Sentence Embeddings for Massively Multilingual Speech and Text Processing

Paul-Ambroise Duquenne

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

Spécialité **Informatique**

École Doctorale Informatique, Télécommunications et Électronique (Paris)

Sentence Embeddings for Massively Multilingual Speech and Text Processing

Représentations vectorielles de phrases pour le traitement massivement multilingue du texte et de la parole

Présentée par

Paul-Ambroise DUQUENNE

Dirigée par

Benoît SAGOT

et encadrée en entreprise par

Holger SCHWENK

Pour obtenir le grade de

DOCTEUR de SORBONNE UNIVERSITÉ

Présentée et soutenue publiquement le 14 Mars 2024

Devant le jury composé de :

Alexandra Birch *Associate Professor, University of Edinburgh*	Rapportrice
Ondřej Bojar *Assistant Professor, Charles University*	Rapporteur
Laurent Besacier *Directeur de recherche, Naver Labs Europe*	Examinateur et Président du jury
Benjamin Piwowarski *Chargé de Recherches HDR, CNRS*	Examinateur
Benoît Sagot *Directeur de recherche, Inria Paris*	Directeur de thèse
Holger Schwenk *Directeur de recherche, META*	Encadrant en entreprise

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Representation learning of sentences has been widely studied in NLP. While many works have explored different pre-training objectives to create contextual representations from sentences, several others have focused on learning sentence embeddings for multiple languages with the aim of closely encoding paraphrases and translations in the sentence embedding space.

In this thesis, we first study how to extend text sentence embedding spaces to the speech modality in order to build a multilingual speech/text sentence embedding space. Next, we explore how to use this multilingual and multimodal sentence embedding space for large-scale speech mining. This allows us to automatically create alignments between written and spoken sentences in different languages. For high similarity thresholds in the latent space, aligned sentences can be considered as translations. If the alignments involve written sentences on one side and spoken sentences on the other, then these are potential speech-to-text translations. If the alignments involve on both sides spoken sentences, then these are potential speech-to-speech translations. To validate the quality of the mined data, we train speech-to-text translation models and speech-to-speech translation models. We show that adding the automatically mined data significantly improves the quality of the learned translation models, demonstrating the quality of the alignments and the usefulness of the mined data.

Then, we study how to decode these sentence embeddings into text or speech in different languages. We explore several methods for training decoders and analyze their robustness to modalities/languages not seen during training, to evaluate cross-lingual and cross-modal transfers. We demonstrate that we could perform zero-shot cross-modal translation in this framework, achieving translation results close to systems learned in a supervised manner with a cross-attention mechanism. The compatibility between speech/text representations from different languages enables these very good performances, despite an intermediate fixed-size representation.

Finally, we develop a new state-of-the-art massively multilingual speech/text sentence embedding space, named SONAR, based on conclusions drawn from the first two projects. We study different objective functions to learn such a space and we analyze their impact on the organization of the space as well as on the capabilities to decode these representations. We show that such sentence embedding space outperform previous state-of-the-art methods for both cross-lingual and cross-modal similarity search as well as decoding capabilities. This

new space covers 200 written languages and 37 spoken languages. It also offers text translation results close to the NLLB system on which it is based, and speech translation results competitive with the Whisper supervised system. We also present SONAR EXPRESSIVE, which introduces an additional representation encoding non-semantic speech properties, such as vocal style or expressivity of speech.

RÉSUMÉ

L'apprentissage de représentations mathématiques des phrases, sous forme textuelle, a été largement étudié en traitement automatique des langues (TAL). Alors que de nombreuses recherches ont exploré différentes fonctions d'objectif de pré-entraînement pour créer des représentations contextuelles des mots à partir des phrases, d'autres se sont concentrées sur l'apprentissage de représentations des phrases par des vecteurs uniques, ou représentations de taille fixe (par opposition à une séquence de vecteurs dont la longueur dépend de la longueur de la phrase), pour plusieurs langues. Le but étant d'encoder par des vecteurs proches entre eux les paraphrases et les traductions d'une même phrase.

Dans cette thèse, nous étudions d'abord comment étendre ces espaces de représentations de phrases à la modalité de la parole afin de construire un espace de représentation de phrases multilingue pour la parole et le texte. Ensuite, nous explorons comment utiliser cet espace de représentation de phrase multilingue et multimodal pour de la recherche de similarité sémantique entre des phrases parlées et écrites à grande échelle. Ceci nous permet de créer automatiquement des alignements entre des phrases écrites et parlées dans différentes langues. Pour des seuils de similarité élevés dans l'espace de représentation, les phrases alignées peuvent être considérées comme des traductions. Si les alignements impliquent d'un côté des phrases écrites et de l'autre des phrases parlées, il s'agit alors de potentielles traductions parole-texte. Si les alignements impliquent des deux côtés des phrases parlées, il s'agit alors de potentielles traductions parole-parole. Pour valider la qualité des données collectées automatiquement, nous entraînons des modèles de traduction de la parole vers le texte et des modèles de traduction parole vers parole. Nous montrons qu'ajouter les données alignées automatiquement améliore significativement la qualité du modèle de traduction appris, démontrant la qualité des alignements et l'utilité des données automatiquement alignées.

