On Generative Spoken Language Modeling from Raw Audio

Kushal Lakhotia[*], **Eugene Kharitonov**[*], **Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte**[§], **Tu-Anh Nguyen**[†], **Jade Copet, Alexei Baevski, Abdelrahman Mohamed, Emmanuel Dupoux**[‡]
Facebook AI Research

Author's version of Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., & Dupoux, E. (2022). On Generative Spoken Language Modeling from Raw Audio, *Transactions of the Association for Computational Linguistics*

Abstract

We introduce *Generative Spoken Language Modeling*, the task of learning the acoustic and linguistic characteristics of a language from raw audio (no text, no labels), and a set of metrics to automatically evaluate the learned representations at acoustic and linguistic levels for both encoding and generation. We set up baseline systems consisting of a discrete speech encoder (returning pseudo-text units), a generative language model (trained on pseudo-text), and a speech decoder (generating a waveform from pseudo-text) all trained without supervision and validate the proposed metrics with human evaluation. Across 3 speech encoders (CPC, wav2vec 2.0, HuBERT), we find that the number of discrete units (50, 100, or 200) matters in a task-dependent and encoder-dependent way, and that some combinations approach text-based systems.[1]

1 Introduction

An open question for AI research is creating systems that learns from natural interactions as infants learn their first language(s): from raw uncurated data, and without access to text or expert labels (Dupoux, 2018). Natural Language Processing (NLP) systems are currently far from this requirement. Even though great progress has been made in reducing or eliminating the need for expert labels through self-supervised training objectives (Brown et al., 2020; Peters et al., 2018; Radford et al., 2019; Devlin et al., 2019; Liu et al., 2019b; Dong et al., 2019; Lewis et al., 2020), the basic units on which these systems are trained are still textual. Yet, young children learn to speak several years before they can read and write, providing a proof of principle that language can be learned without any text. Being able to achieve 'textless NLP' would be beneficial for the majority of the world's languages which do not have large textual resources or even a widely used standardized orthography (Swiss German, dialectal Arabic, Igbo, etc.), and which, despite being used by millions of users, have little chance of being served by current text-based technology. It would also be useful for 'high-resource' languages, where the oral and written forms often mismatch in terms of lexicon and syntax, and where some linguistically relevant signals carried by prosody and intonation are basically absent from text. While text is still the dominant form of language on the web, a growing amount of audio resources like podcasts, local radios, social audio apps, on-line video games provide the necessary input data to push NLP to an audio-based future and thereby expand the inclusiveness and expressivity of AI systems.

Is it possible to build an entire dialogue system from audio inputs only? This is a difficult challenge, but breakthroughs in unsupervised representation learning may address part of it. Unsupervised learning techniques applied to speech were shown to learn continuous or discrete representations that capture speaker invariant phonetic content (Versteegh et al., 2016; Dunbar et al., 2020), despite themselves not being phonemic (Schatz et al., 2021). Recent developments in self-supervised learning have shown impressive results as a pretraining technique (van den Oord et al., 2017; Chung et al., 2019; Hsu et al., 2021), to the extent that Automatic Speech Recognition (ASR) on par with the state of the art from two years back can be built with 5000 times less labelled speech (Baevski et al., 2020b), or even no with no labelled speech at all (Baevski et al., 2021). Of course, ASR still assumes access to text to learn a lan-

[*] equal contribution. [‡] Also at EHESS. [†] Also at INRIA. [§] Work done while at FAIR.

[1] Evaluation code and trained models are here: https://github.com/pytorch/fairseq/tree/master/examples/textless_nlp/gslm. Sample audios are here: https://speechbot.github.io/gslm.

Level	Encoding		Generation		
	Task	Automatic metric	Task	Automatic metric	Human
Language	Spoken LM	**Spot-the-word**, Syntax-Acc	Speech Gen.	**AUC-of-VERT/PPX**, cont-BLEU, PPX@o-VERT	**MMOS**
Acoustic	Acoustic Unit Disc.	**ABX-across**, ABX-within	Resynthesis	**PER-from-ASR**, CER-from-ASR	CER, **MOS**

Table 1: **Tasks and metrics** proposed to evaluate encoding/generation quality of models at the acoustic or language levels. Bold fonts highlights the main metric used for each category (Section 3 for details).

guage model (LM) and the mapping to the audio units. Here, we study the case where the LM is directly trained from the audio units without any recourse to text.

The high level idea (see Figure 1) is that automatically discovered discrete units can be used to encode speech into "pseudo-text" (speech-to-unit, S2u), which is used in turn to train a generative language model (unit-based language model, uLM) and to train a speech synthesizer (unit-to-speech, u2S). This enables learning an LM from scratch without text, and use it to generate speech conditionally or unconditionally, essentially replicating what toddlers achieve before learning to read. Early studies using discrete codes learned from an autoencoder show the feasibility of such an approach, but remain at a level of a demo (van den Oord et al., 2017).

In this paper, we address one major conceptual stumbling block which has, thus far, prevented such early studies from having the transformative impact they could have in language technology: *model evaluation*. We contend that it will be impossible to make progress in this area beyond demos unless proper evaluation methods enabling system comparison are established.

Evaluation for speech generation is difficult due to the continuous, variable and multi-level nature of the speech waveform, and the necessity both to capture fine grained acoustic details to generate intelligible audio and to abstract away from them to learn higher level language concepts. Text-based models do not have this problem, since the input is already expressed in terms of mid-level discrete units (characters or words), and are typically evaluated with unsupervised metrics close to the learning objectives like perplexity or log likelihood. Here, such an approach is not directly applicable even if we rely on discrete pseudo-text units, since such metrics would depend in an unknown fashion on their granularity (number, duration and distribution), making the comparison of models that use different units infeasible.

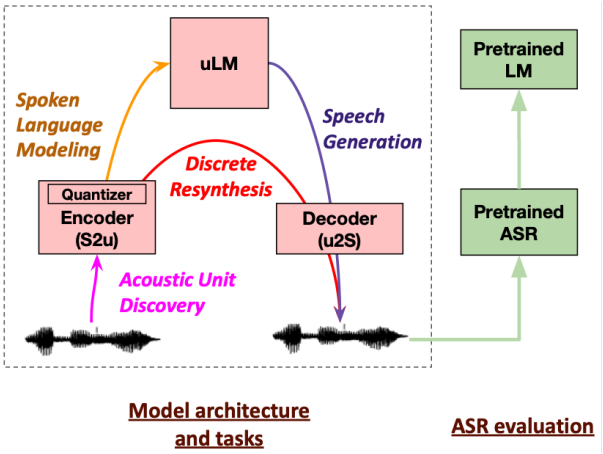

Figure 1: **Setup of the baseline model architecture, tasks and metrics.**

Conceptually, generative spoken language models can be evaluated at two levels, the acoustic and the language levels, and through two modes of operation, encoding and generation, resulting in 2x2 tasks (see Table & Figure 1). *Acoustic Unit Discovery* (encoding at the acoustic level) consists in representing speech in terms of discrete units discarding non-linguistic factors like speaker and noise. *Spoken Language Modeling* (encoding at the language level) consists in learning the probabilities of language patterns. *Speech Resynthesis* (generation for acoustic modeling) consists in generating audio from given acoustic units. This boils down to repeating in a voice of choice an input linguistic content encoded with speech units. *Speech Generation* (generation for language modeling) consists in generating novel and natural speech (conditioned on some prompt or not). Compared to standard text generation, a critical and novel component of the audio variant is clearly the discovery of units since it conditions all the other components. This is why we devote our analyses of model architectures to the unit-to-speech component specifically, and leave it for further work to evaluate how the downstream components can

also be optimized for spoken language generation.

The major contributions of this paper are as follows : (1) we introduce two novel evaluation metrics for the generation mode of spoken language modeling at the acoustic and language levels respectively. Our key insight is to use a generic pretrained ASR system to establish model-independent assessments of the intelligibility (acoustic level) and meaningfulness (language level) of the produced outputs. The ASR system converts the generated waveform back to text, enabling us to adapt standard text-based metrics for these two levels. (2) we validate these metrics through comparison with human evaluation. We show a high degree of concordance between human and machine evaluations of intelligibility and meaningfulness of generated audio. (3) we show that these metrics can be predicted by simpler ones geared to evaluate the encoding mode of the spoken LM. Zero-shot metrics borrowed from previous studies in the Zero Resource Speech Challenges (Versteegh et al., 2016; Nguyen et al., 2020) correlate well with their generative counterpart, offering an easier proxy to rapidly iterate on model selection. (4) we systematically study the effect of the type of encoding units by factorially crossing three recent speech-to-unit encoders, CPC, Wave2vec 2.0 and HuBERT, with three codebook sizes for the discrete units, 50, 100, 200. We keep constant the rest of the system built from out-of-the-box components (standard Transformer for the uLM, Tacotron 2 for u2S). We show that both the encoder type and the number of units matter, and that they matter differently depending on the evaluation task. (5) we open source our evaluation tools and models to help reproducibility and comparability with future work.

