Privacy-preserving Prosody Representation Learning

Kevin Everson, Mari Ostendorf
Electrical & Computer Engineering, University of Washington
{kpever, ostendor}@uw.edu

Abstract

Speech representations that capture prosodic information can be useful for both understanding and generation. However, speaker characteristics are reflected in acoustic-prosodic features (e.g., pitch). To address privacy concerns from the leakage of identity information, we propose a new self-supervised approach to learning prosody representations that incorporates speaker disentanglement strategies. We evaluate our encoder on three tasks to probe representation capabilities, including pitch reconstruction and detection of different prosodic events. Our encoder outperforms raw prosody and HuBERT-base baselines, achieving strong speaker disentanglement without adverse impact on prosody-related downstream tasks.

1 Introduction

Prosody broadly refers to the aspects of speech which complement the lexical content, conveying intent-related or paralinguistic information. For example, local variations in pitch and energy, pausing, and duration lengthening can be used to communicate information focus, sarcasm, self-corrections, or the difference between a statement and a question. A wider pitch range and faster speaking rate can express excitement. Representing prosodic information is essential for both speech generation and understanding, whether explicitly or implicitly in vector space models. In this work, we introduce an explicit representation of prosody that aims to disentangle prosody from lexical content and speaker information, motivated by the need for more reliable understanding and generation of expressive speech and the desire for privacy-preserving speech processing systems.

Early computational work on prosody focused on acoustic cues such as fundamental frequency (F0), energy, and duration statistics, using speaker and phoneme normalization to remove speaker and lexical content information. Challenges with this approach are that automatic phonetic time alignment and F0 extraction algorithms are not very reliable, and energy is sensitive to recording conditions. More recently, researchers have explored self-supervised learning approaches, but often the objective of speaker disentanglement is lost.

Acoustic-prosodic cues are known to carry speaker information. When it is not removed, users are made vulnerable to serious privacy breaches, such as identity theft via deepfake generation. Concerns over these vulnerabilities are exacerbated by the proliferation of speech-based AI assistants and the ever-improving model capabilities. Thus, the disentanglement of speaker characteristics is essential for protecting user privacy in cases where identity information is not needed.

To this end, we train a prosody encoder by distilling information from acoustic-prosodic features via self-supervised learning, leveraging input features that disentangle lexical content and a learning framework for disentangling speaker information via target normalization and an adversarial loss function. We evaluate the efficacy of our prosody representation on three tasks: a standard pitch reconstruction task, as well as prosodic prominence and phrase boundary detection. The speaker identification task is used to assess our speaker disentanglement techniques. Our findings indicate that our encoder yields improved prosody modeling over baselines and that speaker information leakage can be effectively diminished without a negative impact on prosody modeling.

2 Related Work

While there have been a number of prosody representation models developed for specific tasks, e.g. speech synthesis, emotion recognion, and speech understanding tasks, this work focuses on the use of self-supervised learning with untranscribed speech to provide a general representation of prosody. In

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 310–315

addition, we are interested in frame-based representations, since they allow the flexibility of using different levels (e.g. phone, word) via pooling.

Standard acoustic modeling approaches used in speech recognition (e.g. HuBERT (Hsu et al., 2021), wav2vec 2.0 (Baevski et al., 2020)) have proved to be useful for some prosody tasks (Lin et al., 2023), but recently there have been a few efforts to more explicitly target prosodic characteristics of speech. In particular, our work builds on ProsodyBERT (Hu et al., 2023) and PE-Wav2vec (Liu et al., 2024). As in ProsodyBERT, we use a HuBERT architecture with hidden units produced by acoustic-prosodic cues, and augment the standard masked prediction loss with a span boundary loss. However, like PE-Wav2vec, we use the estimated glottal waveform as input, rather than raw prosody features. Different from both, we introduce speaker disentanglement strategies, including speaker-normalization of the prosody features used for masked prediction targets and an adversarial speaker loss function. Other examples of speaker disentanglement in prosody representation learning include information bottlenecks (Qian et al., 2020), pitch-shifting the audio input (Weston et al., 2021), and adversarial losses (Qu et al., 2025).