Ensuite, nous étudions comment décoder ces représentations vectorielles de phrases en texte ou parole dans différentes langues. Nous explorons plusieurs méthodes d'apprentissage de modèles décodeurs et analysons leur robustesse pour décoder des représentations de phrases de langues/modalités non observées pendant leur apprentissage, afin de quantifier leur capacité de généralisation et le transfert entre langues et entre modalités des capacités de décodage. Nous mettons en évidence que l'on peut atteindre des résultats de traduction d'une modalité à l'autre proches de systèmes appris de manière supervisée avec un mécanisme d'attention. La compatibilité des représentations parole/texte

dans différentes langues permet ces très bonnes performances, malgré une représentation intermédiaire composée d'un seul vecteur.

Enfin, nous montrons comment nous avons développé un nouvel espace de représentation de phrases pour la parole et le texte qui améliore l'état de l'art nommé SONAR, grâce aux enseignements tirés de nos travaux précédents. Nous étudions différentes fonctions d'objectif pour l'apprentissage de cet espace et nous analysons leur impact sur l'organisation de l'espace ainsi que sur les capacités de décodage des représentations. Nous montrons que ce nouvel espace de représentation de phrases améliore significativement l'état de l'art pour la recherche de similarité entre langues et entre modalités ainsi que les capacités de décodage de ces représentations. Ce nouvel espace couvre 200 langues écrites et 37 langues parlées. Il offre également des résultats en traduction du texte proche du système de traduction NLLB sur lequel il se base, et en traduction de la parole compétitifs avec le système supervisé Whisper. Nous présentons également SONAR EXPRESSIVE, qui introduit une représentation supplémentaire encodant des propriétés de la parole non sémantiques telles que la voix ou l'expressivité.

REMERCIEMENTS

I would like to express my deepest gratitude to all those who have contributed, either directly or indirectly, to the successful completion of this PhD project. First and foremost, I would like to thank my PhD directors Holger Schwenk and Benoît Sagot who made this project possible.

Thank you Holger for providing me with the opportunity to start an internship with you and for introducing me to the field of sentence embeddings research. I remember the first time we met, when you showed me your live demo on multilingual similarity search at scale on your Linux computer, I was already enthusiastic about the research to come. Thank you for your support and guidance throughout the PhD journey, always encouraging me to scale experiments and making the project grow through collaborations. Thank you Benoît for your trust from day one in this PhD adventure. I remember presenting you my internship project through Zoom, while being on vacation in Corsica, in order to convince you to become my PhD director. What a journey it has been since then! I would like to thank you for the great scientific discussions and advice, your help on the writing of papers, and your guidance as my PhD director. Beyond the research, it has been a real pleasure working with you both.

I would also like to deeply thank the members of the jury. Thank you Alexandra Birch and Ondřej Bojar for your thorough reviews of my PhD manuscript. Thank you also to Laurent Besacier and Benjamin Piwowarski for your participation in the jury as well as in the yearly *Comités de suivi*.

Thank you to all my colleagues who I had the chance to collaborate with. Hongyu for the first collaborations of my PhD with SpeechLASER and SpeechMatrix papers. Kevin for our great collaboration on the SONAR Expressive paper. The former Universal Speech Translation team and especially Ann, Changhan and Juan for the help and discussions around Speech Translation research. All members of the Seamless team, from close EMEA colleagues to cross-Atlantic colleagues with whom I spent great times during offsites. Onur for our collaboration on scaling speech mining for the Seamless project. Alex for your trust and your help to transition full-time at Meta. All the ALMAnaCH team for the great interactions I had when coming to INRIA.

J'ai une pensée toute particulière pour ma femme Domitille qui m'a accompagné dans la thèse. Merci ma Domi pour ton soutien, ton énergie et ton écoute, j'ai une chance immense de t'avoir à mes côtés. Je tiens à remercier mes parents qui m'ont beaucoup transmis, du goût des mathématiques à la recherche

de l'excellence. Merci pour votre soutien tout au long de mes études. Merci à ma sœur Eve-Marie et à mon frère Côme pour notre fratrie unie depuis toujours.