In Section 3, we introduce the ASR, zero-shot and human evaluation metrics, in Section 4 we present the models, in Section 5, we analyze the results and discuss them in Section 6.

2 Related work

Unsupervised speech representation learning aims to distill features useful for downstream tasks, such as phone discrimination (Kharitonov et al., 2021; Schneider et al., 2019) and semantic prediction (Lai et al., 2021; Wu et al., 2020), by constructing pretext tasks that can exploit large quantities of unlabeled speech. Pretext tasks in the literature can be roughly divided into two categories: reconstruction and prediction. Reconstruction is often implemented in the form of auto-encoding (Hsu et al., 2017a), where speech is first encoded into a low-dimensional space, and then decoded back to speech. Various constraints can be imposed on the encoded space, such as temporal smoothness (Ebbers et al., 2017; Glarner et al., 2018; Khurana et al., 2019, 2020), discreteness (Ondel et al., 2016; van den Oord et al., 2017), and presence of hierarchy (Lee and Glass, 2012; Hsu et al., 2017b). Prediction-based approaches task a model with predicting information of unseen speech based on its context. Examples of information include spectrograms (Chung et al., 2019; Wang et al., 2020; Chi et al., 2021; Liu et al., 2020; Chung and Glass, 2020; Liu et al., 2020; Ling et al., 2020; Ling and Liu, 2020), cluster indices (Baevski et al., 2019; Hsu et al., 2021), derived signal processing features (Pascual et al., 2019; Ravanelli et al., 2020), and binary labels of whether a candidate is the target unseen spectrogram (van den Oord et al., 2018; Schneider et al., 2019; Baevski et al., 2020a; Kharitonov et al., 2021; Baevski et al., 2020b).

Speech resynthesis. Recent advancements in neural vocoders enabled generating natural sounding speech and music (Oord et al., 2016; Kumar et al., 2019; Kong et al., 2020). These are often conditioned on the log mel-spectrogram for the generation process. Learning low bitrate speech representations in an unsupervised manner, has attracted attention from both the machine learning and the speech communities (Liu et al., 2019a; Feng et al., 2019; Nayak et al., 2019; Tjandra et al., 2019; Schneider et al., 2019; Baevski et al., 2020a; Chen and Hain, 2020; Morita and Koda, 2020; Tobing et al., 2020). These representations can later be used for generation without text, which is particularly important for low-resource languages (Dunbar et al., 2019, 2020). van den Oord et al. (2017) proposed a Vector-Quantized Variational Auto-Encoder (VQ-VAE) model to learn discrete speech units, which will be later used for speech synthesis using a WaveNet model. Eloff et al. (2019) suggested a VQ-VAE model followed by a FFTNet vocoder model (Jin et al., 2018). Tjandra et al. (2020) suggested to use transformer (Vaswani et al., 2017) together with a VQ-VAE model for unsupervised unit discovery, and van Niekerk et al. (2020) combines vector quantization together with contrastive predictive coding for acoustic unit discovery. Another line

of work use representations from an ASR acoustic model that are combined with identity and prosodic information for voice conversion (Polyak et al., 2020b,a, 2021b). In terms of evaluation, the Zero-Resource challenge (Dunbar et al., 2019, 2020; Nguyen et al., 2020) used bitrate together with human evaluation. In this paper we additionally introduce an ASR based evaluation metric.

3 Evaluation Methods

We present two sets of automatic evaluation metrics; the first ones assess the output of generative speech models (ASR metrics, Section 3.1); the second ones, the encoded representations (zero-shot probe metrics, Section 3.2). Finally, we present the human evaluations (Section 3.3).

3.1 Generation: ASR metrics

We present our new evaluation metrics for generation tasks. The first task, speech resynthesis, involves S2u which encodes input speech into units and u2S which decodes it back to speech. In this task, we wish to evaluate *intelligibility* of the resulting speech. The second task, speech generation, involves the full S2u→uLM→u2S pipeline, and we wish to evaluate *meaningfulness* of the generated speech. Our overall idea is to use ASR to convert the generated speech back to text and then use text-based metrics.

Speech resynthesis intelligibility: ASR-PER. The ideal metric for intelligibility would be to use humans to transcribe the resynthesized speech and compare the text to the original input. An automatic proxy can be obtained by using a state-of-the-art ASR system pretrained on a large corpus of real speech.[2] Our main metric is Phone Error Rate (PER), which only uses an acoustic-model ASR, without fusing with an additional language model (Chorowski and Jaitly, 2016). In preliminary experiments we also experimented with a full ASR with an LM and computed Word Error Rate (WER) and Character Error Rate (CER) to give partial credit. The latter is probably closer to humans intelligibility metrics, as humans cannot turn off their lexicon or language model. We also computed such metrics by training a fitted ASR model for each resynthesis model on a specific training corpus (see Supplementary Section S2.1). The logic of this last test is that it provides a more direct measure of the information lost in the S2u→u2S pipeline, because it could adapt to systematic errors introduced by the u2S model. Since the scores between these different approaches correlated highly, we only report here the results on the PER for a pretrained ASR model which is the simplest to deploy.

Speech generation quality and diversity: AUC on Perplexity and VERT. Text generation evaluation typically involves two axes: the quality of the generated text (with automatic metrics like mean perplexity or negative log likelihood computed on a reference large language model) and the diversity (with metrics like self-BLEU[3], Zhu et al., 2018). Typically, there is a trade-off between these two dimensions based on the temperature hyperparameter used for sampling from the language model, whereby at low temperature, the system outputs good sentences but not varied, and at high temperatures, it outputs varied sentences, but not very good. This results in model comparison being either based on 2D plots with lines representing the trade-off between quality and diversity, or on aggregate metrics like the area under the curve. Preliminary explorations (see Appendix Section 7.2) with our models revealed two problems preventing a straightforward application of such a scoring strategy.

First, we found that for some models, at a low enough temperature, self-BLEU score stopped increasing, but the systems started to repeat more and more words within a sentence (e.g., "the property the property the property"). We therefore introduce a new metric, auto-BLEU, that measures within-sentence diversity. For a single utterance u, auto-BLEU is calculated as the ratio of k-grams $s \in NG_k(u)$ that are repeated at least once:

$$\texttt{auto-BLEU}(u,k) = \frac{\sum_s \mathbb{1}\left[s \in (NG_k(u) \backslash s)\right]}{|NG_k(n)|} \quad (1)$$

As with BLEU score, to get n-gram auto-BLEU we calculate the geometric mean of $\texttt{auto-BLEU}(u,k)$ obtained for $k \in [1,n]$ and average over the set of generated utterances. By calculating the geometric mean of self- and auto-BLEU, we obtain an aggregate metric which we call VERT (for diVERsiTy). We used a bigram version of self- and auto-BLEU.

Second, we found that critical temperatures for which the output was reasonable were not constant

[2] We use a BASE wav2vec 2.0 phoneme detection model trained on LibriSpeech-960h with CTC loss from scratch.

[3] Higher self-BLEU scores indicate lower diversity of the produced text.

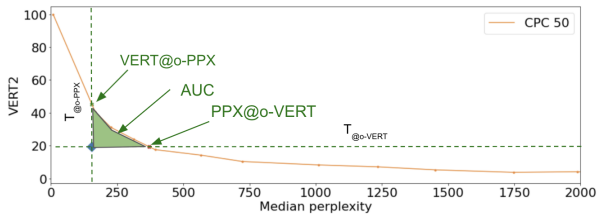

Figure 2: **Comparison of diversity and perplexity of the generated speech.** We plot VERT vs. Median perplexity. The blue diamond corresponds to the oracle reference point. It defines two cut-offs on the curve: VERT @oracle-PPX and PPX @oracle-VERT. The green area corresponds to the AUC metric.

across models. This makes sense, because temperature controls the probability of sampling individual units, and the probabilistic distribution and duration of these units depend on the models. Here, we chose to use the oracle text as an anchor to compute reference temperatures, i.e., the temperatures at which the perplexity or the VERT score reach the values of the oracle text.

