SONAR EXPRESSIVE: Zero-shot Expressive Speech-to-Speech Translation

Paul-Ambroise Duquenne[*]
Meta AI & Inria
padqn@meta.com

Kevin Heffernan[*]
Meta AI
kevinheffernan@meta.com

Alexandre Mourachko
Meta AI
alexmourachko@meta.com

Benoît Sagot
Inria
benoit.sagot@inria.fr

Holger Schwenk
Meta AI
schwenk@meta.com

Abstract

Massively multilingual and multimodal sentence representations like SONAR are usually trained to capture only the meaning of the encoded text or speech. We complement this semantic embedding by a generic speech characteristic embedding which captures the expressive properties of a speech signal. We describe an iterative training procedure which aims to disentangle the semantics and expressive speech properties, and which does not need labeled data. We show the effectiveness of our method on the FLEURS and MEXPRESSO benchmark test sets using multiple metrics which aim to measure the preservation of the meaning and prosody for zero-shot speech-to-speech translation from five languages into English.

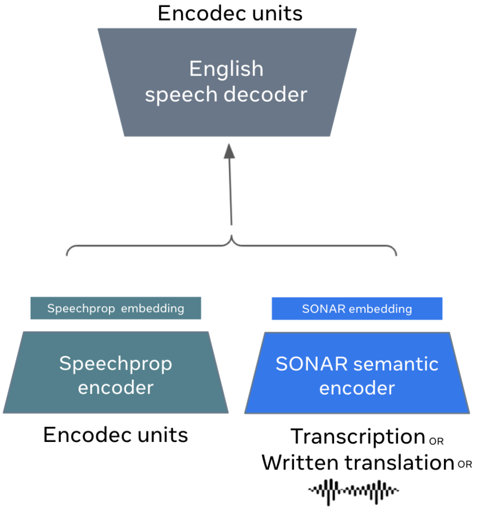

Figure 1: Model architecture for SONAR EXPRESSIVE

1 Introduction

Speech-to-speech translation has made significant progress the last years, and systems trained end-to-end have evolved, e.g. (Jia et al., 2019b, 2022; Lee et al., 2022, 2021; Seamless Communication et al., 2023a), which outperform traditional cascaded approaches like (Nakamura et al., 2006; Lavie et al., 1997). An alternative framework is the T-Modules architecture (Duquenne et al., 2022b). The underlying idea is to connect (independently) trained text/speech encoders and decoders with a fixed-sized multilingual and -modal sentence embedding space. This approach was initially based on the LASER sentence embedding and then extended to a new space named SONAR (Duquenne et al., 2023a). T-Modules has delivered competitive performance in zero-shot speech-to-text (S2TT) and zero-shot speech-to-speech translation (S2ST). However, all those approaches focus on the semantics of S2ST only, i.e. preserving the meaning of the spoken sentence. Still, human oral communication conveys additional information like the speech rate, pitch, prosody, emotion, etc. All these speech characteristics are very important to correctly understand the message and intention in oral communication.

In this work, we extend the T-Modules architecture and introduce an additional embedding to capture generic speech characteristics (see Figure 1). Concretely, we disentangle the content of the message, using the SONAR semantic embedding, from the expressive speech characteristics with a new embedding vector. We describe in detail an iterative procedure how to learn this vector and how to condition a speech decoder on both the semantic and expressive speech properties embeddings. This enables us to decode speech with similar audio style characteristics than the source speech. We present results for zero-shot S2ST from German, French, Italian, Spanish or Chinese into English.

In contrast to some existing work, e.g. (Macary et al., 2021; Duret et al., 2023), we do not require labeled data to specify the type and categories of

[*] Equal contribution

speech characteristics we want to model and maintain, e.g. different emotions like happy or sad, but our generic speech characteristics embedding is automatically derived from raw audio. Compared to other previous works which use unlabeled data to build prosody representations (Qu et al., 2023; Wang et al., 2018) with low dimensional embedding spaces or discrete representations for style conversion in text-to-speech systems, we build a high dimensional embedding space for expressive speech characteristics disentangled from multilingual semantic representations in the frame of speech-to-speech translation. Our speech decoder is also trained on more than 1 million hours of speech when accounting for the raw speech used for the first stage of training.

This paper is structured as follows. In the next section we first summarize the main ideas of SONAR embeddings, the T-Modules architecture, and relate our work to published research. In section 3, we describe the methodology we developed to complement SONAR sentence-level speech embeddings with speech properties embeddings. This approach aims to represent information, such as prosody and expressivity, and to preserve it in the S2ST process. Section 4 describes the multi-stage protocol we followed to train our S2ST models. Section 5 details the multiple metrics used to evaluate our models, including a number of expressivity preservation metrics. We demonstrate that we can perform high-quality, expressivity-preserving zero-shot speech-to-speech translation. This is achieved in a setting where the use of SONAR representations allows for easy cross-modal zero-shot transfer.

