Text-Free Prosody-Aware Generative Spoken Language Modeling

Eugene Kharitonov*, **Ann Lee***, **Adam Polyak**, **Yossi Adi**,
Jade Copet, **Kushal Lakhotia**, **Tu-Anh Nguyen**, **Morgane Rivière**,
Abdelrahman Mohamed, **Emmanuel Dupoux**, **Wei-Ning Hsu**
Facebook AI Research
{kharitonov,annl,wnhsu}@fb.com

Abstract

Speech pre-training has primarily demonstrated efficacy on classification tasks, while its capability of generating novel speech, similar to how GPT-2 can generate coherent paragraphs, has barely been explored. Generative Spoken Language Modeling (GSLM) (Lakhotia et al., 2021) is the only prior work addressing the generative aspects of speech pre-training, which replaces text with discovered phone-like units for language modeling and shows the ability to generate meaningful novel sentences. Unfortunately, despite eliminating the need of text, the units used in GSLM discard most of the prosodic information. Hence, GSLM fails to leverage prosody for better comprehension, and does not generate expressive speech. In this work, we present a prosody-aware generative spoken language model (pGSLM). It is composed of a multi-stream transformer language model (MS-TLM) of speech, represented as discovered unit and prosodic feature streams, and an adapted HiFi-GAN model converting MS-TLM outputs to waveforms. We devise a series of metrics for prosody modeling and generation, and re-use metrics from GSLM for content modeling. Experimental results show that the pGSLM can utilize prosody to improve both prosody and content modeling, and also generate natural, meaningful, and coherent speech given a spoken prompt.[1]

1 Introduction

Natural language processing (NLP) has made tremendous progress recently. One of the most significant findings is that language models (LMs) are natural unsupervised multitask learners (Radford et al., 2018, 2019; Brown et al., 2020) — by simply training a big neural network on next word prediction with a large amount of unlabeled text, it learns to comprehend, answer questions, summarize, and even translate (Radford et al., 2019). Fine-tuning such pre-trained models further leads to the state-of-the-art performance on numerous benchmark tasks (Brown et al., 2020), beating tailor-made models trained from scratch only on labeled data.

Given the impressive performance of pre-trained text language models, it is tempting to approach spoken language processing tasks by first transcribing speech into text with an automatic speech recognition (ASR) system and then utilizing text-based models for comprehension and generation. However, there are a number of caveats for such a framework. First, the majority of the world's languages are primarily spoken and do not have associated texts in large quantities (Lewis et al., 2016). In practice, this limits the reach of NLP techniques to a fraction of the world's languages that have a large presence on the web and for which there exists a widely available high quality ASR system. Second, despite sharing the same vocabulary and syntactic rules, the spoken form and the written form of the same language still vary significantly in terms of sentence lengths, word distributions, presence of disfluencies and back-channelings, and so on (Biber, 1991). This makes language models pre-trained on web text not suitable for processing spoken languages. Third, text does not reflect the rich set of features conveyed by oral languages. Speech carries not only phonetic information, but also non-verbal vocalizations (laughter, voice clicks, filler vocalization, etc), rhythm and intonation (prosody), and emotional markers. All of these features could help, not only with generating more expressive speech (Ren et al., 2020; Łańcucki, 2021), but also with the semantic analysis of the content of the message (Cutler et al., 1997; Tran et al., 2017).

To combat these deficiencies, more recently there is increasing interest in exploring speech pre-training using large quantities of unlabeled speech data (Chung et al., 2019; Schneider et al., 2019;

[1] Audio samples can be found at https://speechbot.github.io/pgslm/. Codes and models are available at https://github.com/pytorch/fairseq/tree/main/examples/textless_nlp/pgslm.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8666 - 8681

Kharitonov et al., 2021; Baevski et al., 2020; Hsu et al., 2021c; Liu et al., 2020; Ling and Liu, 2020; Tjandra et al., 2020; Hsu et al., 2021b,a). However, most of the studies evaluate their models on discriminative tasks, such as ASR and those in the SUPERB benchmark (Yang et al., 2021). To the best of our knowledge, generative spoken language modelling (GSLM) (Lakhotia et al., 2021) is the only prior work that evaluates prompted speech completion, a generative tasks that is similar to the text completion task in GPT-2 (Radford et al., 2019). To remove the reliance on text, GSLM exploits discovered units from self-supervised models to build a unit language model (uLM) and a unit-to-spectrogram (u2S) model. Speech completion can be achieved by first sampling a unit sequence from the uLM with a unit prompt inferred from a speech prompt, and then synthesizing the sampled sequence into speech with the u2S model. Unfortunately, because those discovered units encode mostly phonetic information (Polyak et al., 2021), it suffers from the same prosodic information loss issue as text-based LMs. Therefore, when using that uLM for speech completion, it fails to continue with a coherent tone to the prompt.

In this paper, we introduce a prosody-aware generative spoken language model (pGSLM) that jointly models phonetic content and prosody, in order to leverage prosody for comprehension, and to generate speech coherent with the prompt, which is a precursor for building speech-based dialogue systems. In keeping with our aim of liberating NLP from its over-reliance on text, we follow GSLM and represent the phonetic content with self-supervised units discovered from raw audio. As for prosody, it is represented by the pattern of quantized fundamental frequency (F0) and duration. pGSLM is comprised of two separately trained components: an auto-regressive Multi-Stream Transformer Language Model (MS-TLM) that predicts the next phonetic and prosodic representation given the past ones, and a unit High-Fidelity Generative Adversarial Network (HiFi-GAN) adapted from Polyak et al. (2021) that converts the MS-TLM output into a waveform like a vocoder. To evaluate the proposed model, we adopt metrics from (Lakhotia et al., 2021) for content evaluation, and devise a series of metrics for prosody evaluation. Experimental results demonstrate that 1) joint modeling of prosody improves phonetic content modeling, 2) pGSLM can generate speech continuation coherent with the prompt in term of the content and the prosody, and 3) proper choices of model and prosodic representation is crucial to synthesizing natural, coherent, and expressive speech.

2 Related Work

Our work is related to utilizing prosody for comprehension and predicting prosody for speech synthesis, which we discuss in the following sections.

2.1 Improving Comprehension with Prosody

Prosody, which is often characterized by the rhythm, intonation, and intensity of speech, carries useful information for comprehending speech in addition to the textual content (Cutler et al., 1997). Prior studies have shown that including prosody information can improve the performance from text-only models on speech segmentation (Shriberg et al., 2000), dialogue act classification (Shriberg et al., 1998; Ward and Tsukahara, 2000), syntactic parsing (Tran et al., 2017), speech–language pathology (Cohen et al., 2019), ASR (Ostendorf et al., 2003; Shriberg and Stolcke, 2004), and language modeling (Huang and Renals, 2007; Su and Jelinek, 2008; Ward et al., 2012). These studies provide strong empirical evidences for the benefit of considering prosody in processing spoken languages, especially in the conversational scenarios.