This gives us boundary conditions at which we can compare (the perplexity at oracle diversity and the diversity at oracle perplexity), as well as a method to compute the area under curve (AUC) between these two boundaries (See Figure 2). As AUC decreases, the system gets closer to the oracle point. Thus with AUC, lower is better.

To calculate perplexity of the generated utterances, we use a pre-trained ASR[4] to convert speech to text, and an off-the-shelf Transformer model trained on the English NewsCrawl dataset.[5]

3.2 Encoding: Zero-shot probe metrics

The purpose of the encoding metrics is to evaluate the quality of the learned representations at each linguistic level along the pipeline linking the S2u and the uLM. They are inspired by human psycholinguistics and can be be thought of as *unit tests* providing interpretation and diagnosis. We entirely draw on evaluations from the Zero Resource challenge series (Versteegh et al., 2016; Dunbar et al., 2019; Nguyen et al., 2020)[6] for comparability with published work and refer to these challenges for details. These metrics are "zero-shot" because they do not require training any classifier, and are either based on distances over embeddings, or on computing probabilities over entire utterances. When they have hyperparameters, these are selected using a validation set.

For acoustic-level evaluation, we use the between-speaker **ABX score** to quantify how well-separated phonetic categories are. Briefly, it consists in estimating the probability that two tokens of the same category A (x and a) are closer to one another than a token of A (x) and of B (b). The categories are triphones that only differ in the middle phoneme (like *bit* and *bet*) and the score is averaged over all possible such pairs. For the across-speaker ABX, a and b are spoken by the same speaker and x by a different one, requiring feature invariance over a speaker change. We also include the **bitrate** which has been used in the TTS-without-T challenges (Dunbar et al., 2019) to quantify the efficiency of the discrete units used to resynthetize speech. It is simply the entropy of the sequence of units divided by the total duration.

For language-level evaluation, we use **spot-the-word accuracy** from the Zero Resource 2021 Benchmark (Nguyen et al., 2020). It consists in detecting the real word from a pair of short utterances like 'brick' vs 'blick', matched for unigram and bigram phoneme frequency to ensure that low-level cues do not make the task trivial. This task can be done by computing the probability (or pseudo-probability) of the utterances from the uLM. The test set (sWUGGY) consists of 5,000 word-pseudoword pairs generated by the Google TTS API, filtered for the word being present in the LibriSpeech 960h training set (Panayotov et al., 2015). The ZR21 benchmark also uses higher level metrics, notably, syntactic (based on the sBLIMP dataset), which we did not use because the baselines were too close to chance.

3.3 Human evaluation metrics

As above, we asked humans to evaluate two aspects of speech generation: intelligibility and meaningfulness. Intelligibility was assessed using two metrics: i) Mean Opinion Scores (MOS) in which raters were asked to evaluate subjectively how intelligible a given audio sample is; ii) Character Error Rate (CER) computed from written transcriptions providing an objective intelligibility test. As for meaningfulness, we set up a meaningfulness-MOS (MMOS) in which raters were asked to evaluate how natural (considering both grammar and meaning) a given sample is. For both subjective tests raters evaluate the samples on

[4] We use a LARGE wav2vec 2.0 model, trained on LibriSpeech-960h with CTC loss from scratch. Its decoder uses the standard KenLM 4-gram language model.

[5] github.com/pytorch/fairseq/.../language_model

[6] www.zerospeech.com

a scale of 1-5 with an increment of 1.

For the MMOS, we had to select a temperature to sample from. Preliminary experiments showed that humans preferred lower temperatures (yielding also less diverse outputs, see Supplementary Section S2.2). Here, we settled on selecting the temperature on a model-by-model basis by constructing a continuation task: we take the 1,000 shortest utterances from LibriSpeech test-clean that are at least 6 seconds long, and use the first 3 seconds as prompts for the uLM (after transcribing them into pseudo-texts). For each prompt, we generated 10 candidate continuations of the same length (in seconds) as the utterance which we took the prompt from. We varied temperature (0.3, 0.4, ..., 1.4, 1.5, 1.7, 1.9, 2.1, 2.3, 2.5, 3.0), and selected the one yielding the maximal BLEU-2 score with the reference sentence (after ASR). These temperatures were typically between the two boundary temperatures described above.

We evaluated 100 samples from each of the evaluated methods while we enforced at least 15 raters for each sample. The CrowdMOS package (Ribeiro et al., 2011) was used for all subjective experiments using the recommended recipes for detecting and discarding inaccurate scores. The recordings for the naturalness test were generated by the LM unconditionally and conditionally from a 3 seconds prompt. Participants were recruited using a crowd-sourcing platform.

4 Proposed Systems

Here, we present our S2u (Section 4.1), uLM (Section 4.2) and u2S (Section 4.3) components.

4.1 Speech-to-unit Models

We selected 3 recent state-of-the-art unsupervised Encoders, which we used 'out of the box': we did not retrain them nor change their hyperparameters. We also included a log Mel filter-bank baseline (80 filters, computed every 10ms). We then discretized the embeddings using k-means. We only give a high level description of these models, and refer to the original publications for details.

CPC. Contrastive Predictive Coding (van den Oord et al., 2017) as applied to speech consists of two components: an encoder and a predictor. The encoder produces an embedding z from speech input. The predictor predicts the future states of the encoder based on the past, and the system is trained with a contrastive loss. We use the CPC model from (Rivière and Dupoux, 2020), which was trained on a "clean" 6k hour sub-sample of the LibriLight dataset (Kahn et al., 2020; Rivière and Dupoux, 2020). We extract a representation from an intermediate layer of the predictor, which provides a 256-dimensional embedding (one per 10ms), as in the original paper.

wav2vec 2.0. Similar to CPC, this model uses an encoder and a predictor, which is trained contrastively to distinguish positive and negative samples from discretized and masked segments of the encoder's output. We use the LARGE variant of pretrained wav2vec 2.0 (Baevski et al., 2020b) trained on 60k hours of LibriLight dataset (Kahn et al., 2020). This model encodes raw audio into frames of 1024-dimensional vectors (one per 20ms). To choose the best layer, we extracted frozen representations of the 10 hour LibriLight subset from every layer of the model and trained a linear classifier with the CTC loss to predict the phonetic version of the text labels. Layer 14 obtained the lowest PER on LS dev-other (a similar approach was done in (Baevski et al., 2021) which in this case selected Layer 15).

HuBERT. Unlike CPC and wav2vec 2.0 that use a contrastive loss, HuBERT is trained with a masked prediction task similar to BERT (Devlin et al., 2019) but with masked continuous audio signals as inputs. The targets are obtained through unsupervised clustering of raw speech features or learned features from earlier iterations, motivated by DeepCluster (Caron et al., 2018). We use the BASE 12 transformer-layer model trained for two iterations (Hsu et al., 2021) on 960 hours of LibriSpeech (Panayotov et al., 2015). This model encodes raw audio into frames of 768-dimensional vectors (one per 20ms) at each layer and we extract those from the 6^{th} layer as in the original paper.

LogMel. As a baseline, we consider a Log Mel Filterbank encoder using 80 frequency bands.

Quantization. We use k-means to convert continuous frame representations into discrete representation by training on LibriSpeech clean-100h (Panayotov et al., 2015). We experiment with codebooks that have 50, 100, and 200 units.

4.2 unit-Language Model

We use the Transformer model as implemented in `fairseq` (Ott et al., 2019). We use the *transformer_lm_big* architecture: it has 12 layers, 16 attention heads, embedding size of 1024, FFN size