2 Background and related work

2.1 Speech-to-speech translation

Much of the early work on speech-to-speech translation has focused on developing cascaded-based approaches consisting of a combination of ASR, text-to-text translation (T2TT), and TTS models (Nakamura et al., 2006; Lavie et al., 1997). For example, when using such a multi-stage cascaded system, input speech would first be transcribed using an ASR system, followed by text translation and then finally synthesised into the target speech. These approaches have the advantages of leveraging highly performant individual systems such as large-scale multilingual text translation models (NLLB Team et al., 2022; Fan et al., 2020). However, such cascaded systems can propagate errors. For example, an incorrect transcription by the ASR model will then be passed to each subsequent system and accumulate further errors. In contrast to this, more recent approaches have focused on direct speech-to-speech translation systems which are capable of directly producing target speech representations (Jia et al., 2019b, 2022; Lee et al., 2022). One such common speech representation are speech units. These are discrete representations for speech analogous to word or token representations in text. Examples of such include HuBERT (Hsu et al., 2021) and XLS-R (Babu et al., 2021). In order to generate the target speech from these unit representations, a speech synthesiser is needed (Polyak et al., 2021). However, more recently codec-based representations such as EnCodec (Défossez et al., 2022) allow for units to be converted into the raw wave form without the need for an externally trained vocoder.

2.2 Decoding multilingual and multimodal sentence embeddings

Multilingual speech/text sentence embeddings were recently introduced to encode speech and text sentences from different languages into a shared semantic embedding space (Duquenne et al., 2021; Khurana et al., 2022). These semantic representations were successfully used to perform speech-to-text and speech-to-speech mining (Duquenne et al., 2022a) which significantly helped improve the performance of speech translation models (Seamless Communication et al., 2023a). More recently, Duquenne et al. (2022b, 2023b) showed that these fixed-size sentence representations could be efficiently decoded into text or speech in different languages, training decoders with limited supervised data. In particular, Duquenne et al. (2022b) demonstrated that a speech decoder trained with raw audio leads to good zero-shot text-to-speech and speech-to-speech translation results: raw audio is embedded into the sentence embedding space with a pre-trained frozen encoder, and a decoder is trained to recover the HuBERT units of the input speech. Finally, Duquenne et al. (2023a) introduced SONAR, a state-of-the-art massively multilingual speech/text sentence embedding space with good decoding performances into text in 200 languages. Contrarily to (Duquenne et al., 2022b), no speech decoder was yet trained in this framework.

2.3 Controllable text-to-speech synthesis

Text-to-Speech synthesis is a common task that aims to synthesize speech following a text prompt. Special focus is made on generating natural-sounding speech, with recent works allowing precise control on the output vocal style. Some speech properties are naturally easier to annotate and thus control: specific vocal styles can be selected through labels in (Kim et al., 2021), or pre-trained embeddings (Jia et al., 2018; Casanova et al., 2022). On the contrary, prosody labels are harder to define. Prosody was captured in low-dimensional residual embeddings in (Wang et al., 2018; Akuzawa et al., 2018; Ren et al., 2020). However, this simplistic modeling was proven to struggle when conditioned on noisy references (Hsu et al., 2018).

More recently, novel TTS modeling approaches enabled high quality vocal style transfer. Treating TTS as a language modeling task on Encodec (Défossez et al., 2022), VALL-E (Wang et al., 2023a) is able to preserve the acoustic environment and speaker's emotion of its audio prompt. (Kharitonov et al., 2023) is also using prompting and neural audio codec to tackle TTS while preserving acoustic style of the prompt. Also based on a neural audio codec, and drawing from diffusion-style models, Naturalspeech2 (Shen et al., 2023) can perform singing synthesis. Speech synthesis across languages is even made possible in zero-shot fashion in Voicebox through flow-matching (Le et al., 2023).

2.4 Expressive speech generation and translation

Beyond controllable TTS, expressive speech generation can be achieved through pure speech language modeling as well, without conditioning on text. AudioLM (Borsos et al., 2023) generates tokens inspired from HuBERT (Hsu et al., 2021), followed by vocal style-preserving Soundstream (Zeghidour et al., 2021) tokens.

Vocal style transfer has also recently been a focus for speech-to-speech translation. Vocal style is preserved across languages in Translatotron (Jia et al., 2019b), whose synthesizer is conditioned on a speaker embedding output by speaker encoder trained on the side. Direct S2ST models with consistent vocal style can be trained on vocal style aligned speech generated with controllable TTS models, as shown with Translatotron 2 (Jia et al., 2022) and AudioPaLM (Rubenstein et al., 2023). Polyvoice (Dong et al., 2023) leverages two language models: one for translation and one for speech synthesis. The vocal style characteristics and the speaking style of the original speech is preserved during the synthesizing step. Similarly in (Wang et al., 2023b), speech translation is decomposed into the translation of linguistic contents in a first stage, followed the transfer of vocal style characteristics in later stages.