Merci également à mes amis doctorants, JB, Lina, Tariq et Tu Anh, avec qui j'ai passé de très bons moments pendant les pauses FIFA par exemple. J'ai aussi une pensée pour mes très bons amis qui sont à mes côtés depuis toutes ces années, à Greg, PA, GCO et Inès qui sont aussi tombés dans l'IA et aux autres longues et belles amitiés, en particulier Jérôme, Fara, Cyp, Arnaud, Vinc' et cie.

ACRONYMS

AE	Auto-Encoding
ASR	Automatic Speech Recognition
BLEU	BiLingual Evaluation Understudy
BOS	Beginning-Of-Sentence
CLS	Classification
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
CV	Common Voice
DAE	Denoising Auto-Encoding
DCT	Discrete Cosine Transform
ELMo	Embeddings from Language Models
EOS	End-Of-Sentence
EPST	EuroparlST
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
LSTM	Long Short-Term Memory
MFCC	Mel Frequency Cepstral Coefficients
MLM	Masked Language Modeling
MLP	Multi-Layer Perceptron
MLS	Multilingual LibriSpeech
MoE	Mixture-of-Experts
MSE	Mean-Squared Error
MT	Machine Translation
NLLB	No Language Left Behind
NLP	Natural Language Processing
PCP	Prosodic Consistency Protocol
QA	Question answering
RNN	Recurrent Neural Network
S2S	Speech-to-Speech
S2ST	Speech-to-Speech Translation

S2T	Speech-to-Text
S2TT	Speech-to-Text Translation
S2U	Speech-to-Unit
SNR	Signal-to-Noise Ratio
SOTA	State-Of-The-Art
STFT	Short-Time Fourier Transform
T2ST	Text-to-Speech Translation
TTS	Text-to-Speech
UST	Universal Speech Translation
VAD	Voice Activity Detection
VP	VoxPopuli
WER	Word Error Rate

CHAPTER 1

INTRODUCTION

> La parole est une sorte de tableau dont la pensée est l'original.
>
> *Denis Diderot, L'Encyclopédie Vol. VII*

"Language is a kind of painting of which the thought is the original.", one can read in *L'Encyclopédie* from Denis Diderot. This quote first conveys the idea that while an idea can exist as a thought, one has to express it with language to communicate it to others. Language has developed throughout history to enable humans to communicate between each other and has therefore played a major role in human societies. This expression of thoughts is imperfect, as Diderot highlighted, but tries to be the most *faithful imitation* of thoughts (Diderot and d'Alembert 1751). This is also true with translation as no perfect translation exists. Indeed, multiple possible translations are often commonly accepted for the same source sentence. Moreover, some words or expressions may carry a cultural background that cannot be easily translated. However, translation has also played a major role in human history, allowing people from different parts of the world to communicate, enabling strong diplomatic relationships or the sharing of knowledge between different countries.

Since communication is at the heart of human society, dreams of real-time translation technologies have appeared in several fictions like Star Trek's Universal Translator, a computational device that offers translations between any two languages. To tend towards such technologies, there has been a lot of efforts to create innovations that can support people for translation for many years. For example, as early as 1663, Kircher proposed in *Polygraphia Nova*, a system to translate from one language to another using a shared code, which can be seen as a first attempt to perform word-by-word translation. In 1933, Georges Artsrouni filed and received the first patent for a mechanical translation device. With the premises of computer science, machine translation started to be explored (Booth and Richens 1952). Important improvements in machine translation quality was enabled by machine learning methods which started in 1990 (Brown et al. 1990).

If we analyze again the quote from Denis Diderot, we notice that thought is the origin, and that language is the mean of expression of it, which tries to

be as faithful as possible to the original thought. In this context, the core idea of this thesis is a conceptual space where representations can be instantiated into different languages and modalities and where similar concepts (or ideas practically defined as sentences) have similar representations.

While representation of words have been studied a lot in machine learning, other works also studied representations at the sentence level to build sentence vector representations, also called sentence embeddings (Kiros et al. 2015). We advocate that the sentence is a good scale to build a high-level conceptual representation space. Indeed, translation has been mainly addressed at the sentence level (Cho et al. 2014), even though document-level machine translation should also be addressed for long term consistency and sometimes disambiguation (Barrault et al. 2019). Sentence representations were also explored for classification tasks, as well as semantic similarity estimation for several languages (Artetxe and Schwenk 2019b).

This thesis is situated at the intersection of representation learning of sentences with multilingual speech/text sentence embeddings and multimodal speech/text translation.