Prosody representations have been evaluated in speech synthesis (or voice conversion), intent recognition (sentiment, sarcasm, persuasiveness), and paralinguistic tasks (emotion recognition, health diagnostics). Prosody benchmarks include SUPERB-prosody (Lin et al., 2023) and DAMMP (Weston et al., 2021). Since we are interested in representation of linguistic (rather than paralinguistic) structure, we use prosodic event detection tasks, but also report on the pitch reconstruction task in SUPERB-prosody. Speaker disentanglement is most often evaluated with objective measures such as speaker identification or verification tasks (Lian et al., 2022; Qian et al., 2019), or subjective assessments of source speaker characteristics present in generated speech (Deng et al., 2024; Qian et al., 2020). In this study, we use speaker identification.

3 Prosody Encoder

Our prosody encoder architecture, shown in Fig. 1, is inspired by ProsodyBERT (Hu et al., 2023), which follows HuBERT's approach to self-supervised learning (Hsu et al., 2021) but adds a span-based objective to encourage learning of suprasegmental characteristics. Key differences of our work include: i) use of the estimated glottal waveform as input, and ii) addition of an adversarial speaker identification loss term.

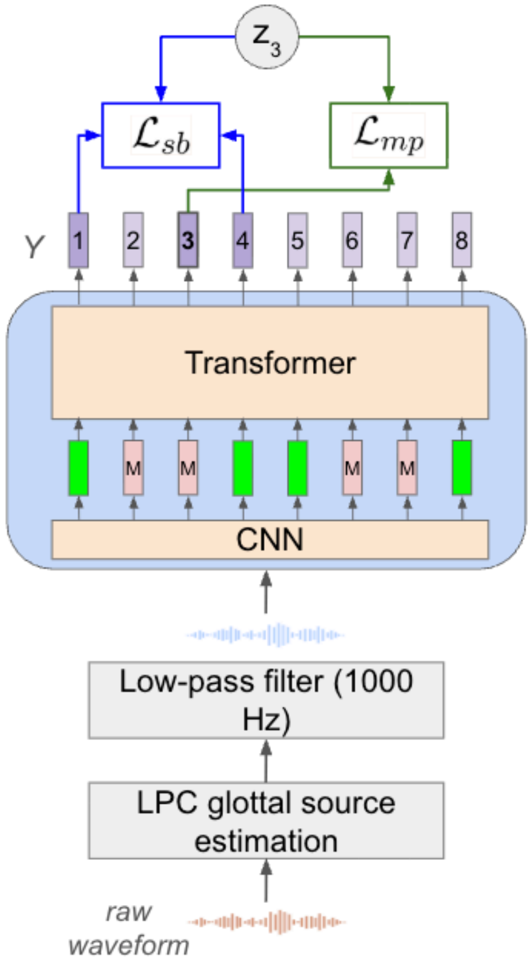

Figure 1: Prosody encoder pretraining, which uses a masked-prediction objective $\mathcal{L}_{mp}$ and a span-boundary objective $\mathcal{L}_{sb}$.

3.1 Model details

Our prosody encoder shares an architecture with HuBERT-base, consisting of a convolutional input module followed by a transformer.

Input. The model input is created using glottal source estimation, which is more robust and computationally efficient than pitch tracking, and captures voice quality information. We extract the glottal source by inverse-filtering with estimated LPC coefficients (Rabiner and Schafer, 2010) (filter order 16, 25ms window, 10ms stride). In some frames with very low energy (non-speech regions), we find that the coefficients produced by LPC were unreliable, leading to input artifacts and training instability. To alleviate this issue, we simply return the raw waveform in frames with energy less than 10^{-4} rather than applying the inverse-filter. A 1 kHz low-pass filter is then applied, yielding the final model input. In anecdotal listening experiments, we find that this low-pass filter reduces lexical information leakage.

Hidden units. Similar to Hu et al. (2023), hidden units $\mathcal{Z} = (\mathrm{z}_1, ..., \mathrm{z}_T)$ are produced via offline k-means clustering of acoustic-prosodic feature vectors $[\mathrm{P}, \log \mathrm{F0}, \Delta \log \mathrm{F0}, \mathrm{c}_1]$, where:

- P (periodicity) indicates the frame-level probability of voiced speech;
- $\log \mathrm{F0}$, normalized by subtracting a speaker's mean log-pitch, where a speaker's mean log-pitch is computed as a weighted average over $\log \mathrm{F0}$ using P as weights;
- $\Delta \log \mathrm{F0}$; and
- c_1, the first mel-frequency cepstral coefficient, representing the overall slope of the spectrum.

Unlike Hu et al. (2023), we normalize F0 features to support speaker disentanglement. c_1 is used in place of energy to reduce sensitivity to factors such as recording conditions. Before clustering, we apply corpus-level z-normalization to each feature.