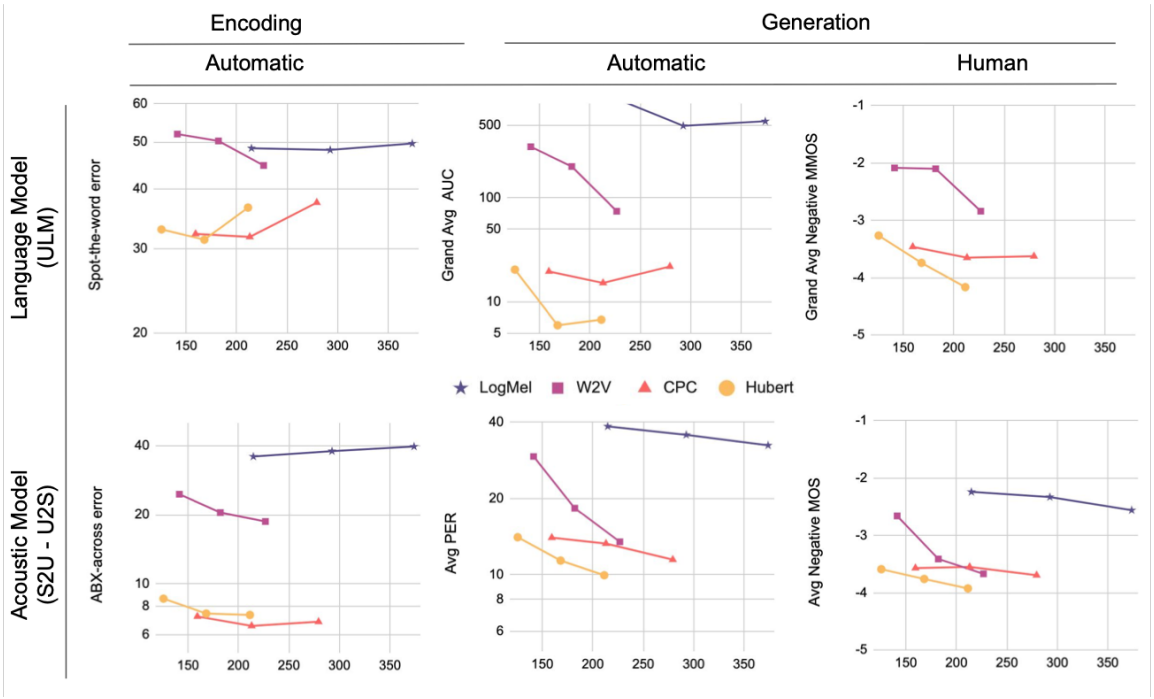

Figure 3: **Overall results with automatic and human metrics**. The results are presented in terms of bitrate for 4 encoders (LogMel, CPC, HuBERT and wav2vec 2.0) varying in number of units (50, 100, 200). For definition of the tasks and metrics, see Table 1 and Figure 1. Negative human opinion scores are shown for ease of comparison with automatic metrics (lower is better). The generation metrics have been averaged across LS and LJ (PER and MOS; resynthesis task) and across prompted and unprompted conditions (AUC and MMOS; speech generation task). The Log Mel Fbank based systems were not evaluated by humans in the speech generation task.

of 4096, and dropout probability of 0.1, and we train it as a causal LM on sequences of pseudo-text units. Each sample contains up to 3,072 units. We use sampling with temperature for generation.

All language models are trained on "clean" 6k hours sub-sample of LibriLight used in (Rivière and Dupoux, 2020), transcribed with corresponding discrete units. In preliminary experiments, we found that removing sequential repetitions of units improves performance, hence we apply it universally.[7] We hypothesise that this simple modification allows to use Transformer's limited attention span more efficiently as in Hsu et al. (2020).

4.3 unit-To-Speech Model

We adapt the *Tacotron-2* model (Shen et al., 2018) such that it takes pseudo-text units as input and outputs a log Mel spectrogram. To enable the model to synthesize arbitrary unit sequences, including those representing incomplete sentences, we introduce two modifications. First, we append a special "end-of-input" (EOI) token to the input sequence, hinting the decoder to predict the "end-of-output" token when attending to this new token. However, this modification alone may not be sufficient, as the decoder could still learn to ignore the EOI token and correlate end-of-output prediction with the learned discrete token that represents silence as most of the speech contains trailing silence. To address this, we train the model using random chunks of aligned unit sequence and spectrogram, and append the EOI token to unit sequence chunks, such that the audio does not always end with silence. We implement chunking in the curriculum learning fashion, where the chunk size gradually grows (starting with 50 frames with an increment of 5 per epoch) to increase the difficulty of the task. For waveform generation, we use the pre-trained flow-based neural vocoder *WaveGlow* (Prenger et al., 2019). This model outputs the time-domain signal given the log Mel spectrogram as input. All u2S models were trained on LJ Speech (LJ) (Ito and Johnson, 2017).

[7]For example, a pseudo-text `10 11 11 11 21 32 32 32 21` becomes `10 11 21 32 21`.

5 Results

In Figure 3, we report the overall results of our models and our LogMel baseline as a function of the number of quantized units on our main automated and human metrics. More detailed results follow in the following sections, including

two character-based toplines: one uses the oracle transcripts for training the LM, the other uses transcripts produced by the pre-trained ASR model.

5.1 Results on the resynthesis task

Overall resynthesis results are shown in the bottom middle and right cells of Figure 3 for our main automatic (PER) and human scores (MOS), respectively, averaged across the LS and LJ evaluation sets. We observe that across all models, increasing the number of units uniformly leads to better scores suggesting that the u2S component can take benefit from extra details of the input to produce a more realistic output. HuBERT and CPC seem to be giving the best results, for both humans and models better capturing phonetic information than other models at equivalent bitrates.

More detailed results are in Table 2 separating the scores for the LJ and LS resynthesis, and adding extra automatic metrics (CER) and human metrics (human CER). On PER, we found a domain effect: resynthesizing input from LJ Speech yields lower PER than from LibriSpeech on all unsupervised models. From the viewpoint of the encoder, LJ Speech is out-of-domain; therefore, one would expect that the units are making more errors than for the trained LibriSpeech. On the other hand, the u2S component has learned from LJ Speech encoded with these units, and might have learned to compensate for these lower quality units. When LibriSpeech is offered as input, the u2S component cannot adapt to this nominally better input and ends up yielding lower quality outputs. This observation is worth further explorations, as other metrics like CER (using an LM) and human evaluations only replicated this for the models with the lowest score (like LogMel and wav2vec). The automatic PER and CER scores and the human MOS and CER scores, all correlate well with one another across the 4×3 models and baselines. Within the LJ or LS domain , the Pearson r ranged from .95 to .99; across domains it was less good (from .79 to .96) illustrating again the existence of a domain effect. Not shown here, we reached similar conclusions with our fitted-ASR metrics, but with less good scores and correlations. Table 2 also shows the results of the two toplines (original text+TTS and ASR+TTS). Interestingly, our best models come within 3% absolute in PER or CER compared to these toplines, are quite close to them in terms of MOS and even beat them in terms of human CER.

5.2 Results on the generation task

The upper mid and right cells of Figure 3 shows generation results averaging across the unconditional and conditional conditions, on automatic and human evaluations respectively. The main result is that there is both an effect of number of units and of system. As for resynthesis, 50 units is always worst, but contrary to resynthesis, 200 units is not always better. Overall, the results on generation are congruent with the idea that speech generation both requires good scores on language modeling and on speech synthesis. The best results for a particular model are then a compromise between the number of units that give both scores to either of these tasks. In terms of systems, the best one here is HuBERT. Regarding human evaluations, they show similar patterns with a clear dispreference for 50 units, and either 100 or 200 being better.

Detailed results are shown in Table 3 with separate statistics for conditional and unconditional generation and additional results with PPX@o-VERT and VERT@o-PPX. As expected, the perplexity metric improved with prompts, but not the diversity score. The human results are congruent with the automatic scores, although they tend to prefer more units, perhaps showing that they cannot fully dissociate their judgment of meaning from their judgment of intelligibility. The three metrics correlate well with one another (r between .86 and .99) and correlate with their counterpart across task (prompted vs. unprompted: r between .82 and .99). Human evaluations correlated well with the automatic metrics (AUC: r=.87; PPX: r=.92; VERT: r=0.75).

5.3 Results for zero-shot probe metrics

In Table 4, we show the results for zero-shot metrics across the different models and baselines. Overall, the performances depend on the linguistic levels while remaining above chance. While performances are excellent at the acoustic level (6.5% error for the best model on ABX-across), they are intermediate at the lexical level (31.3% error for the best model on spot-the-word). Not shown, the syntactic test is close to chance (42% error for the best model on the sBLIMP test). These values are worse than the ASR-topline (3.1% and 29%, for lexicon and syntax resp.), showing room for improvement.