This work shares the same motivation, but differs from the prior work in two crucial aspects. First, this work utilizes discrete units discovered from a self-supervised model and hence does not require any textual supervision, making it applicable to both written and unwritten languages, while in the prior work prosody information is used alongside text. Second, our model can be regarded as the speech version of GPT, which does not require any task-specific labels and can be pre-trained on large quantities of unlabeled speech data. The ability to leverage more data is shown to be the key to achieve good performance in text pre-training.

2.2 Prosody Prediction for Speech Synthesis

The proposed pGSLM model can be re-purposed as a text-to-speech (TTS) model when the phonetic content (represented as a unit sequence) is given and the prosody is generated by the MS-TLM model. This is similar to FastSpeech (Ren et al., 2020) and FastPitch (Łańcucki, 2021) TTS models, where prosodic features are predicted from text and speech are generated conditioning on both

the text and the predicted prosodic features. As FastSpeech and FastPitch are designed to improve the inference-time efficiency from auto-regressive models like Tacotron (Wang et al., 2017), they predict prosodic features and spectrograms without introducing dependency between time steps. In other words, these models assume that the prosody features within an utterance are not correlated across time steps given the text, whereas our proposed MS-TLM does not make such an assumption. We will demonstrate empirically the conditional independence is not a realistic assumption and our model achieves better performance on prosody metrics with auto-regressive modeling.

As for analysis on prosody modeling, we present more extensive metrics by considering both teacher-forcing decoding and sampling, while prior work does not consider the multi-modal nature of prosody and only generate prosody deterministically (Ren et al., 2020). Moreover, we also evaluate prosody in a more disentangled manner by measuring the error of the prosody prediction module alone instead of measuring the error of the prosody extracted from the synthesized waveform: the latter conflates the impact from both the prosody prediction module and the vocoder.

3 Method

In this section, we first describe the phonetic and prosodic representations used in pGSLM, and then introduce the two components it is comprised of: a multi-stream transformer language model and an adapted unit HiFi-GAN.

3.1 Phonetic and Prosodic Representations

We choose units with a vocabulary size of 100 derived from HuBERT (Hsu et al., 2021a), a self-supervised speech model, as the phonetic representation. Specifically, these units are obtained through clustering the 6th transformer layer output of the base HuBERT model provided in (Hsu et al., 2021a) using a k-means algorithm, following the recipe of HuBERT closely. A speech waveform can therefore be encoded into a sequence of discrete units at a frame rate of 50 units per second, or alternatively, into a sequence of (unit, duration) tuples using run-length encoding. HuBERT units were found to perform favorably compared to other self-supervised units such as wav2vec 2.0 (Baevski et al., 2020) and VQ-VAE (van den Oord et al., 2017) in terms of lexical content modeling (Lakhotia et al., 2021) and disentangling prosodic information (Polyak et al., 2021).

We use *unit duration* d and *fundamental frequency* (F0, or pitch) f to derive prosodic representations. Polyak et al. (2021) has shown that pairing HuBERT units with duration and F0 enables high-quality speech re-synthesis that preserves more prosodic information such as intonation compared to re-synthesizing with only units. Similar results are demonstrated in several other studies (Ren et al., 2020; Łańcucki, 2021) in the context of text-to-speech synthesis. Unfortunately, while F0 encodes prosodic information, it also encodes significant amount of speaker information. Figure A.1 in the appendix illustrates how speaker and prosodic information (emotion) are disentangled in raw pitch using a multi-speaker multi-emotion dataset, EmoV (Adigwe et al., 2018). We do not wish to model speaker variation in pGSLM because it is less relevant to spoken language understanding compared to prosody. To that end, we propose to model *speaker-mean normalized log F0*: $lf = \log f - \mathbb{E}_{f' \text{ from the same speaker as } f}[\log f']$, which can be interpreted as the ratio to the mean pitch in the log space: $lf = \log(f/\bar{f})$, where $\bar{f} = \exp \mathbb{E}_{f'}[\log f']$. Specifically, the equation above is used for voiced frames, and the expectation is taken over voiced frames from a speaker. For unvoiced frames, we simply set $lf = 0$.

One may ask why F0 is only normalized by the speaker mean but not the variance. We argue that the variance encodes the "level of expressiveness" and it is desired to preserve it. This is demonstrated empirically in Figure A.2 in the appendix, where speakers from expressive datasets, EmoV and Blizzard 2013 (SynSIG), exhibits larger speaker log F0 standard deviation than those in less expressive datasets, LJSpeech (Ito and Johnson, 2017) and VCTK (Veaux et al., 2016). On the other hand, we also found that variance is more correlated mean in the linear space than in the log space, as shown in Figure A.3. Therefore, we argue that mean-normalized log F0 is a more suitable representation for prosody as it encodes less speaker information while preserving the level of expressiveness.

3.2 Multi-Stream Transformer LM

We adapt the Transformer LM from (Lakhotia et al., 2021) to take multiple streams of input and predict multiple streams of output, and refer to it as the Multi-Stream Transformer Language Model

(MS-TLM). An MS-TLM predicts a sequence of segment representations, which reduces the sequence length significantly and is found beneficial compared to predicting frame sequences (Lakhotia et al., 2021). Each segment is represented with the unit u, duration (in frames) d, and normalized pitch lf. The first two are obtained by run-length encoding the fixed frame rate unit sequence, while a segment lf is computed by averaging those from voiced frames within a segment or set to 0 if the entire segment is unvoiced. An example is provide in Appendix C.

3.2.1 Delayed prosody prediction

Let subscript t be the segment index. At each step, a vanilla MS-TLM takes $(u_{t-1}, d_{t-1}, lf_{t-1})$ as input, linearly projects each of them to the dimension of the transformer, and feeds the summed embeddings to the transformer. The transformer output at that step is projected to the dimension of each stream to predict u_t, d_t, and lf_t independently. The distribution modeled by the synchronous MS-TLM $p(u_{1:T}, d_{1:T}, lf_{1:T})$ can be written as:

$$\begin{aligned}\prod_{t=1}^{T} & p(u_t \mid u_{1:t-1}, d_{1:t-1}, lf_{1:t-1}) \\ & \times p(d_t \mid u_{1:t-1}, d_{1:t-1}, lf_{1:t-1}) \\ & \times p(lf_t \mid u_{1:t-1}, d_{1:t-1}, lf_{1:t-1}). \end{aligned} \quad (1)$$

We see that the factorial assumption here may be too strong, because the duration and the pitch of a segment are highly correlated with the phonetic content of the same segment. To alleviate that without introducing intra-step dependency or interleaving streams (which increases the sequence length and requires determining an order for the three streams a priori), we introduce a delay factor Δ ($\Delta \geq 0$) for prosodic streams, which shift prosodic input and output streams backward by Δ steps, taking $(u_{t-1}, d_{t-\Delta-1}, lf_{t-\Delta-1})$ as input and outputting $(u_t, d_{t-\Delta}, lf_{t-\Delta})$. When $\Delta = 1$, each step of the LM predicts the unit of the current segment and the prosodic representations of the previous segment, of which the lexical unit has been observed already, as shown in Figure 1.