1.1 Introduction to the research problem

Representation of textual data in Natural Language Processing (NLP) has first been explored with non-contextual word embeddings (Mikolov et al. 2013), followed by contextual representations of words (Devlin et al. 2019; Conneau et al. 2020b). A sentence is then represented as a sequence of word embeddings, and these representations have variable lengths depending on the number of words of the encoded sentence (Arora et al. 2017). Representations of sentences by unique vectors (also called fixed-size representations) have also been explored for classification purposes or to efficiently estimate semantic similarity between sentences (from potentially different languages) (Artetxe and Schwenk 2019b; Feng et al. 2020). Those fixed-size representations for sentences are commonly called sentence embeddings and have been extremely useful for bitext mining at scale (Schwenk et al. 2021; Ramesh et al. 2022). Semantic sentence embeddings for multiple languages in the speech modality were left unexplored when this thesis started.

In recent sequence-to-sequence models, composed of an encoder and a decoder, the encoded source sentence is represented as a sequence of contextual representations that a decoder can attend to (Bahdanau et al. 2014; Vaswani et al. 2017). This attention mechanism on all encoder outputs has significantly boosted performances of such models. In that context, decoding fixed-size representations

of sentences has been largely under-explored in NLP with modern architectures like Transformers (Vaswani et al. 2017). Exploring how much information can be encoded and decoded into fixed-size sentence representations is therefore interesting by itself. Moreover, when dealing with multiple modalities, like speech and text, multi-modal sequence-to-sequence models have difficulties to represent similarly text and speech inputs, which hinder efficient cross-modal transfer. Indeed, sequences of contextual representations of speech and text have really different lengths, which is the first reason of this so-called modality gap (Liu et al. 2020b). Exploring fixed-size representations at the sentence level, enables to minimize the modality gap much more easily, with opportunities to explicitly align modalities in the sentence embedding space. Given this research context, we try in this thesis to address the following questions:

How can we build language-agnostic and modality-agnostic sentence embeddings for best semantic similarity estimation between languages and modalities? How much content is preserved and can be recovered from these multilingual and multimodal fixed-size sentence representations? We conclude with perspectives that such compatible representations between speech and text for multiple languages may open.

To answer these questions, throughout the thesis, we extend existing semantic representations of text sentences to the speech modality for multiple languages. We explore decoding of such multilingual and multimodal sentence representations into speech and text in different languages. Finally, we introduce a new sentence embedding space for multilingual speech and text, augmented with a speech-specific representation for non-semantic encoding of speech.

1.2 PhD thesis context

This thesis is part of the CIFRE PhD French program involving both an academic institution (INRIA Paris, ALMAnaCH team) and an industrial lab (Meta AI, FAIR team) and followed an internship at Meta AI. In this section, we quickly contextualize this thesis with related research that was happening in the two labs.

This thesis was initially heavily relying on the existing LASER sentence embedding space developed by Mikel Artetxe and Holger Schwenk at Meta AI, a massively multilingual sentence embedding space with interesting semantic properties that has proven to be useful for bitext mining (Schwenk et al. 2019; Schwenk et al. 2021). The LASER sentence embedding space was also successfully used in a project led by Benoît Sagot with one of his former PhD students Louis Martin to mine paraphrases for multiple languages (Martin et al. 2020). Representation learning has also been addressed in the ALMAnaCH team with

the learning of contextual representations for French text with the well-known CamemBERT (Martin et al. 2019) model.

The beginning of the thesis coincides with the launch of an internal Meta AI project, later released as No Language Left Behind (NLLB), a state-of-the-art machine translation model for massively multilingual text, where bitext mining was scaled to many new languages by leveraging LASER3 (Heffernan et al. 2022). While I did not take part in the NLLB project, I collaborated with the Universal Speech Translation (UST) team at Meta on the SpeechMatrix project, where we scaled speech-to-speech mining to 136 language pairs. Finally, I participated in the recent Seamless Communication project (Seamless Communication et al. 2023a; Seamless Communication et al. 2023b) with the integration of our new SONAR sentence embedding space for large-scale speech mining.

1.3 Contributions

This PhD thesis is structured around exploring massively multilingual speech/text sentence embeddings for large-scale speech mining and how such representations may be decoded into multiple languages in text and speech. We first present the related work in Chapter 2, before presenting our contributions:

- Chapter 3: EMBEDDING SPEECH/TEXT SENTENCES AND MULTILINGUAL SPEECH MINING
  In this chapter, we introduce multilingual and multimodal speech/text sentence embeddings using a teacher-student approach with the existing LASER sentence embedding space. We demonstrate that we can perform semantic similarity estimation between speech and text in different languages and introduce speech mining as an extension of bitext mining for the speech modality. We train speech translation systems using the mined data and demonstrate significant gains with this additional data. Based on these promising results, we scale speech-to-speech mining to 136 language pairs to introduce the SpeechMatrix corpus and train several speech translation systems on this mined data. The work in this chapter has led to two conference publications:

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

  - Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswami, Changhan Wang, Juan Pino, Benoit Sagot, and