3 Methodology

In this work, we aim at training a speech decoder model in the SONAR (Duquenne et al., 2023a) framework. Following the modular training strategy presented by Duquenne et al. (2022b), we trained an English speech decoder on monolingual raw speech data as well as paired speech-text data, to decode SONAR embeddings computed with pre-trained encoders (either speech encoders or text encoder). At inference time, the English speech decoder can decode spoken languages unseen during training to perform zero-shot speech-to-speech translation.

In addition to semantic conditioning of the speech decoder with SONAR sentence embeddings, we introduce an additional supposedly disentangled fixed-size representation to capture the prosody and expressive content of speech that is not represented by SONAR semantic embeddings. This additional embedding is called SPEECHPROP embedding, as it is supposed to encode prosody and expressive properties of the speech modality. We define our combined system comprising the SPEECHPROP and expressivity-aware speech decoder as SONAR EXPRESSIVE. Contrarily to Duquenne et al. (2022b) which uses HuBERT discrete units as target for their unit decoder, we used EnCodec units in order to be able to generate diverse speech (Défossez et al., 2022). HuBERT units were built to be more or less agnostic to the speaker and are often referred as semantic speech tokens. On the other hand, EnCodec units are trained to build compressed representations of audio, carrying much more acoustic information. Moreover, EnCodec model comes with a decoder, which can generate speech waveforms from units whereas a separate HiFi-GAN vocoder has to be trained when using HuBERT units.

3.1 Architecture

In addition to the pre-trained SONAR semantic encoders for speech and text which are frozen during the speech decoder training, our model is composed of an auto-regressive EnCodec decoder, a non auto-regressive EnCodec decoder and the SPEECHPROP encoder. The auto-regressive and non-auto-regressive decoders follows the architecture introduced by (Wang et al., 2023a), with the exception that, for simplicity, units from different codebooks are gathered into a common vocabulary on which the softmax operation is applied. Following (Wang et al., 2023a), the auto-regressive decoder predicts the EnCodec units from the first codebook, while the non auto-regressive decoder takes as input the sum of embeddings of units from the first $n-1$ codebooks to predict EnCodec units of codebook n. During training, the value of n is uniformly sampled between 2 and 8 for each training step. The SPEECHPROP encoder is a Transformer encoder taking as input the sum of EnCodec units embeddings. Its ouputs are mean-pooled to form the SPEECHPROP embedding. We use 16 transformer layers for the decoders and 12 transformer layers for the SPEECHPROP encoder. Finally, the SPEECHPROP embedding and the SONAR semantic embeddings are concatenated so that decoders can perform cross-attention on these representations to predict target EnCodec units (see Figure 1).

Training of new EnCodec model The original EnCodec model was trained on both speech and music with a 75Hz frame rate. In this paper, we introduce a slightly modified EnCodec model which is trained only on multilingual speech and with a 25Hz frame rate compared to the original 75Hz frame rate which makes speech unit sequences tree times shorter, improving memory usage during training. We followed the original EnCodec model design: 128 dimensions for the representation space, 1024 codes in each of 8 codebooks, but a modified subsampling scheme in order to have 25Hz frame rate. To achieve this lower frame rate, the SEANet encoder/decoder has following ratio: [8,5,4,4] which effectively downsamples 16kHz into 25Hz (16000/(8*5*4*4)=25). We used some natural multilingual speech data to train this new EnCodec model without integrating other audio training data like music in the training contrarily to the original model training.

Multilingual SONAR speech encoder In this work, we focus on handling six source languages: English, German, French, Italian, Spanish and Mandarin Chinese. This choice is motivated by the availability of evaluation data to measure various prosodic features of speech (see Section 5.2). In principle, our approach is generic and could be applied to any language. We use the original SONAR English speech encoder[1] and train a new single speech encoder for the remaining five languages. We follow the recipe of Duquenne et al. (2023a) and train on public ASR data only. Table 1 provides an S2TT evaluation on the FLEURS test set when connecting this speech encoder to the SONAR text decoder. Despite being zero-shot for speech-to-text translation, our results compare favorably to a large system like Whisper v2 large which was trained on large amounts of labeled data.

Model	cmn	deu	fra	ita	spa
Ours	17.1	31.6	30.2	25.4	24.1
Whisper	18.4	34.6	32.2	23.6	23.3

Table 1: Evaluation of our multilingual speech encoder on S2TT FLEURS test set (sacreBLEU scores).