3.2.2 Quantizing prosodic representations

A straightforward solution to encode prosody streams d and lf is to represent them as continuous values and minimize an L1 or L2 loss for training, similar to FastSpeech2 (Ren et al., 2020) and FastPitch (Łańcucki, 2021). Doing so assumes that the duration and the pitch of a segment follow

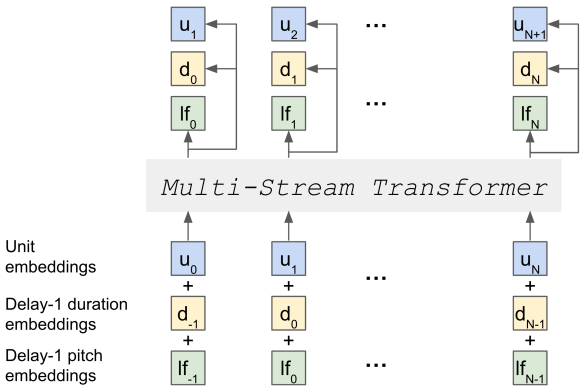


Figure 1: Delayed multi-stream transformer language model with prosody stream delay $\Delta = 1$.

a unimodal distribution (Laplace for L1 and Gaussian for L2) given the context. If the underlying distribution is multimodal with wide spread, the learned distribution would be significantly underfitting with a mean far from the modes. Empirically, we found that such modeling indeed leads to predicting lf values very close to 0 for all segments, and the generated prosody sounds dull and boring.

Inspired by WaveNet (Oord et al., 2016), we represent prosodic features as discrete random variables through quantization. It is straightforward to quantize d since it encodes integer values originally (length in frames). We set the maximum length to be 32 and the bin width to be 1, resulting in 32 bins. We quantize speaker-mean normalized log F0 lf into $K = 32$ bins such that each bin with boundaries $[b_{i-1}, b_i]$ contains the same probability mass: $\mathbb{P}(lf \in [b_{i-1}, b_i]) = 1/K$.

3.2.3 Training objective

The training loss is a weighted sum of three per-stream losses. Omitting dependency on the context for brevity, MS-TLM defines a distribution $p(u_t, d_t, lf_t)$ of the potential values for a timestep t. Then, denoting ground-truth per-channel values as u_t^*, d_t^*, lf_t^*, we get:

$$\begin{aligned} L(p(u_t, d_t, lf_t), u_t^*, d_t^*, lf_t^*) &= L_u(p(u_t), u_t^*) \\ &+ \alpha \cdot L_d(p(d_t), d_t^*) + \beta \cdot L_{lf}(p(lf_t), lf_t^*) \end{aligned} \quad (2)$$

In all experiments, we use cross-entropy as the loss on the predictions of the unit channel (L_u). Whenever we operate on quantized prosody values (both duration and F0), we also use cross-entropy as losses L_d and L_{lf}. In the case of continuous-valued prosody streams, we treat predicted values $p(d_t)$ and $p(lf_t)$ as the mode of Laplacian distributions and maximize the log likelihood of the model,

which is equivalent to minimizing an L1 loss. In preliminary experiments, we found that the results are relatively robust to variations of the relative weights α and β, hence we fix them $\alpha = \beta = 0.5$ in all our experiments.

3.2.4 Sampling from a model

To generate new utterances, potentially conditioned on a prompt, we run autoregressive generation where at each step we sample units, duration, and normalized log F0 values, append them to the context and feed them back. In the case of discrete channels (units, also duration/pitch in the case of discrete-valued models), we sample from the corresponding multinomial distribution. As commonly done in language modelling (Lakhotia et al., 2021), we perform sampling with temperature by scaling the logits by the temperature parameter. We fine-tune the temperature on the validation data.

For MS-TLM that models normalized log F0 as continuous variables, we draw samples from a Laplacian distribution with its location parameter set to the predicted value, because the model assumes the output distribution is Laplacian (see §-3.2.3). For duration, to avoid sampling invalid values, we sample from a Laplacian distribution truncated at zero and round it to the nearest positive integer.

3.3 Waveform Generation with Unit Hifi-GAN

Given $(u_{1:T}, d_{1:T}, lf_{1:T})$ generated from the MS-TLM, we adapt the discrete unit-based HiFi-GAN vocoder from (Polyak et al., 2021) to generate waveform. The original vocoder proposed in (Polyak et al., 2021) takes in frame-level discrete unit, pitch and speaker embedding as input and applies VQ-VAE quantization on the pitch. As MS-TLM predicts quantized speaker-mean normalized log F0 on the segment level, we modify the training of the vocoder so that it takes frame-level segment-average pitch as input, where the pitch values for frames within a segment are set to the same value. We apply the same quantization described in § 3.2.2 instead of VQ-VAE on the pitch. The unit Hifi-GAN and the MS-TLM are trained separately.

4 Experimental Setup

4.1 Data, Model, and Training

In our experiments, we train MS-TLM models on two English datasets: LibriSpeech (Panayotov et al., 2015) and a 6K-hour subset (Rivière and Dupoux, 2020) of Libri-Light (Kahn et al., 2020) which we refer to as LL-6K. Both datasets represent audio books and we use LibriSpeech dev-clean and test-clean as validation and test sets. As described in Section 3.1, we use HuBERT-based unit representations. However, to investigate whether our proposed models can work with other types of units, we also experiment with CPC (Rivière and Dupoux, 2020; Oord et al., 2018) and ground-truth phone representations. We experiment with a vocabulary of 100 units when working with Hubert and CPC, following the same protocol and using the same pre-trained models as Lakhotia et al. (2021). On the other hand, frame-level phone transcripts are obtained through forced-alignment using the `tri6b` model from Kaldi's LibriSpeech recipe (Povey et al., 2011). The position- and context-independent phones without lexical stress markers are used, which include 41 units (39 phones, one silence `SIL`, and one spoken noise `SPN`). The frame rate of CPC and phone units is 100Hz, and is 50Hz for HuBERT units.