The metrics correlate well: the ABX score predicts the lexical score ($r = 0.85$) and the syn-

Table 2: **Results on the resynthesis task** for 3 unsupervised models plus one LogMel baseline and 3 unit sizes. Bitrates are in bit/sec, PER are for a pretrained phone recognition model without lexicon and LM, CER are derived from a full ASR model (lower is better). Human MOS (upper is better) and CER (computed from transcription, lower is better) are provided (the 95% confidence interval was on average .32 for MOS and 1.8 for human CER)

Systems			End-to-end ASR-based metrics				Human Opinion			
S2u architect.	Nb units	Bit-rate	PER↓ (LJ)	PER↓ (LS)	CER↓ (LJ)	CER↓ (LS)	MOS↑ (LJ)	MOS↑ (LS)	CER↓ (LJ)	CER↓ (LS)
Toplines										
original wav			-	-	-	-	4.83	4.30	8.88	6.73
orig text+TTS			7.78	7.92	8.87	5.14	4.02	4.03	13.25	10.73
ASR + TTS	27		9.45	8.18	9.48	5.30	4.04	4.06	15.98	11.56
Baselines										
LogMel	50	214.8	27.72	49.38	27.73	52.05	2.41	2.07	43.78	66.75
LogMel	100	292.7	25.83	45.58	24.88	48.71	2.65	2.01	37.39	62.72
LogMel	200	373.8	19.78	45.16	17.86	46.12	2.96	2.16	23.33	62.6
Unsupervised										
CPC	50	159.4	10.87	17.16	10.68	12.06	3.63	3.51	13.97	19.92
CPC	100	213.1	10.75	15.82	9.84	9.46	3.42	3.68	13.53	14.73
CPC	200	279.4	**8.74**	14.23	9.20	8.29	3.85	3.54	**9.36**	14.33
HuBERT-L6	50	125.7	11.45	16.68	11.02	11.85	3.69	3.49	14.54	13.14
HuBERT-L6	100	168.1	9.53	13.24	9.31	7.19	3.84	3.68	13.02	11.43
HuBERT-L6	200	211.3	8.87	**11.06**	**8.88**	**5.35**	**4.00**	**3.85**	11.67	**10.84**
wav2vec-L14	50	141.3	24.95	33.69	25.42	32.91	2.45	2.87	46.82	54.9
wav2vec-L14	100	182.1	14.58	22.07	13.72	17.22	3.50	3.32	23.76	28.1
wav2vec-L14	200	226.8	10.65	16.34	10.21	10.50	3.83	3.51	13.14	15.27

tax score ($r = 0.71$). Across the different models, CPC gets the best units (ABX score) and HuBERT gets the best LM scores. In addition, we see a clear effect of number of units (Figure 3). For wav2vec, the performances on all metrics increase with more units, whereas, for CPC and HuBERT a U-shaped pattern emerges on most metrics, with best scores for units of intermediate sizes. It is interesting that the models with the highest bitrate do not always have the best results. This means that encoding too much acoustic information can be detrimental to linguistic encoding in the uLM. See Appendix Section 7.1 showing that ABX has good correlations with automatic and human metrics ($r > .88$).

6 Discussion and Conclusion

We introduced *Generative Spoken Language Modeling* as a new unsupervised task bridging the gap between speech and natural language processing and related it conceptually to previously studied unsupervised tasks: Acoustic Unit Discovery, Spoken Language Modeling, Discrete Speech Resynthesis and Text Generation. We introduced a suite of metrics, baselines, and first results on Librilight that sets the playing field for future work. For comparability, we open source our evaluation stack and the best of our baseline models.

Our main contributions are as follows: (1) we established a set of easy to use automatic ASR-based metrics for model comparison at two critical levels for this task: intelligibility of the speech output and meaningfulness in terms of higher linguistic content. We assessed the first through ASR-based PER and CER metrics; and the second using text-generation-based metrics (AUC for PPX/VERT). (2) We found that these two sets of metrics correlated well with human judgement and (3) that they can be approximated with their inference-mode counterparts, which are faster to computed using zero-shot probe tasks. (4) Applying these metrics to pipeline models based on current speech representation learning models and out-of-the-box LM and TTS components, we found that our basic premise is fulfilled: it is possible to train a language model from quantized units derived from audio and using it to generate new speech. The generated speech is English sounding, with recognizable phonemes and words and locally acceptable syntax (see transcribed ex-

Table 3: **Results on the generation task** for three unsupervised models plus the LogMel baseline and 3 unit sizes. PPX@-o-VERT and VERT@o-PPX are reported as PPX and VERT. '-' : missing or non calculable results. Human MMOS are also provided (the 95% confidence interval was on average .29 for uncond. and .61 for cond.).

Systems		Generation based metrics						Human Opinion	
Encoder	Nb	unconditional			prompt			uncond.	prompt
architect.	units	PPX↓	VERT↓	AUC↓	PPX↓	VERT↓	AUC↓	MMOS↑	MMOS↑
Controls									
oracle text		154.5	19.43	-	154.5	19.43	-	4.02	4.26
ASR + LM		178.4	21.31	0.18	162.8	20.49	0.04	3.91	4.38
Baseline									
LogMel	50	1588.97	-	1083.76	-	-	-	-	-
LogMel	100	1500.11	95.50	510.26	-	-	-	-	-
LogMel	200	1539.00	-	584.16	-	-	-	-	-
Unsupervised									
CPC	50	374.26	46.26	19.68	323.9	39.92	18.44	3.31	3.61
CPC	100	349.56	41.797	15.74	294.7	42.93	14.06	3.65	3.65
CPC	200	362.84	40.28	16.46	303.5	43.42	26.67	3.58	3.67
HuBERT-L6	50	376.33	43.06	19.27	339.8	45.85	21.03	3.53	3.00
HuBERT-L6	100	**273.86**	**31.36**	**5.54**	**251.2**	**33.67**	**5.88**	3.95	3.53
HuBERT-L6	200	289.36	33.04	7.49	262.4	34.30	6.13	**4.01**	**4.32**
wav2vec-L14	50	936.97	-	307.91	1106.3	-	330.8	2.26	1.91
wav2vec-L14	100	948.96	79.51	208.38	775.1	-	205.7	2.28	1.92
wav2vec-L14	200	538.56	61.06	61.48	585.8	-	91.07	2.64	3.04

Table 4: **Results for zero-shot probe metrics** for 3 unsupervised models plus one LogMel baseline and 3 unit sizes. ABX within and across speakers, spot-the-word and acceptability judgments are error rates (lower is better); chance is 50%.

Metrics		S2u		uLM	
	Nb	ABX	ABX	spot-the-	accept.
System	units	with.↓	acr.↓	word↓	judg.↓
Toplines					
ASR+LM		-	-	3.12	29.02
Baselines					
LogMel	50	23.95	35.86	48.52	46.78
LogMel	100	24.33	37.86	48.12	46.83
LogMel	200	25.71	39.65	49.62	47.76
Unsupervised					
CPC	50	5.50	7.20	32.18	45.43
CPC	100	**5.09**	**6.55**	31.72	44.35
CPC	200	5.18	6.83	37.40	45.19
HuBERT-L6	50	7.37	8.61	32.88	44.06
HuBERT-L6	100	6.00	7.41	**31.30**	**42.94**
HuBERT-L6	200	5.99	7.31	36.52	47.03
wav2vec-L14	50	22.30	24.56	51.92	45.75
wav2vec-L14	100	18.16	20.44	50.24	45.97
wav2vec-L14	200	16.59	18.69	44.68	45.70

amples in the Appendix and audio snippets here: https://speechbot.github.io/gslm). Our automatic metrics confirm the quality of the representations and outputs at the acoustic/phonetic level, but show that improvements are needed at the language level. It is to be expected that performance will increase with larger training sets beyond our 6k hours, as has been noted in the case of text. (5) We also uncovered specific issues regarding the number of quantized units. For speech resynthesis, the optimum number of units was always 200 by a large margin, reflecting the well known bitrate/intelligibility trade-off (Dunbar et al., 2019). However, for language modeling, this was not necessarily the case, as the more detailed acoustic information may introduce too numerous phonetic details that have no impact at the level of lexical and syntactic representations. (6) Finally, we found that the choice of units also affected the temperature parameter which is used to control the trade-off between quality and diversity in text-based language model. To address this effect, we proposed a method to normalize the temperature by using an oracle text to build perplexity and diversity anchor points.

Obviously, this is only a first step towards building textless NLP applications that could be applied to any language, even low resource ones. To reach this long term goal, three important challenges need to be addressed.

First, even though we did compare three different encoders and obtained different results we cannot conclude that one encoder is definitely su-

Table A1: Unconditional uLM (trained on CPC-100 units) samples, transcribed by ASR, at different temperatures.