4 Training setup

4.1 Multi-stage training

Semantic vectors are computed from source instances: source inputs to SONAR encoders are different for the different stages of training detailed in the following paragraphs. On the other hand, SPEECHPROP embeddings are computed from target speech during unsupervised fine-tuning in order to extract the missing information to predict output speech from both SONAR embeddings and SPEECHPROP embeddings. SPEECHPROP embeddings are computed from source speech during inference. More details about training configurations are given in the following parts.

SONAR and SPEECHPROP embeddings are concatenated as inputs to the decoders and we used cross-entropy loss on EnCodec units as our objective function. This conditioning is replaced by zero vectors with a probability of 0.1 during training, in order to also train the decoders in an uncon-

[1] https://github.com/facebookresearch/SONAR

ditional setting, to be used to compute classifier-free (CF) guidance during inference.

Initial experiments showed that introducing the SPEECHPROP embedding from scratch leads to a state where the speech decoders only rely on the SPEECHPROP vector to predict output units (auto-encoding EnCodec units with SPEECHPROP encoding) and ignoring the SONAR embedding. To overcome this collapse, we introduce a multi-stage training strategy which can be divided into pre-training and fine-tuning stages.

Pre-training with raw speech. Only the decoders are trained during this stage, taking as input SONAR embeddings only, the SPEECHPROP embedding is replaced by a vector filled with zero values. Pre-training is itself composed of 2 sub-stages: a first pre-training phase using only raw monolingual speech data, following the training method introduced in Duquenne et al. (2022b). We first start with approximately 1 Million hours of raw English speech data originating from a publicly available repository of web data (Seamless Communication et al., 2023b). Raw speech data is then segmented using the SHAS neural segmenter (Tsiamas et al., 2022). These raw speech segments are embedded into the SONAR space with a pre-trained English SONAR speech encoder, frozen during this training. The speech decoders learn to recover the EnCodec units of input speech only based on the SONAR speech embeddings. This training stage enables learning an initial conditioning to SONAR embedding as well as internal language modeling of EnCodec units (this step can be seen as auto-encoding with a frozen encoder). We trained this first stage of pre-training for 300k gradient updates which corresponds to 1.5 epochs on our training data.

Pre-training with S2TT data. The second step of pre-training consists in using public repositories of ASR data totalling approximately 42k hours of English speech. Transcriptions from a 2k hours subset were translated into our 5 languages of focus. These multilingual transcripts are used to compute SONAR embeddings to condition the decoder that learns to predict the EnCodec units of the corresponding speech. Similarly to the previous pre-training stage, only the decoders are trained. Multilingual inputs are used in order to make the speech decoders more robust to other languages, rather than overfitting on English embeddings, as motivated in Duquenne et al. (2022b) for their text decoders. This second phase of training is also called pre-training, as the speech decoders are only attending to SONAR embeddings, and the SPEECHPROP embeddings are not yet introduced. It has the advantage to pre-train the speech decoders to rely on multilingual semantic SONAR embeddings to predict EnCodec units. We trained this second stage of pre-training for 100k gradient updates.

Fine-tuning Now that the speech decoders has learned to rely on SONAR embeddings to predict EnCodec units, we introduced the SPEECHPROP embeddings, in order to make the decoders not only rely on semantic information to predict target EnCodec units but also prosody and expressive speech properties that should be found in the output speech. Target EnCodec units are then fed to the SPEECHPROP encoder, and both the SPEECHPROP encoder and the decoders are fine-tuned. In order to efficiently fine-tune the speech decoders, and avoid over-fitting, we only fine-tuned the cross-attention weights of the decoders. Moreover, to encourage the decoders to continue relying on semantic embeddings during this fine-tuning stage, we introduce a regularization method that we called *random-cropping* of target speech. Instead to feeding the entire sequence of EnCodec units to the SPEECHPROP encoder, only random crops of the target EnCodec units are fed. The lengths and positions of the crops are randomly sampled, with minimum length of 10 EnCodec units and maximum length set to the length of the target sequence. This minimum length ensured stable training of the SPEECHPROP encoder. This fine-tuning stage is performed on the same publicly available data with automatic multilingual transcripts as semantic conditioning. We trained the model for 40k gradient updates during this fine-tuning stage.

5 Evaluation

5.1 Datasets

We evaluate our models on both the FLEURS (Conneau et al., 2023) and MEXPRESSO benchmark datasets (Seamless Communication et al., 2023b). FLEURS is a partially n-way parallel speech dataset in 102 languages built on top of the FLoRes-101 machine translation dataset (Goyal et al., 2022). MEXPRESSO is a multilingual expressive speech-to-speech translation dataset which contains speech for five target languages recorded in six different vocal styles: default (neutral), happy, sad, confused, enunciated, and whispering. There are four speakers for each language. As interpretation of each vocal style can vary from speaker to speaker (e.g. happy can be expressed with different levels of intensity, intonation, rhythm, pause, etc), English speech was first recorded independently. In order to gather alignments in other target languages, bi-lingual speakers (native in the target language) listened to the English-side of each utterance before recording, in order to ensure they expressed the same interpretation of vocal style. An overview of the benchmark datasets is shown in Table 2.