We experiment with MS-TLM of two sizes: base and large. The base one has 6 layers, 8 attention heads per layer, embedding size of 512. Its FFN layer has 2048 units. The large variant has 12 layers, each with 16 heads, embedding size of 1024 and the FFN layer is of dimensionality 4096. We set attention dropout and dropout probabilities to 0.1 for both alternatives. On top of that, we apply sequence-level and span-level (Baevski et al., 2020) input dropout to the two prosody streams. Specifically, each stream is zero-ed out with a probability of 0.2, and 2% of the steps are selected as starts, from which 5 steps of that stream is zero-ed out. Optimization is done using Adam (Kingma and Ba, 2014) with a peak learning rate of 5e-4. Learning rate ramps up linearly for the first 4K updates, and then decays to 0 with an inverse square-root schedule. We train the base model for 70 epochs, and large model for 100 epochs. Each GPU's batch contains up to 3072 (u, d, lf) segments and we used 8 (16) GPUs to train base (large) MS-TLM. For each update, we aggregated gradients from 8 batches.

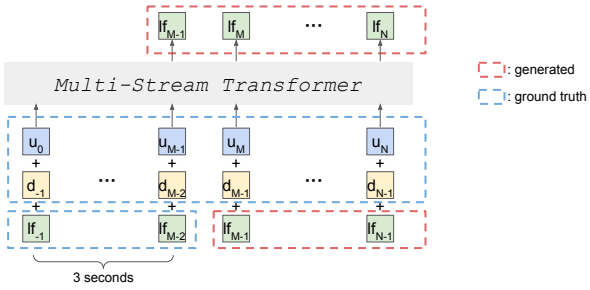


Figure 2: Per-stream prosody continuation task. lf is the target stream for continuation in this example.

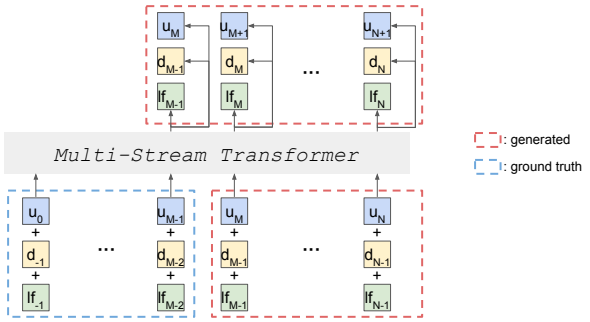


Figure 3: Speech continuation task.

4.2 Prosody and Content Evaluation

Our overall goal is to find models that can freely generate meaningful content and consistent as well as diverse prosody. In this Section, we define a set of metrics that measure models' performance over each stream individually and combined, in both the teacher-forcing mode and the inference mode.

4.2.1 Teacher-forcing metrics

A simple way to evaluate models is to measure its loss on hold-out data in a setup where for each step the full ground truth context is provided. For the unit stream, we measures Negative Log-Likelihood (NLL), equivalent to cross-entropy. For the duration and pitch streams we use Mean Absolute Error (MAE), equivalent to L1 loss. When the pitch values are quantized, we de-quantize predictions to the means of the respective buckets.

4.2.2 Per-stream prosody continuation

We next evaluate the model's ability to complete a stream in isolation. Specifically, we provide a 3s prompt for all streams, and then sample auto-regressively the target stream while feeding the ground truth value for the other streams, as depicted in Figure 2. The prompts are inferred from the utterances in the validation set. When prosodic features are quantized, we sample with a temperature $\tau \in \{0.0, 0.25, 0.5, 0.7, 1.0, 1.3\}$, and when they are continuous, we sample with a scale $b \in \{0.0, 0.05, 0.125, 0.25, 0.5, 0.7, 1.0, 1.3\}$ for duration and $b \in 0.01 \times \{2^{-6}, 2^{-5}, \cdots, 2^{0}\}$ for pitch. The temperature/scale is chosen to minimize the Min-MAE for the corresponding stream, which we describe next. We chose different sweeping ranges for continuous pitch and duration because they have different inherent standard deviations.

Correctness (Min-MAE) A prompt might have multiple meaningful continuations in the content space (Lakhotia et al., 2021). Similarly, a single sentence can have multiple correct prosodic profiles. To account for that, for each prompt we generate $n = 20$ samples so that a model has a chance to cover most modes of the underlying distribution, and report the minimal MAE (min-MAE) against the reference among the n samples.

Consistency (Corr.) To quantify the models' capability to generate consistent prosody, we measure Pearson correlation between the mean values of a stream in the prompt and in the generated continuation. Clearly, if the prompt has a distinct tempo or a pitch, a good continuation should reflect this. The same setup as the min-MAE metric is used ($n = 20$) with one exception: we only consider sequences that are at least 6s long.

Expressiveness (Std.) To measure how expressive the generated prosody is, we calculate the standard deviation of the generated values and expect a good model to exhibit a similar level of that as the ground truth. The same setup as in "Min-MAE" is used.

4.2.3 Speech continuation

Lastly, we evaluate the model's ability to carry out prompted speech completion, where all three streams are sampled given a 3s prompt using the temperature/scale parameter determined from per-stream continuation (§ 4.2.2) as illustrated in Figure 3. We sample the MS-TLM auto-regressively until it emits the EOS unit or reaches the length of the reference. The MS-TLM output is synthesized into a waveform using the adapted HiFi-GAN.

Content (Max-Word-Cont-BLEU2) We re-use the maximum word-level continuation BLEU2 proposed by Lakhotia et al. (2021) to quantify how well a model can complete a prompt in terms of the textual content. We transcribe the waveform with an off-the-shelf wav2vec 2.0-based ASR (Baevski

et al., 2020) (same as (Lakhotia et al., 2021)) and compute the BLEU2 score for each of the $n = 20$ continuations against the reference completion. The highest one is used as score for a prompt.

Human evaluation (MOS, MMOS, PMOS) We ask humans to evaluate three aspects of speech continuation: sound quality, meaningfulness (how natural the text content is considering both grammar and meaning), and prosody (how consistent and natural the intonation and the rhythm is). We follow the human evaluation protocol used by Lakhotia et al. (2021) closely, where raters evaluate subjective quality of the recordings using headphones on a scale between 1 to 5 with an increment of 1, the higher the better. Only Native English speakers were recruited as raters for all three studies. The same 100 prompts as (Lakhotia et al., 2021) from LibriSpeech test-other are used, and each system generates one continuation per prompt. Each continuation is evaluated by at least 5 raters for each aspect. The CrowdMOS package (Ribeiro et al., 2011) was used for all experiments using the recommended recipes for outlier removal. All participants were recruited using the Amazon Mechanical Turk platform. The metrics on the three aspects are denoted as MOS, M-MOS, and P-MOS.