Temp	Example
Very low temperature samples (stuttering zone)	
0.3	the property by james resell red for liberata or by jason downy the property by jason downy the property the property the property the property
0.3	and to take in another path and to take in another path and to take in another path and to take in another path and to take in another path and to take in another path and to take in another path and take in a
Low temperature samples (obsessive zone)	
0.7	chapter nineteen of the life of the upper part of the ocean this is ali bravos recording only bravos recordings are in the public domain i for more information or to volunteer
0.7	this is a lipper vox are courting oliver vox or courting are in the public domain for afraid art to volunteer pleases it lipper vox dot or this
Mid temperature samples	
1.0	but it is attendant from the people to defend himself from this information pride of the potential in criminal activity a curiosity and impetuosity of the world a war soon acquired
1.0	finally we ought to have a strong plan a without positively the best type of the public with which we ascend it or extend it our business and as we are a persons of the most strong designs and other affairs of the case we
High temperature samples (babble zone)	
1.5	ation of pure blue he said at once a licking streamy at her warm spot of half performed note was a raging oath let it as bir of amole in mood strolling er crass
1.5	at the swing here as to motions out of the events not time and abe he was any stump headed and flow any he's the kiln are tama why do ye take the floor

perior to the others. Our point here was merely to use previously published pretrained encoders, and study systematically the effect of number of units on these encoders. A fuller study including a wider set of encoders and a proper hyperparameter search (including the selection of the embedding layer and the clustering algorithm) would be needed in order to determine which of them is most appropriate for speech generation.

Second, it is to be expected that to further improve generation results, more needs to be done than applying this pipeline to larger training sets. Contrary to text, speech unfolds through time and varies continuously in phonetic space. Speech also contains multilayered representations (phonetic, prosodic, speaker identity, emotions, background noise, etc.). However, both our TTS and our LM were out-of-the-box systems typically used for text applications. More work is needed to adapt these architectures to the richness and variability of the speech signal (see Polyak et al., 2021a, for first steps towards integrating prosody into discrete units). The metrics and baselines we introduced here provide landmarks against which we will measure future progress.

Third, the automatic metrics that we defined here depend on textual resources to build the evaluation ASR and LM models, and on linguistic resources to build the zero-shot metrics. How could this ever be applied to low-resource languages? Note that the linguistic resources we require are used only for model selection, not model training. Our metrics allow for fast iterations in architecture and hyperparameter search, but the overall algorithm is totally unsupervised. Therefore, an important next step is to extend this work to other languages, in order to find a common architecture/hyperparameter set that gives good results in held out languages (high or low resource). The hope is that once good learning models are tuned using a diverse sample of high resource languages, the same models could be deployed in languages where no such resources are available, and work in a purely unsupervised fashion.

7 Appendix

7.1 Zero-shot metrics correlation results

In Figure A1, we present the Pearson correlations between the zero-shot metrics and the human and automatic metrics on downstream tasks. The fact that the ABX metric correlates well with these downstream metrics makes it a useful proxy metric for preliminary model and unit size selection, as it is much less costly than generating TTS output and running human or ASR evaluations.

7.2 Effect of temperature on outputs

In this section, we describe preliminary experiments we conducted to test the effects of temperature on the generated outputs. As shown in Table A1, the temperature defined qualitatively 4 operating zones. With the lowest temperature, we get

	Zero-shot			ASR-based				Human			
	ABX within	ABX across	spot-the word	avg PER	avg CER	AUC uncond	AUC prompted	avg CER	avg MOS	MMOS uncond	MMOS prompted
ABX within				0.904	0.896	0.893	0.806	0.901	0.883	0.935	**0.881**
ABX across	0.970			**0.944**	**0.938**	**0.962**	**0.910**	**0.905**	**0.924**	**0.941**	**0.881**
spot-the-word	0.937	0.853		0.767	0.760	0.753	0.639	0.806	0.743	0.902	0.808
Accept judgement	0.672	0.708	0.698	0.612	0.612	0.703	0.681	0.591	0.573	0.370	0.068

Figure A1: **Patterns of correlations between the zero-shot metrics and the automatic and human metrics.** Color scale indicates strength of the Pearson correlation coeficient (we used negative MOS and MMOS to enforce less is better for all metrics).

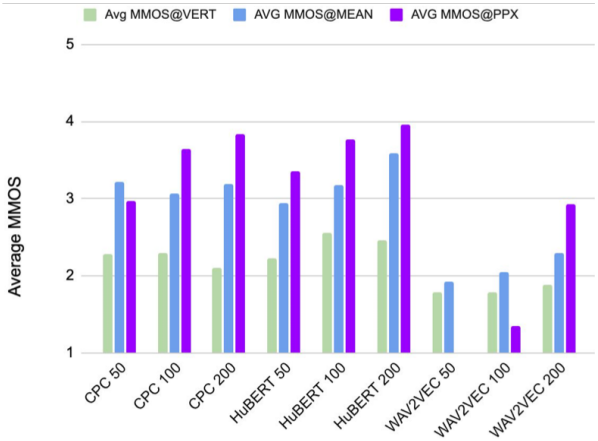

Figure A2: **MMOS for unconditional (no prompt) and conditional generated speech** sampled at the three reference temperatures (oracle VERT, oracle PPX, and average temperature) (preliminary experiments).

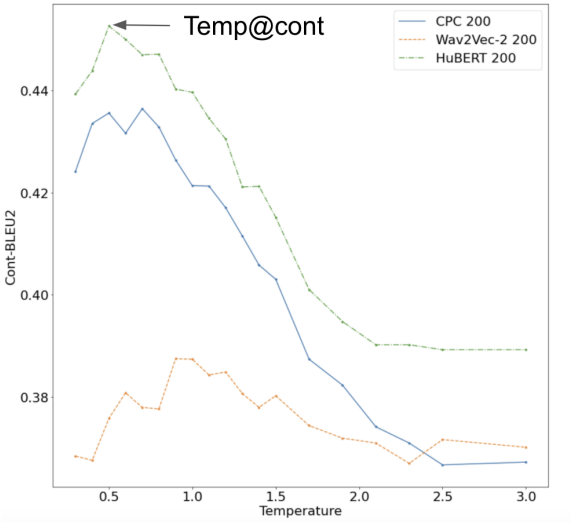

Figure A3: **Method for selecting the continuation temperature for MMOS judgements**.

repetitive outputs, where the system keeps repeating the same few words. At a slightly higher temperature, the system outputs complete sentences, but they are sampled from a narrow set of topics. At the highest temperature, the system utters an unstructured bag of words. In the mid-temperature range, we observe relatively coherent and varied outputs. This is the range we want to select for our systems. As described in Figure 2, the lowest bound was set by using the oracle PPX (temperature range between 0.2 and 0.65. across unsupervised models) and the highest bound by using the oracle VERT (temperature range between 1.1 to 1.4). In Figure S1 we present human opinion results for samples from these two temperatures, plus an extra mean temperature falling in between. Humans typically preferred the lower temperature.

In Figure A3, we illustrate the continuation method for selecting a single temperature for human meaningfulness judgments in a model-neutral way as explained in Section 3.3. It consists in generating possible continuations of each prompt and computing the BLEU-2 score[8] with oracle continuation. The temp@cont temperature is defined as the temperature maximizing this score. Computing these estimates with 10 continuations gave continuation temperatures varying between 0.5 and 0.9 across models and unit sizes. These are the temperatures we used for the MMOS results reported in the main paper.

Acknowledgments

We thank Michael Auli and Alexis Conneau for their useful input on wav2vec, and Lior Wolf, Pierre Emmanuel Mazaré and Gargi Gosh for their support for this project. We would also like to thank the reviewers and editors for their thorough review, and constructive feedback.

[8]We used NLTK to compute BLEU (Bird et al., 2009).

References

Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2019. Effectiveness of self-supervised pre-training for speech recognition. *CoRR*, abs/1911.03912.

Alexei Baevski, Wei-Ning Hsu, and Alexis Conneau. 2021. Unsupervised speech recognition. *arXiv preprint arXiv:2012.15454*.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020a. vq-wav2vec: Self-supervised learning of discrete speech representations. *International Conference on Learning Representations (ICLR)*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, pages 12449–12460.

Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural language processing with python.

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*.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149.

Mingjie Chen and Thomas Hain. 2020. Unsupervised acoustic unit representation learning for voice conversion using WaveNet auto-encoders. In *Proc. INTERSPEECH*, pages 4866–4870.

Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Shang-Wen Li, and Hung-yi Lee. 2021. Audio ALBERT: A lite BERT for self-supervised learning of audio representation. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350.

Jan Chorowski and Navdeep Jaitly. 2016. Towards better decoding and language model integration in sequence to sequence models. In *Proc. INTERSPEECH*, pages 523–527.

Yu-An Chung and James Glass. 2020. Improved speech representations with multi-target autoregressive predictive coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2353–2358. Association for Computational Linguistics.

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 Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32, pages 13063–13075. Curran Associates, Inc.

Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W. Black, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2019. The Zero Resource Speech Challenge 2019: TTS without T. In *Proc. INTERSPEECH*, pages 1088–1092.