	FLEURS		mEXPRESSO	
	dev	test	dev	test
cmn → eng	1.27	3.07	3.51	6.40
deu → eng	1.26	3.15	4.85	7.21
fra → eng	0.80	1.95	5.31	6.82
ita → eng	1.55	3.52	5.86	6.64
spa → eng	1.35	3.09	5.20	6.94

Table 2: Num. of source hours per benchmark dataset.

5.2 Metrics

In order to ensure our expressive translation system is able to maintain content translation quality, we first evaluate using ASR-BLEU. In order to measure this, we transcribe using the publicly available Whisper model,[2] and then compare the transcriptions to the ground truth using sacreBLEU.[3]

As there are multiple dimensions of prosody which can be captured by our SPEECHPROP vector, it is not straight-forward to find one prosody-based metric which is able to adequately cover each dimension of vocal style. We therefore choose to evaluate the prosodic qualities of our translation system using a suite of expressivity metrics, each of which is described below.

Speaker style similarity. Speaker style embeddings of both source and target speech are extracted using a pre-trained WavLM-based speaker style encoder (Chen et al., 2022). We then measure speaker style similarity as the cosine between source and target (Le et al., 2023).

AUTOPCP. In order to estimate the quality of sentence-level prosodic similarity, we use AUTOPCP (Seamless Communication et al., 2023b). This is a neural model trained to predict Prosodic Consistency Protocol (PCP) scores (Huang et al., 2023), which are measured on a likert scale between 1 and 4 (where 4 is the highest possible score), and have been found to correlate with human judgments of prosodic similarity.

Speech rate and pause alignment. As rhythmic patterns in the utterance are also an important aspect of expressivity, we aim to capture such characteristics by comparing both the rate of speech and the pause alignment. The speech rate is calculated by measuring the number of syllables spoken per second. We then report the Spearman correlation of the number of syllables spoken between the source and generated audios.[4] Complementary to the speech rate, another aspect of rhythm are the lengths of silence left between words. We therefore also report a pause alignment score measuring how well silences are preserved between the source and translation. Silences were captured using Silero VAD (Silero, 2021). For both speech rate and pause alignment metrics, we used the Rythmic Toolkit implementation (Seamless Communication et al., 2023b).

5.3 Inference

We used top-k sampling to generate EnCodec units during inference. EnCodec units from the first codebook are generated in an auto-regressive manner with the auto-regressive decoder, while EnCodec units from other codebooks are iteratively predicted by the non auto-regressive decoder, as presented in (Wang et al., 2023a).

Moreover, we used classifier-free (CF) guidance on logits as done in (Gafni et al., 2022; Kreuk et al., 2022), thanks to both conditional and unconditional training of the decoders. We used $k = 10$ for top-k sampling and a classifier-free guidance

Setup	cmn	deu	fra	ita	spa
Top-k sampling	5.26	17.64	16.99	12.53	14.98
+ CF guidance	9.25	24.11	22.70	17.15	18.53

Table 3: ASR-BLEU performance with and without classifier-free guidance on FLEURS.

[2] large-v2 model.

[3] 13a tokenizer.

[4] For Mandarin, characters are treated as syllables.

	FLEURS									mEXPRESSO								
	pre-train$_1$			pre-train$_2$			fine-tune			pre-train$_1$			pre-train$_2$			fine-tune		
SPEECHPROP	✗			✗			✓			✗			✗			✓		
Semantic input	eng text	xxx text	xxx speech	eng text	xxx text	xxx speech	eng text	xxx text	xxx speech	eng text	xxx text	xxx speech	eng text	xxx text	xxx speech	eng text	xxx text	xxx speech
cmn	51.63	3.50	4.58	65.65	13.82	9.25	57.24	11.49	7.86	82.53	7.07	8.84	80.89	18.33	14.81	69.84	12.77	10.40
deu	52.61	15.60	14.89	67.25	27.74	24.11	59.50	24.90	22.16	82.53	19.71	14.82	80.79	32.13	24.24	71.31	25.61	19.43
fra	51.84	18.00	13.44	65.59	29.14	22.70	54.15	23.86	19.82	81.81	22.82	14.92	80.70	35.01	23.28	67.99	26.97	18.64
ita	51.37	11.32	11.33	66.78	20.34	17.15	62.38	18.90	16.53	81.28	25.65	15.21	80.91	38.72	25.03	68.74	30.71	20.21
spa	53.38	12.44	11.55	66.37	20.79	18.53	61.62	19.52	17.49	81.51	33.28	23.67	80.77	44.66	33.31	68.92	36.52	27.23

Table 4: ASR-BLEU performance.

scale of 3. We report the difference in ASR-BLEU for the model trained to predict EnCodec units from semantic vectors only (pre-train stage 2) with and without classifier-free guidance and show the importance of such method when predicting EnCodec units for direct speech-to-speech translation.