5 Results

5.1 Prosodic Inputs Are Useful for Content and Prosody Modeling

In Table 1 we report teacher-forcing metric calculated on LibriSpeech dev-clean dataset for a diverse set of models. In rows 1-8, we report metric values for base MS-TLM models that are trained on LibriSpeech 960h transcribed into HuBERT-100 units. In rows 9-12 we consider large MS-TLM models trained on HuBERT transcripts of LL6k. Rows 13 & 14 and 15 & 16 contain metric values for models that are trained on LibriSpeech 960h transcribed using CPC and ground-truth phonetic units.[2] The row 1 corresponds to the prosody-ignorant baseline model of (Lakhotia et al., 2021).

On comparing two models that only predict units (rows 1 and 5) we see that by simply adding prosodic channels to the input of the model, we obtain considerably lower level of negative log-likelihood of the units (u NLL: 1.522 vs. 1.336). The same trend persist for the models that predict prosodic channels, too. For instance, this holds in the case of the continuous-F0 models (rows 9 & 11: 1.513 vs. 1.421) and, equally for the quantized F0 HuBERT-based models (rows 10 and 12: 1.522 vs. 1.406). Moreover, this holds for the CPC-based models (row 13 & row 14) and even for the models trained on phone transcripts (rows 15 & 16). Hence we conclude that prosodic input *universally* improves speech "content" modelling.

Our results in Table 1 also allow us to investigate whether shifting prosody streams w.r.t. the unit stream ($\Delta > 0$) is useful. On comparing rows 6 & 7 we see that this is indeed the case: at an expense of some increase in u NLL (e.g., 1.337 vs. 1.441) we obtain considerable relative improvement in d

ID	Input	Output	Quant?	Δ	u NLL↓	d MAE↓	lf MAE↓
Base MS-TLM, HuBERT units, trained on LS960							
1	u	u	n/a	n/a	1.522	n/a	n/a
2	u	(u,d,lf)	✓	0	1.525	0.759	0.115
3	u	(u,d,lf)	✓	1	1.517	0.586	0.112
4	u	(u,d,lf)		1	1.514	0.562	0.093
5	(u,d,lf)	u	✓	0	**1.336**	n/a	n/a
6	(u,d,lf)	(u,d,lf)	✓	0	1.337	0.722	0.052
7	(u,d,lf)	(u,d,lf)	✓	1	1.441	0.551	0.049
8	(u,d,lf)	(u,d,lf)		1	1.447	**0.536**	**0.046**
Large MS-TLM, HuBERT units, trained on LL6k							
9	u	(u,d,lf)		1	1.513	0.563	0.095
10	u	(u,d,lf)	✓	1	1.522	0.586	0.116
11	(u,d,lf)	(u,d,lf)		1	1.421	**0.527**	**0.043**
12	(u,d,lf)	(u,d,lf)	✓	1	**1.406**	0.543	0.047
Base MS-TLM, CPC units, trained on LS960							
13	u	(u,d,lf)	✓	1	1.511	1.302	0.122
14	(u,d,lf)	(u,d,lf)	✓	1	**1.353**	**1.181**	**0.045**
Base MS-TLM, Phone units, trained on LS960							
15	u	(u,d,lf)	✓	1	1.559	2.748	0.150
16	(u,d,lf)	(u,d,lf)	✓	1	**1.485**	**2.419**	**0.079**

ID	Input	Output	Quant?	τ	u NLL stddev	d MAE stddev	lf MAE stddev
Base MS-TLM, HuBERT units, trained on LS960							
2	u	(u,d,lf)	✓	0	0.0004	0.00226	0.0033
6	(u,d,lf)	(u,d,lf)	✓	0	0.0022	0.00306	0.00037

Table 1: (Top) Teacher-forcing metrics on Librispeech dev-clean. Exp 1 is identical to the uLM presented in (Lakhotia et al., 2021). We can observe that models with both phonetic and prosodic input (u, d, lf) consistently outperforms their counterpart model with only phonetic input u. This trend holds for different lexical representations (HuBERT, CPC, phone), both continuous and discrete prosodic features, and different delay factors τ. (Bottom) We train Exp ID 2 and 6 with five random seeds and measure the standard deviation on all three metrics. Results show that the gap between the models with and without prosodic input is significant relative to the standard deviation.

[2]Note: the metric values in this section are only comparable within the same unit type. To compare across unit types, one can synthesize the MS-TLM output into waveform and transcribe the speech with an ASR systems to compute metrics in the word or character space.

ID	Input	Output	Quant?	d			lf			Max-Word-Cont-BLEU2 ↑
				min-MAE ↓	Corr.↑	Std.↑	min-MAE ↓	Corr.↑	Std.↑	
	ground truth		n/a	.000	.463	1.32	.000	.520	.163	1.000
	resynthesized		✓	.000	.464	1.32	.000	.315	.145	.943
Large MS-TLM, HuBERT units, trained on LL6k, $\Delta = 1$										
9	u	(u, d, lf)		.542	.176	.942	.084	.093	.081	.488
10	u	(u, d, lf)	✓	.542	.086	**.965**	.096	.217	.147	.489
11	(u, d, lf)	(u, d, lf)		.539	**.344**	.940	.081	**.494**	.076	.498
12	(u, d, lf)	(u, d, lf)	✓	**.536**	.242	.946	**.077**	.324	**.149**	**.499**

Table 2: Continuation metrics on Librispeech test-clean. The temperature and the scale parameters used for sampling are selected using the dev-clean set. We compare the four large MS-TLM models from Table 1 here. For each utterance in the test set, we use the first 3 seconds as the prompt and sample 20 continuations.

ID	Input	Output	Quant?	Mean Opinion Score		
				MOS	M-MOS	P-MOS
resynthesized				3.21±0.09	3.95±0.32	3.87±0.45
Large MS-TLM, HuBERT units, trained on LL6k, $\Delta = 1$						
9	u	(u, d, lf)		3.16±0.19	3.80±0.25	3.69±0.42
10	u	(u, d, lf)	✓	2.66±0.18	3.36±0.40	3.15±0.52
11	(u, d, lf)	(u, d, lf)		3.31±0.23	3.76±0.27	**3.78±0.46**
12	(u, d, lf)	(u, d, lf)	✓	**3.43±0.20**	**4.04±0.20**	3.75±0.48

Table 3: Human evaluation on sound quality (MOS), meaningfulness (M-MOS), and prosody (P-MOS). ± indicates 95% CI.

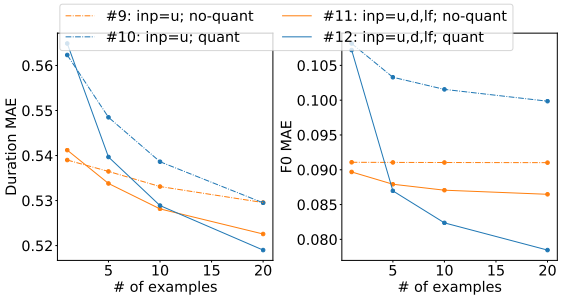


Figure 4: Minimal-MAE of duration (left) and pitch (right) with respect to different number of samples for the four large models (ID 9-12).