Holger Schwenk (July 2023a). "SpeechMatrix: A Large-Scale Mined Corpus of Multilingual Speech-to-Speech Translations". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 16251–16269. URL: https://aclanthology.org/2023.acl-long.899

- Chapter 4: DECODING SENTENCE EMBEDDINGS AND ZERO-SHOT CROSS-MODAL MACHINE TRANSLATION
  Then, we explore how to efficiently decode these fixed-size representations into multiple languages and modalities and how we can perform zero-shot cross-modal machine translation in this framework. We demonstrate that we can combine independently trained encoders and decoders from different languages and modalities in a zero-shot way to perform cross-modal translation. As a second step, we explore multilingual training in such modular framework, to benefit from cross-lingual learning. The work in this chapter has led to two conference publications:

  - Paul-Ambroise Duquenne, Hongyu Gong, Benoit Sagot, and Holger Schwenk (Dec. 2022). "T-Modules: Translation Modules for Zero-Shot Cross-Modal Machine Translation". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5794–5806. URL: https://aclanthology.org/2022.emnlp-main.391

  - Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot (2023c). "Modular Speech-to-Text Translation for Zero-Shot Cross-Modal Transfer". In: *Proc. INTERSPEECH 2023*, pp. 32–36

- Chapter 5: SONAR: UTTERANCE-LEVEL REPRESENTATIONS FOR MASSIVELY MULTILINGUAL SPEECH AND TEXT
  Finally, in our last chapter, we draw conclusions from the two first chapters to introduce SONAR, a state-of-the-art massively multilingual speech/text sentence embedding space for both cross-lingual and cross-modal similarity search as well as decoding capabilities. We complement these semantic sentence representations with a modality specific representation encoding non-semantic speech properties of an audio signal. The work in this chapter has led to two preprint publications:

- Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot (2023d). *SONAR: Sentence-Level Multimodal and Language-Agnostic Representations*. URL: https://arxiv.org/abs/2308.11466

- Paul-Ambroise Duquenne, Kevin Heffernan, Alexandre Mourachko, Benoît Sagot, and Holger Schwenk (2023b). *SONAR EXPRESSIVE: Zero-shot Expressive Speech-to-Speech Translation*

Finally, we put these contributions into perspective and discuss potential future research directions in Chapter 6.

CHAPTER 2

RELATED WORK

In this chapter, we detail the related work of this thesis. We first introduce machine learning and the specific fields of Natural Language Processing (NLP) and Speech Processing. Then, we present different methods to learn representations for sentences from contextual representations to sentence embeddings for multiple languages. Finally, we present related work about multimodal communication tasks including Machine Translation (MT), Automatic Speech Recognition (ASR), Speech-to-Text Translation (S2TT), Speech-to-Speech Translation (S2ST) and Expressive speech generation. Throughout this chapter, we also introduce the different evaluation strategies commonly used to evaluate the presented tasks.

2.1 From machine learning to natural language processing and speech processing

2.1.1 Machine learning and neural networks

Machine learning is the general field of algorithms that learn from data. Supervised learning algorithms are trained to solve specific tasks, which can be modeled as a mapping between input features X and their label y. For example, a translation task can be modeled as a mapping of source sentences in a given language to target sentences in another language. Machine learning can then be seen as the algorithms to find (or learn) the best mapping between X and y, called the model, based on some labeled training data. The best model is searched inside a model class. In neural network learning, the model class is defined by the architecture of the network as well as parameters of the model. The architecture is often fixed before training, and the machine learning algorithm is searching for best parameters to solve the task on the training data. This is actually an optimization problem, where the parameters are chosen to minimize a loss function. This optimization is practically approximated in neural network learning with gradient descent (Cauchy et al. 1847; Rumelhart et al. 1986).

The goal of supervised learning algorithms is to solve a specific task for any data drawn from an unknown distribution p_{data}. As this theoretical distribution is unknown, only a set of instances is sampled and labeled as the training set. During the training process, the parameters are found to minimize the loss on the training set.

However, the generalization error or test error of the model is measured on instances unseen during training called the test set, independently drawn from p_{data}. Indeed, the training process may lead to a low training error but a high test error, which is commonly called over-fitting. This happens when the model class is too complex and without enough training data: a complex model is learned to approximate noise and outliers of the training set rather than the true underlying pattern of labels and instances from p_{data}. Regularization techniques can help avoid over-fitting by adding constraints to encourage simpler models (parameter norm penalties like weight decay), training the model on several similar tasks to increase generalization, early stopping of the training process, adding noise in the training (e.g. dropout), using data augmentation techniques etc. Collecting more training data or simplifying the model class are also two other ways to avoid over-fitting.