Ewan Dunbar, Julien Karadayi, Mathieu Bernard, Xuan-Nga Cao, Robin Algayres, Lucas On-

del, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2020. The Zero Resource Speech Challenge 2020: Discovering discrete subword and word units. In *Proc. INTERSPEECH*, pages 4831–4835.

Emmanuel Dupoux. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

Janek Ebbers, Jahn Heymann, Lukas Drude, Thomas Glarner, Reinhold Haeb-Umbach, and Bhiksha Raj. 2017. Hidden Markov Model variational autoencoder for acoustic unit discovery. In *Proc. INTERSPEECH*, pages 488–492.

Ryan Eloff, André Nortje, Benjamin van Niekerk, Avashna Govender, Leanne Nortje, Arnu Pretorius, Elan van Biljon, Ewald van der Westhuizen, Lisa van Staden, and Herman Kamper. 2019. Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks. In *Proc. INTERSPEECH*, pages 1103–1107.

Siyuan Feng, Tan Lee, and Zhiyuan Peng. 2019. Combining adversarial training and disentangled speech representation for robust zero-resource subword modeling. In *Proc. INTERSPEECH*, pages 1093–1097.

Thomas Glarner, Patrick Hanebrink, Janek Ebbers, and Reinhold Haeb-Umbach. 2018. Full Bayesian Hidden Markov Model variational autoencoder for acoustic unit discovery. In *Proc. INTERSPEECH*, pages 2688–2692.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Wei-Ning Hsu, David Harwath, Christopher Song, and James Glass. 2020. Text-free image-to-speech synthesis using learned segmental units. *arXiv preprint arXiv:2012.15454*.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. 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.

Wei-Ning Hsu, Yu Zhang, and James Glass. 2017a. Learning latent representations for speech generation and transformation. In *Proc. INTERSPEECH*, pages 1273–1277.

Wei-Ning Hsu, Yu Zhang, and James Glass. 2017b. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems*, volume 30, pages 1878–1889. Curran Associates, Inc.

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

Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu. 2018. FFTNet: A real-time speaker-dependent neural vocoder. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2251–2255.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-light: A benchmark for ASR with limited or no supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673.

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.

Sameer Khurana, Shafiq Rayhan Joty, Ahmed Ali, and James Glass. 2019. A factorial deep Markov Model for unsupervised disentangled representation learning from speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6540–6544. IEEE.

Sameer Khurana, Antoine Laurent, Wei-Ning Hsu, Jan Chorowski, Adrian Lancucki, Ricard Marxer, and James Glass. 2020. A convolutional deep Markov Model for unsupervised speech representation learning. In *Proc. INTERSPEECH*, pages 3790–3794.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, pages 17022–17033.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. 2019. MelGAN: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems*, volume 32, pages 14910–14921. Curran Associates, Inc.

Cheng-I Lai, Yung-Sung Chuang, Hung-Yi Lee, Shang-Wen Li, and James Glass. 2021. Semi-supervised spoken language understanding via self-supervised speech and language model pretraining. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7468–7472.

Chia-ying Lee and James Glass. 2012. A nonparametric Bayesian approach to acoustic model discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40–49.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Shaoshi Ling, Yuzong Liu, Julian Salazar, and Katrin Kirchhoff. 2020. Deep contextualized acoustic representations for semi-supervised speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6429–6433. IEEE.

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.

Alexander H Liu, Yu-An Chung, and James Glass. 2020. Non-autoregressive predictive coding for learning speech representations from local dependencies. *arXiv preprint arXiv:2011.00406*.

Andy T. Liu, Po chun Hsu, and Hung-Yi Lee. 2019a. Unsupervised end-to-end learning of discrete linguistic units for voice conversion. In *Proc. INTERSPEECH*, pages 1108–1112.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Takashi Morita and Hiroki Koda. 2020. Exploring TTS without T using biologically/psychologically motivated neural network modules (ZeroSpeech 2020). In *Proc. INTERSPEECH*, pages 4856–4860.

Shekhar Nayak, C Shiva Kumar, G Ramesh, Saurabhchand Bhati, and K Sri Rama Murty. 2019. Virtual phone discovery for speech synthesis without text. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *Advances in Neural Information Processing Systems (NeurIPS) – Self-Supervised Learning for Speech and Audio Processing Workshop*.

Benjamin van Niekerk, Leanne Nortje, and Herman Kamper. 2020. Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 Challenge. In *Proc. INTERSPEECH*, pages 4836–4840.

Lucas Ondel, Lukáš Burget, and Jan Černockỳ. 2016. Variational inference for acoustic unit discovery. *Procedia Computer Science*, 81:80–86.

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

Aaron van den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315.

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*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. Fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT*, pages 48–53.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Jongseok Park, Kyubyong & Kim. 2019. g2pe. `https://github.com/Kyubyong/g2p`.

Santiago Pascual, Mirco Ravanelli, Joan Serra, Antonio Bonafonte, and Yoshua Bengio. 2019. Learning problem-agnostic speech representations from multiple self-supervised tasks. In *Proc. INTERSPEECH*, pages 161–165.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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

Adam Polyak, Lior Wolf, Yossi Adi, Ori Kabeli, and Yaniv Taigman. 2021b. High fidelity speech regeneration with application to speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7143–7147.

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

Adam Polyak, Lior Wolf, and Yaniv Taigman. 2020b. TTS skins: Speaker conversion via ASR. In *Proc. INTERSPEECH*, pages 786–790.

R. Prenger, R. Valle, and B. Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621.

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.

M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio. 2020. Multi-task self-supervised learning for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993.

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.

Thomas Schatz, Naomi H Feldman, Sharon Goldwater, Xuan-Nga Cao, and Emmanuel Dupoux. 2021. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7).

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. INTERSPEECH*, pages 3465–3469.

J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu. 2018. Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Transformer VQ-VAE for unsupervised unit discovery and speech synthesis: ZeroSpeech 2020 Challenge. In *Proc. INTERSPEECH*, pages 4851–4855.

Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura. 2019. VQVAE unsupervised unit discovery and multi-scale Code2Spec inverter for Zerospeech Challenge 2019. In *Proc. INTERSPEECH*, pages 1118–1122.

Patrick Lumban Tobing, Tomoki Hayashi, Yi-Chiao Wu, Kazuhiro Kobayashi, and Tomoki Toda. 2020. Cyclic spectral modeling for unsupervised unit discovery into voice conversion with excitation and waveform modeling. In *Proc. INTERSPEECH*, pages 4861–4865.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Maarten Versteegh, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2016. The Zero Resource Speech Challenge 2015: Proposed approaches and results. *Procedia Computer Science*, 81:67–72.

W. Wang, Q. Tang, and K. Livescu. 2020. Unsupervised pre-training of bidirectional speech encoders via masked reconstruction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6889–6893.

Anne Wu, Changhan Wang, Juan Pino, and Jiatao Gu. 2020. Self-supervised representations improve end-to-end speech translation. In *Proc. INTERSPEECH*, pages 1491–1495.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *SIGIR*, page 1097–1100.

Supplementary Materials

Here, we provide supplementary information not appearing in the TACL version for lack of space.

S1 Implementation Details

This section provides information about model training.

S1.1 Speech Decoder

We train all speech decoder models referred in Section 4.3 on 8 32-GB GPUs using data distributed training with a batch size of 32 per GPU for 500 epochs. We compute validation loss every 5000 steps, and select model with the lowest loss. For chunking (described in Section 4.3), we initialize the chunk size to 50 and increment it by 5 per epoch. Note that we only do chunking of Unit-To-Speech scenario only.

S1.2 Training of ASR Models for evaluation

S1.2.1 Frozen ASR Model

The frozen ASR model is trained using LARGE wav2vec model architecture with Connectionist Temporal Classification (CTC) loss (Graves et al., 2006) from scratch (not using the pretrained model) on LibriSpeech 960hours dataset.

S1.2.2 Frozen Phoneme Recognition Model

The frozen phone recognition model is trained using BASE wav2vec model architecture with Connectionist Temporal Classification (CTC) loss (Graves et al., 2006) from scratch on LibriSpeech 960hours dataset. We use g2p-en (Park, 2019) for obtaining gold phoneme transcriptions.

S1.2.3 Fitted ASR Model

For speed of training, we design a SMALL version of wav2vec architecture with a Transformer encoder of 6 layers, 4 attention heads, embedding of size 256 and FFN of size 1024. The model is always trained with CTC loss using synthesized speech from the speech decoder.