MAE (0.722 → 0.551). The trend follows when further increasing Δ. We also observe that having prosody in the context is beneficial when modeling prosody itself. Indeed, this is the case across all pairs of models (rows 9 & 11, 10 & 12) according to d MAE and lf MAE metrics. Moreover, this holds for the types of units that differ from HuBERT (CPC: rows 13 & 14, phonetic units: rows 15 & 16).

5.2 Prosodic Inputs Are Useful for Speech Generation

In our next experiment we study how the number of sampled prompt continuation affects prosody accuracy metrics (MAE). We report results for the four large models (rows 9-12) in Figure 4. From these results we observe that models that operate on quantized prosodic streams greatly benefit from sampling multiple candidates. In contrast, the two continuous-valued models seem to benefit little if at all (in the case of the F0 stream). We hypothesise that this striking difference is due to the ability of the multinomial-sampled trajectories to cover multiple mode of the underlying distribution, while the continuous-valued models produce samples that are "averaged" to the median of the underlying distribution due to the L1 loss.

In Table 2 we report the continuation metrics for four large MS-TLM models, trained on HuBERT transcripts of LL-6k (they correspond to rows 9-12 in Table 1).[3] These models differ in whether they have prosodic input or not (rows 11 & 12 vs. 9 & 10) and if the prosodic channels are discretized or not (10 & 12 vs. 9 & 11).

Firstly, on comparing models with and without prosodic input, we observe that having prosody in input improves the accuracy of the prosody continuation (in terms of MAE). This holds for predicting duration (e.g., 0.542 and 0.536 for rows 10 and 12). We see a higher relative difference for lf (e.g., 0.096 vs. 0.077, same models). Our proposed models are also able to leverage provided prosody input to maintain high consistency of the prosody continuation, as measured by the correlation metrics. For example, for the continuous-prosody models the correlation values grows from 0.176 to 0.344 for the duration prediction and from 0.093 to 0.494 for the F0 channel. Having prosody input also turns out to be important for the word-level BLEU metric: models 11 and 12 outperform their counterparts without prosody inputs, 9 and 10.

Next, when contrasting discrete- and continuous-prosody models the following picture emerges. For

[3] Audios samples of speech continuation are included in the supplementary material.

both duration and F0 channels, discrete models achieve lower min-MAE errors. Further, both discrete models generate considerably more diverse F0 values than either of the continuous models (up to 2x higher std). Among the models with prosody inputs, the one with discrete prosody get higher variability in the d channel. In contrast, the correlation metrics favor the prosody-aware continuous model. From the point of view of the word-level BLEU scores, both models are very close with the quantized model (row 12) being slightly ahead. We attribute this difference between the models to the ability of discrete-valued MS-TLM to better describe multi-modal distributions, as we saw above in the experiment reported in Figure 4.

Table 3 presents the human evaluation results. The model with prosody input and quantized prosody performs significantly better than the rest on MOS and M-MOS, and is on par with the variant with prosody input and continuous prosody on P-MOS. Note that when not having the prosody input, the model with quantized prosody performs significantly worse on all metrics, demonstrating the importance of auto-regressive generation for discrete representation.

To summarize, we conclude that (i) including prosody input allows better modelling of speech, and (ii) architectures that operate with quantized prosody values, generally, perform better on our introduced metrics.

6 Conclusion and Future Work

In this work, we propose a text-free prosody-aware generative spoken language model, pGSLM, which models textual content and prosodic information explicitly and does not use any text supervision by leveraging self-supervised units. Through extensive evaluation on a diverse set of metrics, we demonstrated that prosody not only improves content modeling, but also enables better prompted speech generation that is aware of both the content and the prosody from the prompt for the first time in the literature. We conducted a number of ablation studies to validate the effectiveness of model design choices.

As for broader impacts, this work serves as the foundation for building better conditional speech generation applications where prosody is essential, such as in the conversational scenarios. In addition, the proposed model could also serve as a pre-trained model for other classification tasks, such as emotion recognition or syntactic parsing from speech, or as a pre-trained model for generative tasks such as text-to-speech synthesis with more expressive and coherent prosody. Finally, the proposed prosody metrics (teacher-forcing duration and pitch MAE, continuation correctness/consistency/expressiveness) may also be used for evaluation of text-to-speech synthesis systems that can produce diverse prosody for a given text input.

References

Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Wei Chu and Abeer Alwan. 2009. Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3969–3972. IEEE.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An unsupervised autoregressive model for speech representation learning. In *Proc. INTERSPEECH*, pages 146–150.

Jacob T Cohen, Alma Cohen, Limor Benyamini, Yossi Adi, and Joseph Keshet. 2019. Predicting glottal closure insufficiency using fundamental frequency contour analysis. *Head & neck*, 41(7):2324–2331.

Anne Cutler, Delphine Dahan, and Wilma Van Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and speech*, 40(2):141–201.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021a. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, et al. 2021b. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. In *Interspeech*.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021c. HuBERT: How much can a bad teacher benefit ASR pre-training? In *Neural Information Processing Systems Workshop on Self-Supervised Learning for Speech and Audio Processing Workshop*, pages 6533–6537.

Songfang Huang and Steve Renals. 2007. Using prosodic features in language models for meetings. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 192–203. Springer.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Jacob Kahn et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP*.

Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux. 2021. Data augmenting contrastive learning of speech representations in the time domain. *arXiv preprint arXiv:2007.00991*, pages 215–222.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, et al. 2021. Generative spoken language modeling from raw audio. *arXiv preprint arXiv:2102.01192*.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.

M. Paul Lewis, Gary F. Simon, and Charles D. Fennig. 2016. *Ethnologue: Languages of the World, Nineteenth edition*. SIL International. Online version: http://www.ethnologue.com.

Shaoshi Ling and Yuzong Liu. 2020. DeCoAR 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659*.

A. T. Liu, S. Yang, P. Chi, P. Hsu, and H. Lee. 2020. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423.

Tomohiro Nakatani, Shigeaki Amano, Toshio Irino, Kentaro Ishizuka, and Tadahisa Kondo. 2008. A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments. *Speech Communication*, 50(3):203–214.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Mari Ostendorf, Izhak Shafran, and Rebecca Bates. 2003. Prosody models for conversational speech recognition.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. In *Proc. INTERSPEECH*.

Adam Polyak, Lior Wolf, Yossi Adi, and Yaniv Taigman. 2020. Unsupervised cross-domain singing voice conversion. In *Proc. INTERSPEECH*, pages 801–805.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer. 2011. CROWDMOS: An approach for crowdsourcing mean opinion score studies. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2416–2419.