On the other hand, training error may still be high at the end of the training process which is called under-fitting. This happens when the model class is too simple and no model in this model class can correctly approximate the underlying mapping between X and y. Exploring different model classes is important to avoid under-fitting.

Once the model is trained and evaluated on a separate test set, it can be used to make predictions. Machine learning has many different applications like image classification (L. Chen et al. 2021), recommender systems, machine translation (Brown et al. 1990) and speech recognition (Bahl et al. 1987).

2.1.2 Natural language processing

Among machine learning applications, several ones deal with text data and are often grouped as NLP tasks. This set of tasks has specificities that we are going to develop in this section.

The first specificity of NLP tasks is the nature of the textual data. Indeed, neural networks take vector representations as input and output. To apply neural networks to NLP tasks, one should first represent text as vectors. There has been a long history of vector representations of words or subwords (a word split in several parts) since the beginning of NLP research. Text is then represented by sequences of vector representations of each word or subword.

A first method to represent words as vectors is called "one-hot" representations. Based on a word vocabulary of size N, a word is represented by a vector of dimension N filled with zeros except at index i (filled with 1), which corresponds to the index of the encoded word in the vocabulary list. In this way, each word of the vocabulary can be easily represented. However, the dimensionality of the vector space increases with the size of the vocabulary, leading to high dimensional representation spaces for words. Another issue with such representations is that representations of similar words have nothing in common when represented as one-hot vectors, which forbids models to generalize well when trained on a subset of possible words.

To overcome this issue, distributional approaches were introduced. The main idea is that similar words are appearing in similar contexts (Harris 1954; Firth 1957). Extracting contexts and co-occurrences of words in easily-available raw text corpora can be used to build vector representations of words, called word embeddings.

An initial distributional approach, commonly referred as count-based methods, counts co-occurrence of a word with other words found in its contexts in the raw text corpus. As a result, words which have similar statistics of neighbouring words in the raw corpus will be represented similarly if one takes these counts as a vector representation for each word (Church and Hanks 1990).

Another distributional approach can be summarized as predictive methods. Word2vec (Mikolov et al. 2013) is maybe the most well-known predictive approach to build word embeddings, where a neural network is trained to either predict a word from its neighbouring words (CBoW method) or to predict its neighbouring words from the word itself (Skip-gram method). The projection hidden layer is used as words embeddings. Mikolov et al. (2013) have shown that such embedding spaces have interesting semantic properties. Indeed, interestingly, some simple linear operations called delta vectors in the embedding space highlight the good semantic organization of the space. For example, one can compute embeddings for the words "man", "woman", "king" and "queen" among other words of a vocabulary, as e_{man}, e_{woman}, e_{king} and e_{queen}. Computing the delta vector $e_{delta} = e_{king} - e_{man}$ and adding it to e_{woman} results in a word embedding close to e_{queen}. This is illustrated, among other examples, in Figure 2.1.

Many other predictive methods exist to build word embeddings such as (Bojanowski et al. 2017; Pennington et al. 2014). All the word embedding methods presented so far are called non-contextual embeddings, as each word is represented the same way in different contexts. However, in many cases, the meaning of a word is ambiguous if taken separately from its context. This is why contextual representations of words were introduced, and we will develop such methods in Section 2.2.1.

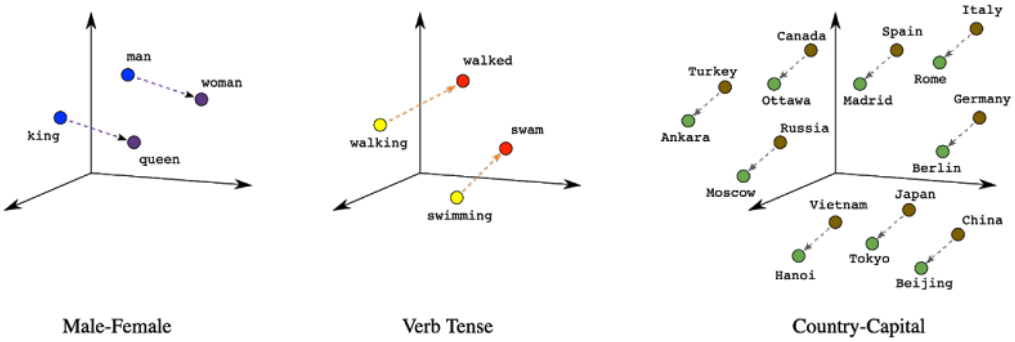

Figure 2.1. – **Semantic transformations in Word2vec embedding space.** Semantic transformations can be modeled as delta vectors in the word embedding space. *Image from Google Cloud blog on Word2vec*.