5.4 Results

In order to first measure the content translation quality of SONAR EXPRESSIVE, we calculated ASR-BLEU results for each target language across both the FLEURS and mEXPRESSO benchmark datasets during each stage of model training. We condition the speech decoder with various semantic embeddings in order to analyze the cross-lingual and cross-modal transfer, given the combination of different semantic SONAR encoders with our speech decoder. We namely use three such embeddings: one extracted from target English text, one extracted from source non-English transcription, and one from non-English source speech. These three different setups are respectively performing TTS, T2ST and zero-shot S2ST. Finally, in order to determine the effect of the SPEECHPROP vector, we also generate audio without this embedding. Results are shown in Table 4.

First, we notice that SONAR EXPRESSIVE is performing TTS very capably in terms of ASR-BLEU. TTS results are already surprisingly good with the pre-train$_1$ model, reaching for instance more than 80 BLEU on the French → English split of mEXPRESSO. This again highlights the zero-shot cross-modal transfer happening in the SONAR framework as the pre-train$_1$ speech decoder was only trained to decode speech embeddings. We see that TTS results are better after the second stage of pre-training on FLEURS, which can be explained by the length distribution of FLEURS compared to the training data of the pre-train$_2$ speech decoder which includes longer audios from ASR training set compared to the training data of the pre-train$_1$ speech decoder which contains speech instances are $\sim$3 seconds in average. This is to compare with TTS after the second stage of pre-training on mEXPRESSO, which does not improve compared to the first stage of pre-training. Indeed, the average duration on mEXPRESSO is 3.51 seconds (target-side) whereas the average duration on FLEURS is 9.78 seconds (target-side). After the second stage of pre-training, we get important performance boost on TTS ASR-BLEU on FLEURS, with more 10 ASR-BLEU gains.

When starting to introduce SPEECHPROP embeddings during finetuning, we notice some loss in ASR-BLEU. This could be explained by the fact that during training, the model starts to rely on the SPEECHPROP embeddings of the cropped target to predict the whole target. But it could also come from the ASR-BLEU metric itself that relies on an automatic transcription. The speech recognition system may perform worse on more expressive speech compared to more normalized English generated speech output by the pre-training-only based model.

TTS task should be seen as a topline for T2ST and S2ST translation results. It highlights the ability of the speech decoder to output diverse speech sentences while conditioned on fixed-size sentence embeddings.

When switching from TTS to T2ST, we notice a clear loss for the pretrain$_1$ model, showing that it had over-fitted on English embeddings, while the goal is to have a speech decoder robust to embeddings from other languages. However, we notice that the second stage of pre-training helps improve the robustness of the speech decoder to other languages, which validates the incorporation of S2T data in the training. For example on French, we see a +11 ASR-BLEU gain when comparing pre-train$_1$ and pre-train$_2$ models.

	FLEURS			mExpresso		
	pre-train$_1$	pre-train$_2$	fine-tune	pre-train$_1$	pre-train$_2$	fine-tune
SpeechProp	✗	✗	✓	✗	✗	✓
cmn	0.05	0.06	0.30	0.02	0.04	0.30
deu	0.02	0.04	0.39	0.02	0.03	0.25
fra	0.0	0.03	0.29	0.0	0.03	0.21
ita	0.02	0.05	0.27	0.0	0.02	0.22
spa	0.02	0.04	0.28	-0.01	0.02	0.23

Table 5: Speaker style similarity performance in zero-shot S2ST.

	FLEURS			mExpresso		
	pre-train$_1$	pre-train$_2$	fine-tune	pre-train$_1$	pre-train$_2$	fine-tune
SpeechProp	✗	✗	✓	✗	✗	✓
cmn	0.06	0.08	0.24	0.13	0.12	0.54
deu	0.19	0.20	0.64	0.10	0.08	0.62
fra	0.04	0.15	0.36	0.08	0.13	0.43
ita	0.14	0.23	0.31	0.09	0.11	0.49
spa	0.17	0.30	0.42	0.10	0.13	0.56

Table 6: Speech rate Spearman correlation in zero-shot S2ST.

Finally, we introduce zero-shot speech-to-speech translation results. We observe reasonable ASR-BLEU results after the first stage of pre-training only. It is important to highlight that this model was trained only to decode English SONAR speech embeddings into English EnCodec units (which can be seen as auto-encoding with a frozen semantic encoder). Therefore, the speech-to-speech translation results shown for this first stage of pre-training are zero-shot cross-lingual for non-English spoken languages. Second, we notice that adding multilingual text inputs in the training during pretraining stage 2 significantly improves ASR-BLEU results. It confirms that multilingual inputs, even though coming from another modality, help to make the speech decoder more robust to multilingual inputs from the speech modality. The disparity in results between mandarin and the other target languages during the first stage of pre-training may be due to fact that the representations from the SONAR speech encoder for mandarin are less strong compared to other languages (17.1 S2T BLEU for cmn compared to 31.6 S2T BLEU for deu).