Morgane Rivière and Emmanuel Dupoux. 2020. Towards unsupervised learning of speech features in the wild. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 156–163.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Elizabeth Shriberg and Andreas Stolcke. 2004. Prosody modeling for automatic speech recognition and understanding. In *Mathematical Foundations of Speech and Language Processing*, pages 105–114. Springer.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication*, 32(1-2):127–154.

Elizabeth Shriberg, Andreas Stolcke, Daniel Jurafsky, Noah Coccaro, Marie Meteer, Rebecca Bates, Paul Taylor, Klaus Ries, Rachel Martin, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech*, 41(3-4):443–492.

Yi Su and Frederick Jelinek. 2008. Exploiting prosodic breaks in language modeling with random forests. In *Speech Prosody*. Citeseer.

SynSIG. Blizzard Challenge 2013. https://www.synsig.org/index.php/Blizzard_Challenge_2013. Accessed: 2021-08-23.

Andros Tjandra, Ruoming Pang, Yu Zhang, and Shigeki Karita. 2020. Unsupervised learning of disentangled speech content and style representation. *arXiv preprint arXiv:2010.12973*.

Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf. 2017. Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information. *arXiv preprint arXiv:1704.07287*.

Aaron van den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. In *NeurIPS*.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2016. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.

Nigel G Ward, Alejandro Vega, and Timo Baumann. 2012. Prosodic and temporal features for language modeling for dialog. *Speech Communication*, 54(2):161–174.

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, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: speech processing universal performance benchmark. *CoRR*, abs/2105.01051.

A Analysis of Log F0 Distribution

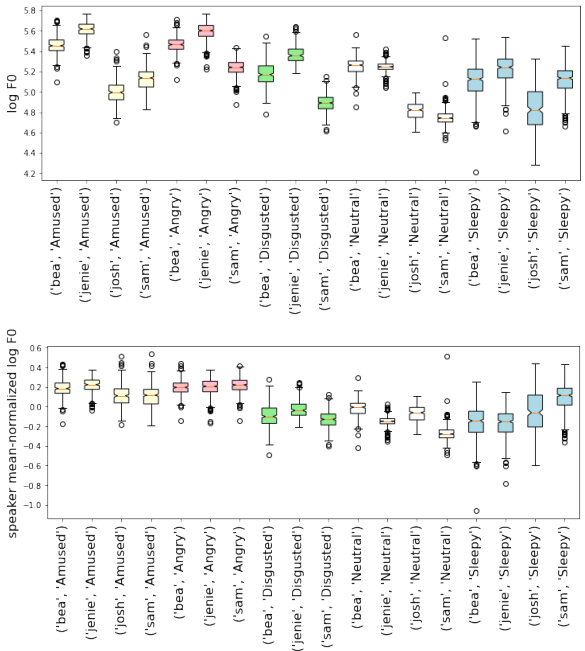


Figure A.1: Log F0 distribution of each (speaker, emotion) combination in the EmoV dataset without speaker mean normalization (top) and with speaker normalization (bottom). Each color corresponds to one emotion.

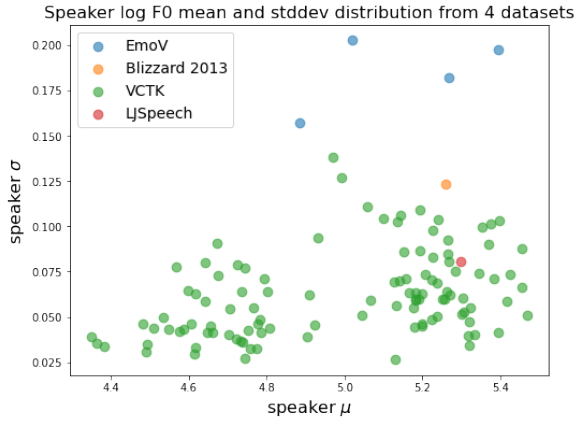


Figure A.2: Speaker log F0 mean and standard deviation distributions from two expressive datasets (EmoV and Blizzard 2013) and two plain datasets (LJSpeech and VCTK). Each point corresponds to one speaker.

B HiFi-GAN Adaptation Analysis

Table 4 presents an analysis of HiFi-GAN performance when using different quantized pitch representations. Similarly to (Polyak et al., 2020) we report voice decision error (VDE) (Nakatani et al., 2008), which measures the portion of frames with voicing decision error and F0 Frame Error

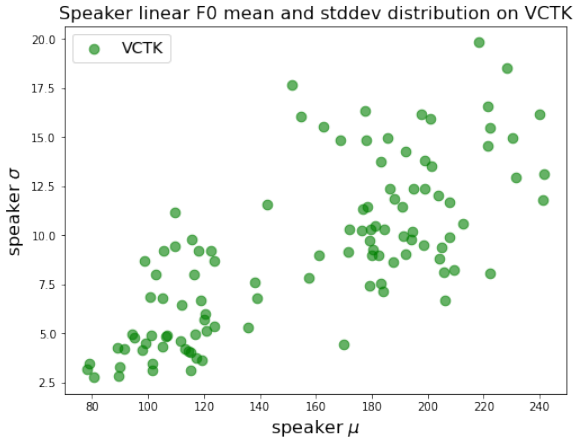


Figure A.3: Speaker linear F0 mean and standard deviation distributions on VCTK. Each point corresponds to one speaker.

(FFE) (Chu and Alwan, 2009), which measures the portion of frames that contain a deviation of more than 20% in pitch value or have a voicing decision error. Results show that the chosen quantizer achieve favorable performance in terms of VDE and comparable results in terms of FFE without having to pre-train a F0 VQ-VAE quantizer.

C Example of Converting Frame-Level to Segment-Level Representations

Assume we have an utterance of six frames: [(13, 1.5), (13, 2.5), (13, 0.0), (21, 0.0), (27, 1.3), (27, 3.5)] where the first number in each tuple denotes the unit of the frame and the second number denotes the speaker normalized log F0 of the frame. In particular, the third and the fourth frame are unvoiced and their lf values are set to 0.0.

The segment level representation of the utterance is [(13, 3, 2.0), (21, 1, 0.0), (27, 2, 2.4)]. The first segment (13, 3, 2.0) is labeled with unit $u = 13$, duration $d = 3$ frames, and an average normalized log F0 $lf = (1.5+2.5)/2 = 2.0$ for the two voiced frames. The second segment contains only one unvoiced frame, and hence lf is set to 0. Finally, the last segment contains two voiced frames, and therefore $d = 2$ and $lf = (1.3 + 3.5)/2 = 2.4$.

D Effects of F0 Representation on MS-TLM

Table 5 compares content modeling performance when using different pitch representations. Results show that using mean normalized pitch information is better than using raw pitch, and using log pitch is better than using linear pitch.