S2 Supplementary Results

In Section S2.1, we present the full set of metrics used the resynthetized task, and in section S2.2, those for the generation task.

S2.1 Speech Resynthesis Results on fitted ASR metrics

In the table S1 we present the full set of metrics that we have used to evaluate synthesized speech using fitted ASR models. All the speech synthesis models are trained on LJSpeech and the synthesized speech is evaluated on LibriSpeech and LJ Speech. Note that we always use dev_clean set for LibriSpeech, and a random hold-out set of 1000 samples for LJ Speech that were not seen during training or validation.

S2.2 Generation Task

Table S2 shows the results of the meaninfulness opinion score (MMOS) detailing the three temperature settings (oracle PPX, oracle VERT and the average temperature between these two endpoints). As seen on Figure S1, human judgements are usually higher for the oracle PPX than the other two temperatures, and the pattern of results across systems is globally comparable across temperatures. Figure S2 shows the table of correlation coefficients across human and automatic metrics.

Table S1: Full results using fitted ASR and Phoneme Recognition models of three unsupervised unit discovery models as a function of number of quantized units on 4 sets of metrics: PER and CER without LM & Lexicon (first 2 columns), and WER and CER with LM & Lexicon (last 2 columns). For comparison we also show the metrics for our supervised topline system.

Systems	PER		CER		WER		CER (with LM, Lex)	
	LJ	LS	LJ	LS	LJ	LS	LJ	LS
Gold Text + TTS	8.47	11.95	5.32	10.75	20.56	17.60	9.44	8.62
Pretrained ASR + TTS	-	-	8.80	11.88	23.86	21.10	11.28	10.34
LogMelFbank + KM50	34.32	60.76	25.29	48.54	69.36	92.52	39.62	63.56
LogMelFbank + KM100	24.94	50.58	24.01	47.40	65.27	92.13	36.43	62.64
LogMelFbank + KM200	22.55	53.21	20.99	46.37	57.61	90.98	30.72	59.01
CPC + KM50	16.04	26.72	13.27	27.61	39.16	63.10	18.52	33.26
CPC + KM100	19.23	32.05	12.48	25.93	35.77	58.09	16.63	29.57
CPC + KM200	15.92	26.78	**9.98**	23.22	**31.23**	53.56	**14.30**	26.73
HuBERT L6 + KM50	15.37	24.58	12.72	25.66	39.15	58.31	18.30	28.91
HuBERT L6 + KM100	16.57	25.83	11.74	24.56	31.27	51.67	14.60	26.36
HuBERT L6 + KM200	15.88	24.11	11.20	**21.62**	32.91	**50.53**	16.00	**26.20**
wav2vec L14 + KM50	23.89	36.18	27.60	39.55	71.19	86.27	44.52	57.51
wav2vec L14 + KM100	16.62	27.76	17.11	30.67	48.84	71.53	24.51	38.28
wav2vec L14 + KM200	**13.79**	**24.06**	11.25	23.54	34.55	55.53	15.96	27.48

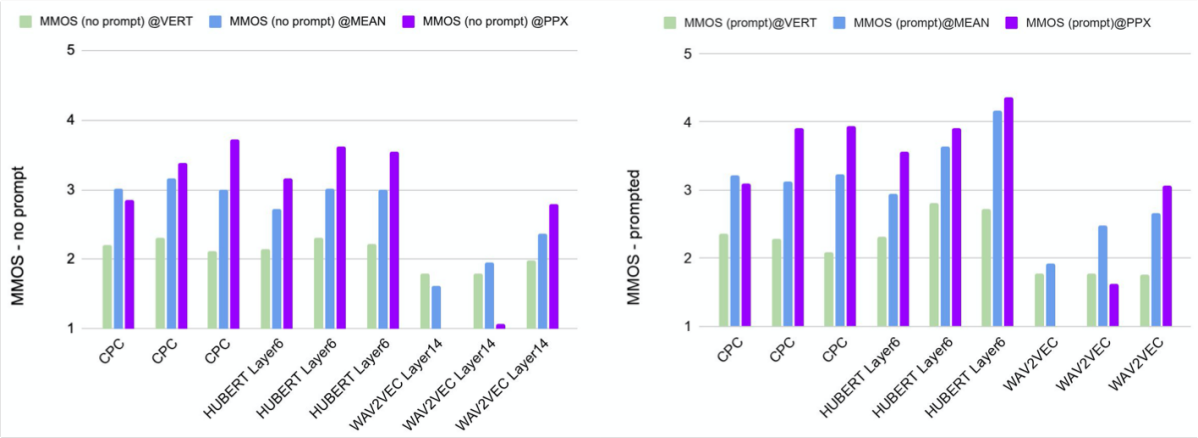

Figure S1: **MMOS for unconditional (no prompt) and conditional generated speech** sampled at the three reference temperatures (oracle VERT, oracle PPX, and average temperature).

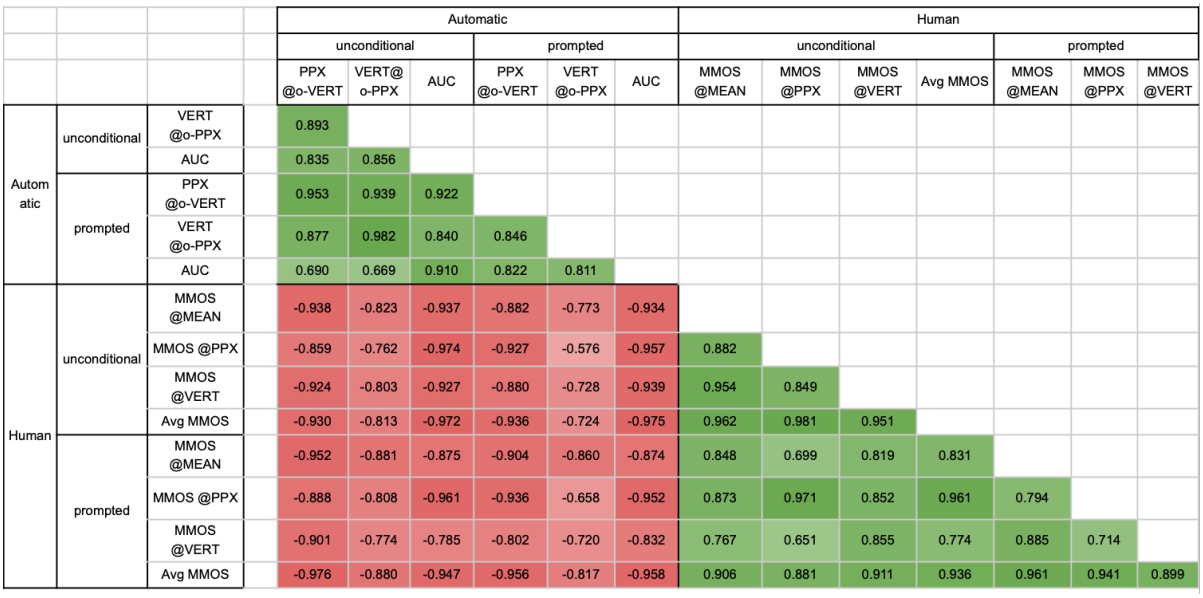

Figure S2: **Patterns of correlations between automatic and human metrics for generation.** Color scale indicates strength and direction of the Spearman correlation coeficient.

Table S2: Full Human evaluation MMOS results for three unsupervised models and 3 unit sizes. For each model and unit size, samples were generated using three different temperatures corresponding to: VERT, PPX and the average temperature of both.

Systems		MMOS					
Encoder architect.	Nb units	unconditional PPX	unconditional VERT	unconditional MEAN	prompt PPX	prompt VERT	prompt MEAN
Controls							
oracle text		-	-	-	-	-	4.44
ASR + LM		3.72	3.23	3.57	3.24	3.02	3.24
Unsupervised							
CPC	50	2.85	2.20	3.02	3.10	2.35	3.22
CPC	100	3.38	2.31	3.16	3.90	2.28	3.12
CPC	200	3.72	2.11	3.00	3.94	2.08	3.23
HuBERT-L6	50	3.16	2.15	2.72	3.56	2.31	2.94
HuBERT-L6	100	3.62	2.31	3.02	3.90	2.81	3.64
HuBERT-L6	200	3.55	2.22	3.00	4.36	2.71	4.16
Wav2vec-L14	50	-	1.79	1.62	-	1.77	1.92
Wav2vec-L14	100	1.07	1.79	1.95	1.62	1.77	2.48
Wav2vec-L14	200	2.80	1.99	2.37	3.06	1.76	2.65