The differences in ASR-BLEU between TTS, T2ST and S2ST suggest that incorporating more S2TT or even S2ST data in the training could boost ASR-BLEU performances. This is left to future work.

In order to determine the dimensions of expressivity captured by the SpeechProp embedding, we begin by examining its effect on speaker similarity. Results are shown in Table 5. As we expected, models which were not trained with SpeechProp embeddings generate output speech with a very low speaker style similarity given an input speech. Introducing the SpeechProp embeddings into the training during the fine-tuning stages significantly boosts speaker style similarity between the source and target generated speech across all languages. In particular, we observe a large speaker style similarity increase for German of 0.04 → 0.39 between stages pretrain$_2$ and fine-tuning.

In order to evaluate the rhythmic capabilities of Sonar Expressive, we report both the speech rate Spearman correlation and pause alignment results in Tables 6 and 7 respectively. Similar to our observations on speaker style similarity, we notice large increases across both metrics and all languages once the SpeechProp embedding is introduced.

Results from sentence-level prosodic similarity using the AutoPCP metric are shown in Table 8. As defined by the Prosody Consistency Protocol (cf. subsection 5.2), a score of 1 corresponds

	Fleurs			mExpresso		
	pre-train$_1$	pre-train$_2$	fine-tune	pre-train$_1$	pre-train$_2$	fine-tune
SpeechProp	✗	✗	✓	✗	✗	✓
cmn	0.02	0.19	0.45	0.15	0.07	0.34
deu	0.01	0.24	0.49	0.03	0.14	0.34
fra	0.01	0.30	0.49	0.06	0.12	0.39
ita	0.00	0.18	0.42	0.05	0.14	0.32
spa	0.00	0.31	0.49	0.04	0.14	0.33

Table 7: Pause alignment results in zero-shot S2ST.

	Fleurs			mExpresso		
	pre-train$_1$	pre-train$_2$	fine-tune	pre-train$_1$	pre-train$_2$	fine-tune
SpeechProp	✗	✗	✓	✗	✗	✓
cmn	1.54	2.42	2.90	2.29	2.46	3.24
deu	1.30	2.28	2.92	1.99	2.41	3.11
fra	1.69	2.67	3.10	1.92	2.43	3.13
ita	1.16	2.40	2.87	1.99	2.41	3.23
spa	1.78	2.64	3.01	2.05	2.51	3.18

Table 8: AutoPCP results in zero-shot S2ST.

to “very different” prosody, 2 to “some similarities, but more differences”, 3 to “some differences, but more similarities”, and 4 to “very similar” We notice that our speech-to-speech models with SpeechProp embeddings produces expressive speech with a predicted PCP score of around 3. This grade is qualified in the evaluation protocol as having “some differences, but more similarities”, highlighting the expressivity preservation of the output translated speech.

5.5 Data generation with Sonar Expressive

Back-translation has been heavily used to augment training datasets for machine translation (Schwenk, 2009; Sennrich et al., 2015; Edunov et al., 2018; NLLB Team et al., 2022), using generated translations as input to train machine translation systems. In the same spirit, pseudo-labeling with cascade systems for speech-to-text and speech-to-speech translation to overcome training data scarcity was also widely explored (Pino et al., 2020; Jia et al., 2019a; Dong et al., 2022). Finally, generated data was also used to fine-tune Large Language Models (LLMs) in order to better align with human preferences. For example, Touvron et al. (2023), used a pre-trained language model to generate several answers. Each answer is ranked by a reward model, and the top predictions are used as gold labels to fine-tune the model. They refer to this technique as Rejection Sampling fine-tuning.

Inspired by such methods, we used Sonar Expressive to generate expressive speech translations for Seamless Communication et al. (2023b). In order to generate new data, we leverage the same publicly available data which was used for pre-training (cf. subsection 4.1). SHAS segments from each target language were then translated into English text using Sonar encoders/decoders, and then we expressively decoded each segmented into English speech.