F0 scale	F0 norm.	F0 res.	F0 quant.	FFE↓	VDE↓
(Polyak et al., 2021)					
lin	mean+std	frm	VQ-VAE (V)	0.223	0.169
lin	mean+std	frm	VQ-VAE (B)	0.198	0.181
lin	mean+std	frm	naive	0.172	0.138
lin	mean+std	seg	VQ-VAE (V)	0.149	**0.116**
lin	mean	frm	VQ-VAE (V)	0.220	0.178
log	mean+std	frm	VQ-VAE (V)	0.388	0.188
(This work)					
log	mean	seg	naive	**0.134**	0.118

Table 4: Speech resynthesis results on the Blizzard 2013 validation set with segment-level pitch information. All the HiFi-GAN models are trained on the Blizzard 2013 training set using the same HuBERT units but different quantized pitch representations. **"scale"** denotes the F0 scale, which is linear (*lin*) or logarithmic (*log*). **"norm."** denotes the normalization method applied to F0, which is normalizing by *mean* or by mean and standard deviation (*mean+std*). **"res."** denotes the resolution of pitch in the training time, where "*frm*" refers to using frame-level pitch and "*seg*" refers to using segment-level pitch (pitch values are set to the average for all frames within a segment). **"quant."** denotes the F0 quantizer, where *VQ-VAE (V)* is a neural pitch quantizer pre-trained on VCTK, *VQ-VAE (B)* is one pre-trained on Blizzard, and *naive* is the one adopted in this work.

E More Details of Human Evaluation

The instruction page displayed to the raters are shown in Figure E.1. We modify the *Introduction*, *Task Instruction*, *Example* in the instruction page for MOS, MMOS, and PMOS correspondingly. The text used for each metric are detailed in Table 6

F0 scale	F0 norm.	u NLL
linear	none	1.763
linear	mean	1.564
log	none	1.461
(This work)		
log	mean	**1.447**

Table 5: Content modeling performance of base MS-TLM models trained on LS960 with HuBERT units using different pitch representations without quantization. **"scale"** denotes the F0 scale, which is linear (*lin*) or logarithmic (*log*). **"norm."** denotes the normalization method applied to F0, which is not normalizing (none) or normalizing by mean.

Instructions for speech quality evaluation

Introduction

Your task is to evaluate the subjective quality of the speech from short (2-8 second) audio files. Each HIT can be completed in roughly around 120 seconds.

Payment

We have methods that analyze the consistency of your answers with respect to themselves, to those of your fellow workers and to references we know to be accurate. We will use these methods to rank the submitted assignments according to quality.

For this experiment we will pay a base reward of $0.20/HIT for every accepted HIT. You will receive a bonus of:

- $0.05/HIT (for a total of $0.25/HIT) if you submit 12 or more HITs or
- $0.10/HIT (for a total of $0.30/HIT) if you submit 12 or more HITs and your results are among the top 50% or
- $0.20/HIT (for a total of $0.35/HIT) if you submit 12 or more HITs and your results are among the top 10%.

Bonuses will be paid up to 7 days after submission, because we can only rank the submissions once we have a statistically significant number of answers. The base reward will always be paid within 48 hours of submission.

Task Instructions

In this task you will hear samples of speech recordings. The purpose of this test is to evaluate the quality of each file in terms of its overall sound quality and the amount of mumbling and unclear phrases in the recording.

Please keep in mind that speech samples can be distorted and noisy, however these are only specific examples.

Please use a headset for listening and adjust your volume level to your comfort during this training, and do not change later during the experiment.

You should give a score according to the following scale, known as the MOS (mean opinion score) scales:

Score	Quality of the audio
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

IMPORTANT: Note that some of the sentences truncate unexpectedly, which is OK.

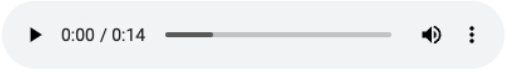
Example

The following recording represents a speech recordings with a reference score of: **5.0 (Excellent)**.

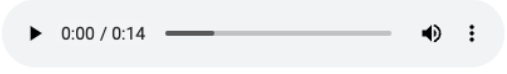
The following recording represents a speech recordings with a reference score of: **5.0 (Excellent)**.

The following recording represents a speech recordings with a reference score of: **3.0 (Fair)**.

The following recording represents a speech recordings with a reference score of: **1.0 (Bad)**.



The following recording represents a speech recordings with a reference score of: **1.0 (Bad)**.



Approval/Rejection Policy

Your results will be collected and evaluated for consistency. We (the requesters) have an estimate of each file's subjective quality that conforms with the references above. Thus, we can detect if someone submits random scores or does not rate according to these instructions, which can lead to work being rejected. You can rest assured that your work will be approved if you rate according to the instructions above.

Answers will be either reviewed or automatically approved within 48 hours.

Figure E.1: The instruction page for human evaluation on MOS.

Metric	Introduction	Metric-Specific Task Instruction
MOS	Your task is to evaluate the **subjective quality** of the speech from short (2-8 second) audio files. Each HIT can be completed in roughly around 120 seconds.	...The CONTINUATION has been generated by a computer and your task will be to concentrate specifically on it and evaluate its **quality** in terms of the **sound clarity** on a 1 to 5 scale, (irrespective of the intonation or meaning)
MMOS	Your task is to evaluate the **subjective meaningfulness** of the speech from short (2-8 second) audio files. Each HIT can be completed in roughly around 120 seconds.	...The CONTINUATION has been generated by a computer and your task will be to concentrate specifically on it and evaluate its **meaning** in terms of **grammar and content** on a 1 to 5 scale, (irrespective of sound clarity or intonation)
PMOS	Your task is to evaluate the **subjective prosodic coherence** of the speech from short (2-8 second) audio files. Each HIT can be completed in roughly around 120 seconds.	...The CONTINUATION has been generated by a computer and your task will be to concentrate specifically on it and evaluate its **naturality** in terms of **intonation and rhythm** on a 1 to 5 scale, (irrespective of sound clarity or meaning)

Table 6: The *Introduction* and Task Instruction used for the three human evaluation metrics. All the *Task Instruction* starts with "In this task, you will hear samples of speech recordings, composed of the following parts: [PROMPT] beep [PROMPT CONTINUATION]. [PROMPT] is a sentence beginning, said by one voice. ("ANOTHER PREACHER AFTER REPROACHING HIM" in a male voice). Beep is a short tone. [PROMPT CONTINUATION] a longer sentence in a different voice, which starts with the initial prompt ("ANOTHER PREACHER AFTER REPROACHING HIM" in a female voice) and adds a few more words that continue this prompt ("TO HIS FACE WITH HIS MISGOVERNMENT ORDERED THIS PSALM TO BE SUNG" in the same female voice)."