Another specificity of NLP tasks is the variable lengths of inputs and outputs. Simple neural network architectures, like Multi-Layer Perceptron (MLP), were adapted to perform sequence modeling like Recurrent Neural Network (RNN). RNN architectures (Rumelhart et al. 1986) are a specific type of neural network sharing weights for each time-step of the input sequence and taking as input both the current time-step input representation as well as the previous recurrent hidden representation. When dealing with long-term dependencies, these architectures often suffer from either vanishing or exploding gradients (Hochreiter 1991; Bengio et al. 1994) which makes the learning process with gradient descent difficult. To overcome these issues, other architectures like Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) were introduced.

Sequence-to-sequence architectures were also introduced to produce variable-length outputs, and successfully used for tasks like MT. Sequence-to-sequence architectures were initially composed of an RNN encoder and an RNN decoder. The decoder takes as first recurrent input the last recurrent hidden representation of the encoder, which implies that the full input sequence should be encoded in the last hidden representation of the encoder. Attention mechanism was introduced so that the decoder may take as input a learned weighted sum of encoder hidden representations for each time-step (Bahdanau et al. 2014). This attention mechanism was extended to encoder and decoder internal model architectures with multi-head self-attention in the Transformer architecture (Vaswani et al. 2017) in addition to cross-attention between the decoder and the encoder outputs.

While contextual representations and pre-trained language models are also key in NLP (and will be presented in a following section), these non-contextual word representations and specific architectures represent the basics of the NLP field which aim at addressing several tasks like text classification, sentiment

analysis, question answering, information retrieval, dialogue generation, machine translation or automatic summarization.

2.1.3 Speech processing

Dealing with speech is really different as input features are continuous, contrarily to text data.

In some modern end-to-end systems, 1-D raw waveforms are directly used as input. However, historically and in many current systems, pre-computed features of speech are used, like mel-filterbanks or Mel Frequency Cepstral Coefficients (MFCC), which tend to replicate the processing of speech in the human inner ear (Stevens and Volkmann 1940). The main steps to extract those features is a pre-emphasis step to amplify high frequencies, followed by Short-Time Fourier Transform (STFT) on successive windowed signals (a Hamming window is used to avoid spectral leakage). Finally, filter-banks are applied on a mel-scale to mimic the non-linear human ear perception. This last representation of speech can be de-correlated using Discrete Cosine Transform (DCT) which gives MFCC features, that were successfully used in many speech processing models.

In addition to RNN models presented in Section 2.1.2, Convolutional Neural Network (CNN) (LeCun et al. 1989) have been successfully used to down-sample speech as well as computing local features. The CNN architecture is also based on the parameter-sharing idea, as same learned convolutional filters are applied at different places of the input. The CNN architecture was used for the first time for speech processing in the time-delay neural network (TDNN) architecture (Waibel et al. 1989; Lang et al. 1990). Nowadays, speech processing mainly uses a combination of convolutional and transformer layers, either using CNN as a feature extractor before feeding a transformer model like in (Baevski et al. 2020), or stacking self-attention layers and convolutional layers like in the Conformer architecture (Gulati et al. 2020).

Speech recognition has been the main area of research in speech processing and is described in Section 2.4.2, from statistical speech recognition to neural approaches and the recent architectures and objective functions for speech recognition that converge to methods similar to NLP ones. Another important application of speech processing is speech synthesis or Text-to-Speech (TTS) synthesis. The tasks of directly translating speech in a given spoken language into text or speech in another language is addressed as Direct S2TT or S2ST. These tasks are also developed in Section 2.4.2 and Section 2.4.3 and have seen large improvements with the emergence of pre-trained models for the speech modality, introduced in the next section. Other notable speech processing tasks are speaker

diarization (distinguishing between different speakers in an audio recording), keyword spotting (detecting keywords in an audio), language identification (classifying the language of an audio) or emotion recognition (classifying between different emotions conveyed in an audio).

2.2 Contextual representations and pre-trained models

In this section, we develop the main research on contextual representations for text and speech and for several languages. These contextual representations are obtained from pre-trained models which are at the core of many research works in NLP and speech processing.

2.2.1 Contextual representations and pre-trained models for text

In Section 2.1.2, we introduced non-contextual word embeddings which have been the basis of word representations in NLP. However, one would be interested in representing a word in a way that takes into account the context of this word. Indeed, some words with the same spelling may have really different meanings depending on their context. To introduce contextual representations of words, Peters et al. (2018) used internal states of a pre-trained language model based on the LSTM architecture as word representations, called Embeddings from Language Models (ELMo). They used a bidirectional LSTM architecture, which means that the model processes the input in both forward and backward directions. These representations of words are contextual as the same word with different right and left contexts will be represented differently.