3.2 Self-supervised learning objective

The training loss function is the weighted sum:

$$\mathcal{L} = \mathcal{L}_{mp} + \alpha_{sb}\mathcal{L}_{sb} + \alpha_{spk}^{adv}\mathcal{L}_{spk}^{adv},$$

with hyperparameters α_{sb} and α_{spk}^{adv}.

The first component, $\mathcal{L}_{mp}$, is the standard frame-level masked-prediction objective. For index i in a masked span starting at index j and ending at index k, a linear layer produces frame-level distributions over the k hidden unit labels. The masked-prediction objective given by the cross-entropy:

$$\mathcal{L}_{mp} = \log p_{mp}(\mathrm{z}_i \mid \mathbf{y}_i),$$

where $\mathcal{Y}$ is the transformer output when randomly sampled spans of the convolutional encoder output are masked. We use the same span-level masking algorithm introduced in Hsu et al. (2021).

To encourage learning of suprasegmental patterns, we add a span boundary objective (Joshi et al., 2020). Specifically, we add a linear layer to predict the frame-level hidden unit at masked index i, given the *nearest unmasked frames* on the left and right (at $j-1$ and $k+1$, respectively), as well as the distance (in frames) from i to each. The span boundary objective is also given by the cross-entropy:

$$\mathcal{L}_{sb} = \log p_{sb}(\mathrm{z}_i \mid \mathbf{y}_{j-1}, \mathbf{y}_{k+1}, i-j+1, k+1-i).$$

Lastly, we include an adversarial speaker identification objective $\mathcal{L}_{spk}^{adv}$. The frame-level features $\mathbf{y}_i$ are passed to a gradient-reversal layer, then projected via a linear prediction layer to produce a distribution over speaker labels:

$$\mathcal{L}_{spk}^{adv} = \log p_{spk}(\mathrm{spk}|\mathbf{y}_i).$$

Due to the gradient reversal layer, the linear prediction parameters are updated in the direction that minimizes $\mathcal{L}_{spk}^{adv}$, while the encoder parameters are updated in the opposite direction.

The prosody encoder is trained on the transcribed portion of the GigaSpeech corpus (Chen et al., 2021), containing 11K hours of English-language spontaneous and read speech from YouTube, podcasts, and audiobooks. Since speaker labels are not provided, we use pseudo-labels created by extracting utterance-level embeddings from a pretrained speaker encoder, then clustering with 1000 clusters per (Malinen and Fränti, 2025). Pitch and periodicity are extracted with torchcrepe. We train using the fairseq toolkit on four NVIDIA (A40 or L40) GPUs for 500K steps, with a variable batch size averaging ∼30 per GPU[1]. The checkpoint with the lowest validation loss is frozen and used for downstream tasks.

4 Experiments and Results

We evaluate the encoder on a series of downstream tasks to assess its prosody modeling capacity and speaker disentanglement. The (frozen) encoder features are input to a task-specific model, which is fine-tuned on a single A40 or L40 GPU using the s3prl toolkit[2]. In line with our efforts to preserve privacy, we use only the final encoder output layer, in contrast to other SUPERB-based (Yang et al., 2021) setups, which use all intermediate layers. All reported results are on test splits, using the best checkpoint based on development splits.

4.1 Evaluation models

The proposed prosody encoder is compared to two baselines: the standard pretrained HuBERT-base model and the raw speaker-normalized prosody feature vectors ($[\mathrm{P}, \log \mathrm{F0}, \Delta \log \mathrm{F0}, \mathrm{c}_1]$). For ablations, we train versions of our proposed prosody encoder without the adversarial speaker loss and without the speaker normalization of $\log \mathrm{F0}$ when producing hidden units.

[1] github.com/kpeverson/speaker_disentangled_prosody
[2] github.com/kpeverson/s3prl_tobi

`such supervision | according to ash | is a sensible | cost effective alternative | to incarceration |`

Figure 2: Prosody event example: "|" indicates a phrase boundary; highlighted text indicates prominent syllables.

4.2 Prosody prediction evaluations

In addition to the standard pitch reconstruction task, we evaluate on two prosodic event detection tasks, chosen because the self-supervised learning approach was designed with the goal of characterizing linguistic (vs. paralinguistic) prosodic information.

Pitch reconstruction. We evaluate on the LibriTTS (Zen et al., 2019) pitch reconstruction task using the setup and splits from SUPERB-prosody (Lin et al., 2023). After extracting frame-level F0 offline using pYAAPT, $\log \text{F0}$ is predicted over voiced frames from the prosody encoder features using a linear layer. Mean-squared error (MSE) is used as the training objective and reported metric.