6 Conclusion

We trained a speech decoder in the SONAR framework which is capable of decoding both multimodal and multilingual SONAR sentence embeddings into expressive speech. We showed that the expressive and prosodic content of the input speech can be encoded into a separate SpeechProp embedding which is disentangled from the Sonar semantic representations. Our multi-stage training approach shows that by initially training on unlabeled monolingual speech data only, and later introducing non-expressivity aligned S2T data, we are capable of generating expressively-aligned target speech in a zero-shot cross-modal way. We validated our approach with various expressivity preservation metrics. Moreover, the trained SPEECHPROP embedding appears to be language-agnostic and could potentially be applied to other spoken languages. We also show that this fixed-sized bottleneck representation for expressive and prosodic speech properties manages to capture locations of locally uttered pauses, and also has knowledge of speech rates. Since the speech decoder is based on SONAR, we can use it to decode any language or modality in a zero-shot way with reasonable results using pretrained SONAR encoders. As future work, we would like to extend speech decoders to more languages and explore multilingual speech decoders based on shared EnCodec units.

7 Acknowledgments

We would like to thank Ilia Kulikov for the training of the EnCodec model used for this project, Loic Barrault for his help on SONAR text decoding at scale, and Hady Elsahar for his help on SHAS segmention.

References

Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. 2018. Expressive speech synthesis via modeling expressions with variational autoencoder. *arXiv preprint arXiv:1804.02135*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Qianqian Dong, Zhiying Huang, Qiao Tian, Chen Xu, Tom Ko, Yunlong Zhao, Siyuan Feng, Tang Li, Kexin Wang, Xuxin Cheng, Fengpeng Yue, Ye Bai, Xi Chen, Lu Lu, Zejun Ma, Yuping Wang, Mingxuan Wang, and Yuxuan Wang. 2023. Polyvoice: Language models for speech to speech translation. *arXiv preprint arXiv:2306.02982*.

Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, Qibing Bai, and Yu Zhang. 2022. Leveraging pseudo-labeled data to improve direct speech-to-speech translation. *arXiv preprint arXiv:2205.08993*.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. 2022a. Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations. *arXiv preprint arXiv:2211.04508*.

Paul-Ambroise Duquenne, Hongyu Gong, Benoît Sagot, and Holger Schwenk. 2022b. T-Modules: Translation modules for zero-shot cross-modal machine translation. In *EMNLP*, page 5794–5806.

Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. 2021. Multimodal and multilingual embeddings for large-scale speech mining. *Advances in Neural Information Processing Systems*, 34.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023a. SONAR: sentence-level multimodal and language-agnostic representations.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023b. Modular speech-to-text translation for zero-shot cross-modal transfer. In *Interspeech*.

Jarod Duret, Yannick Estève, and Titouan Parcollet. 2023. Learning multilingual expressive speech representation for prosody.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, and Ruoming Pang. 2018. Hierarchical generative modeling for controllable speech synthesis. *arXiv preprint arXiv:1810.07217*.

Wen-Chin Huang, Benjamin Peloquin, Justine Kao, Changhan Wang, Hongyu Gong, Elizabeth Salesky, Yossi Adi, Ann Lee, and Peng-Jen Chen. 2023. A holistic cascade system, benchmark, and human evaluation protocol for expressive speech-to-speech translation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019a. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184. IEEE.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning*, pages 10120–10134. PMLR.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019b. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer learning from speaker verification to multi-speaker text-to-speech synthesis. *Advances in neural information processing systems*, 31.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *arXiv preprint arXiv:2302.03540*.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. SAMU-XLSR: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *arXiv preprint arXiv:2205.08180*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan. 1997. Janus-iii: Speech-to-speech translation in multiple languages. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 99–102. IEEE.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *CoRR*, abs/2306.15687.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022. Direct speech-to-speech translation with discrete units. In *ACL*, pages 3327–3339.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2021. Textless speech-to-speech translation on real data.

M. Macary, M. Tahon, Yannick Estève, and Antony Rousseau. 2021. On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. 2006. The atr multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling human-centered machine translation.

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation. *arXiv preprint arXiv:2006.02490*.

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. *arXiv preprint arXiv:2104.00355*.

Leyuan Qu, Taihao Li, Cornelius Weber, Theresa Pekarek-Rosin, Fuji Ren, and Stefan Wermter. 2023. Disentangling prosody representations with unsupervised speech reconstruction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara N. Sainath, Johan Schalkwyk, Matthew Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirovic, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Havnø Frank. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Holger Schwenk. 2009. Investigations on large-scale lightly-supervised training for statistical machine translation. In *IWSLT*, pages 182–189.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023a. SeamlessM4T-massively multilingual & multimodal machine translation.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet Artyom Kozhevnikov, Gabriel Mejia, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao Ann Lee Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023b. Seamless: Multilingual expressive and streaming speech translation.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Silero. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ioannis Tsiamas, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation. *arXiv preprint arXiv:2202.04774*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Yongqi Wang, Jionghao Bai, Rongjie Huang, Ruiqi Li, Zhiqing Hong, and Zhou Zhao. 2023b. Speech-to-speech translation with discrete-unit-based style transfer. *arXiv preprint arXiv:2309.07566*.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, pages 5180–5189. PMLR.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.