Technology Research*, 2019.

[24] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech 2015*, 2015, pp. 3214–3218.