Since average pitch varies with speaker and we aim to remove speaker information, we also include a modified evaluation in which the predicted and ground-truth $\log \text{F0}$ contours are shifted to 0-mean.

Phrase boundary detection. We introduce two tasks from the BU Radio Corpus (Ostendorf et al., 1995), containing 11 hours of read speech from seven professional radio announcers with word- and phoneme-level forced alignments, plus ToBI annonations (Beckman and Hirschberg, 1994) which include break index labels with word boundary timestamps. The corpus is partitioned such that $\sim$80/10/10% of each speaker's files are present in the training, development, and test splits, respectively.

We detect full phrase boundaries (i.e. words with "4" break index labels, as indicated by "|" symbols in Fig. 2) from prosody features within $\pm$100 ms of each word boundary. Because changes in prosody can indicate phrase boundaries, we separately pool (via self-attention) the features before and after the timestamp, and concatenate the two resulting vectors before applying the classifier. This task is trained using cross-entropy loss, and evaluated using F1 score and accuracy.

Syllable prominence detection. The tone tier of the BU Radio Corpus ToBI annotations includes timestamps of accented syllables (containing "*" labels, indicated by highlighted text in Fig. 2). To detect these, we pool the frame-level features from each syllable using self-attention and apply a linear classification layer to the syllable-level features. We use the same splits as in the phrase boundary detection task, and again use cross-entropy loss as the training objective and evaluate on F1 and accuracy.

4.3 Speaker disentanglement evaluation

To evaluate the amount of speaker information captured by our prosody modeling and the efficacy of our disentanglement methods, we also fine-tune on the VoxCeleb1 (Nagrani et al., 2017) speaker identification task.

4.4 Results and discussion

Results on the prosody modeling evaluations are reported in Table 1. Among the baseline systems, HuBERT-base achieves stronger results despite the lack of prosody-specific training strategies. The variants of our encoder outperform the baselines. The advantage of our system is greatest on the syllable prominence detection task, achieving a relative F1 improvement of 15% over HuBERT-base.

Notably, the use of speaker-normalized $\log \text{F0}$ and the adversarial speaker objective together did not adversely affect downstream task performance; other than the variant with the adversarial speaker objective and without speaker-normalization of $\log \text{F0}$, the variants achieved similarly strong results on the ToBI detection tasks. At the same time, the variant with both speaker-disentanglement strategies performed the best in the 0-mean pitch reconstruction setup, indicating the strongest modeling of local pitch dynamics.

The speaker identification (SID) results on the VoxCeleb1 test set are shown in Table 2. We note that the accuracy of our HuBERT-base implementation is lower than the 0.81 reported in Yang et al. (2021), since we only use the outputs of the final layer. Our prosody encoder variants indicate much less speaker information (lower SID accuracy) than HuBERT-base, and disentanglement strategies further diminish the speaker information. In terms of SID accuracy, the adversarial objective yields a 46% relative reduction, and the two strategies together yield a 66% relative reduction.

5 Conclusion

We propose a prosody encoder that takes the estimated glottal source as input, and is trained to

Model	Speaker-norm. log F0	$\mathcal{L}_{spk}^{adv}$	ToBI detection tasks: Phrase boundary F1 (↑)	ToBI detection tasks: Phrase boundary acc. (↑)	ToBI detection tasks: Syl. prominence F1 (↑)	ToBI detection tasks: Syl. prominence acc. (↑)	Pitch recons.: Standard MSE (↓)	Pitch recons.: 0-mean MSE (↓)
most freq. class	–	–	*0.00*	*0.87*	*0.00*	*0.70*	–	–
HuBERT-base	–	✗	0.79	**0.95**	0.74	0.85	0.056	0.011
Raw prosody	✓	–	0.49	0.88	0.66	0.83	–	–
Ours	✗	✗	**0.82**	**0.95**	**0.86**	**0.92**	0.027	0.012
	✓	✗	**0.82**	**0.95**	**0.86**	**0.92**	0.048	0.012
	✗	✓	0.73	0.93	0.82	0.89	**0.024**	0.012
	✓	✓	**0.82**	**0.95**	**0.85**	**0.91**	**0.025**	**0.008**

Table 1: Prosody modeling test evaluation results.

Model	Speaker-norm. log F0	$\mathcal{L}_{spk}^{adv}$	Accuracy (↓)
HuBERT-base	–	✗	0.64
Ours	✗	✗	0.41
	✓	✗	0.42
	✗	✓	0.22
	✓	✓	**0.14**

Table 2: VoxCeleb1 test speaker identification results.

predict hidden units produced from raw acoustic-prosodic features. Disentanglement techniques are employed to reduce the amount of speaker information present in the output features. Our findings indicate that prosody modeling is not negatively affected by these speaker disentanglement techniques, an encouraging result given the dual importance of capturing the intent-related information conveyed by prosody while protecting user privacy.

6 Limitations

We use pseudo-labels from utterance-level speaker embeddings, which places limitations on speaker-normalization of log F0 and the adversarial speaker objective. Ground-truth labels would likely be more effective. Chen et al. (2021) indicated plans to update the Gigaspeech corpus metadata with speaker information, but this has not yet been released.

The assessment of the representation learning approach is limited in a few respects. We focused primarily on local prosodic events, but it may be useful to also evaluate on paralinguistic tasks. For comparison to prior work, the understanding tasks leverage hand transcriptions. Evaluation with automatically recognized transcripts would be more informative but would require a more sophisticated scoring algorithm. It would also be useful to evaluate on a speech generation task, which would provide an opportunity for subjective human assessments. Our model is not causal, which is needed for streaming generation scenarios, but translation of our approach to a causal framework is straightforward.

Comparisons to existing prosody models, namely ProsodyBERT (Hu et al., 2023) and PE-Wav2Vec (Liu et al., 2024), were limited due to lack of publicly available code. Our attempts to train similar models resulted in uncompetitive performance on downstream tasks.

7 Ethical considerations

The ethical concern of protecting privacy is an important consideration in the development of this model. Our approach was developed assuming relatively small amounts of data are available for a speaker. It is possible that speaker recognition algorithms could be developed that are effective given larger amounts of data.

Acknowledgments

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0424C0070. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Mary E Beckman and Julia Hirschberg. 1994. The ToBI Annotation Conventions. *Ohio State University*.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. *Interspeech 2021*.

Yimin Deng, Jianzong Wang, Xulong Zhang, Ning Cheng, and Jing Xiao. 2024. Learning Expressive Disentangled Speech Representations with Soft Speech Units and Adversarial Style Augmentation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation learning by Masked Prediction of Hidden Units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Yushi Hu, Chunlei Zhang, Jiatong Shi, Jiachen Lian, Mari Ostendorf, and Dong Yu. 2023. ProsodyBERT: Self-Supervised Prosody Representation for Style-Controllable TTS.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. Place: Cambridge, MA Publisher: MIT Press.

Jiachen Lian, Chunlei Zhang, and Dong Yu. 2022. Robust Disentangled Variational Speech Representation Learning for Zero-shot Voice Conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6572–6576. IEEE.

Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G Ward. 2023. On the Utility of Self-supervised models for Prosody-related Tasks. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1104–1111. IEEE.

Zhao-Ci Liu, Liping Chen, Ya-Jun Hu, Zhen-Hua Ling, and Jia Pan. 2024. PE-Wav2vec: A Prosody-Enhanced Speech Model for Self-Supervised Prosody Learning in TTS. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4199–4210.

Mikko I. Malinen and Pasi Fränti. 2025. Fixed-sized clusters k-Means. *arXiv preprint*. ArXiv:2501.16113 [cs].

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Interspeech 2017*, interspeech2017. ISCA.

Mari Ostendorf, Patti J Price, and Stefanie Shattuck-Hufnagel. 1995. The Boston University Radio News Corpus. *Linguistic Data Consortium*, pages 1–19.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. 2020. Unsupervised Speech Decomposition via Triple Information Bottleneck. In *International Conference on Machine Learning*, pages 7836–7846. PMLR.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5210–5219. PMLR.

Leyuan Qu, Cornelius Weber, Wei Wang, Jia Jin, Yingming Gao, Taihao Li, and Stefan Wermter. 2025. Disentanglement of Prosody Representations via Diffusion Models and Scheduled Gradient Reversal. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):15043–15054.

Lawrence Rabiner and Ronald Schafer. 2010. *Theory and Applications of Digital Speech Processing*. Prentice Hall Press.

Jack Weston, Raphael Lenain, Udeepa Meepegama, and Emil Fristed. 2021. Learning De-identified Representations of Prosody from Raw Audio. In *International Conference on Machine Learning*, pages 11134–11145. PMLR.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, and 1 others. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech 2019*, pages 1526–1530.