Devlin et al. (2019) introduced the Masked Language Modeling (MLM) self-supervised task, where a model is trained to predict masked input tokens from context words on raw text data. They introduced the popular BERT model, pre-training a Transformer encoder with MLM task as well as a next sentence prediction task on large amount of unlabeled text from BookCorpus (Zhu et al. 2015) and English Wikipedia, totalling 3.3 Billion words. In addition to providing good contextual word embeddings, BERT can be used as a pre-trained model that can be fine-tuned on new NLP downstream tasks. With their BERT model, Devlin et al. (2019) improved the previous state-of-the-art on several Question answering (QA) and language inference tasks. Several variants of BERT were explored based on these promising results. Liu et al. (2019), with their Roberta

model, optimized training hyper-parameters, used more data and trained for more training steps. They also removed the next sentence prediction pre-training task which they showed was unnecessary for good performance on downstream tasks. Joshi et al. (2020) introduced SpanBERT with masking of contiguous spans instead of single subword tokens which brought better performances in downstream tasks, especially tasks requiring span selection. The ELECTRA work (K. Clark et al. 2020) introduced a more sample-efficient pre-training task which consists in discriminating between corrupted and original tokens. The corrupted tokens are sampled from another small MLM generator trained with maximum likelihood, but only the discriminator is kept after pre-training to be fine-tuned for downstream tasks.

A multilingual BERT model, called mBERT, was also introduced. mBERT has been pre-trained on Wikipedia text data for 104 languages. Similarly to mBERT, XLM (Conneau and Lample 2019) model presents multilingual pre-training either with unpaired multilingual text data or with additional bitext training data, introducing a Translation Language Modeling (TLM) loss as an extension of the MLM loss for bitext data. The unsupervised version of XLM was scaled with XLM-R (Conneau et al. 2020b). These methods demonstrate really good zero-shot cross-lingual transfer on XNLI (Conneau et al. 2018b) benchmark. Additionally, some monolingual pre-trained models were also introduced as specialized models for some languages like CamemBERT (Martin et al. 2019) for French and AraBERT (Antoun et al. 2020) for Arabic.

2.2.2 Contextual representations and pre-trained models for speech

Contextual representations of speech data was also explored, especially in recent years, after witnessing the important impact that self-supervised pre-training methods had on NLP.

The suite of wav2vec papers focused on speech self-supervised pre-training, starting back in 2019. The first wav2vec paper (Schneider et al. 2019) presented a pre-training method based on a noise contrastive binary classification task in order to improve speech recognition downstream tasks while using less training data. Later, vq-wav2vec (Baevski et al. 2019) introduced self-supervised learning of discrete representations of speech with Gumbel-Softmax or online k-means. Then, they apply a NLP pre-training method on quantized speech, similar to span-BERT. Finally, Baevski et al. (2020) presented Wav2vec2, pre-training a CNN feature encoder and a Transformer encoder with a contrastive objective over masked latent representations which are quantized and jointly learned. They demonstrated that

such pre-training method enables to outperform previous best semi-supervised baselines in speech recognition after fine-tuning.

Wav2vec2 pre-training was extended to the multilingual setting, introducing XLS-R (Conneau et al. 2020a) for 53 spoken languages. The multilingual pre-trained model is then fine-tuned on multilingual speech recognition tasks, and the authors demonstrated that it outperforms monolingual models trained independently for low-resource languages. XLS-R multilingual training was scaled on both model size and number of languages by Babu et al. (2021). They released 2B parameter model variants pre-trained on half a million hours of publicly available speech for 128 languages. They fine-tuned and evaluated their models on a broad range of speech tasks like speech recognition, speech translation, language identification and improve previous state-of-the-art on these tasks. Finally, in the MMS project, Pratap et al. (2023) significantly scaled the scope of languages and pre-trained a 1B XLS-R model on 491K hours in 1,406 languages.

Another main self-supervised approach for speech that builds contextual representations out of audio inputs is HuBERT (Hsu et al. 2021). HuBERT is based on a Transformer architecture and trained for several iterations with a masked prediction task inspired by BERT (Devlin et al. 2019), masking continuous input speech features. Discrete targets are extracted after an offline clustering ran before each training iteration. For the first iteration, k-means clustering is ran with 100 clusters on 39-dimensional MFCC features, whereas other iterations are running k-means with 500 clusters on hidden representations from the model of the previous iteration at some intermediate transformer layer. Fine-tuned HuBERT matches or outperforms Wav2vec2 results on speech recognition tasks and the pre-trained model has also proven to provide good contextual discrete representations of speech, heavily used in textless spoken language modeling (Lakhotia et al. 2021).

Following Wav2vec2 and HuBERT works, w2v-BERT (Y.-A. Chung et al. 2021) presented a self-supervised pre-training method that combines contrastive learning and masked prediction learning. The constrastive learning loss, inspired by Wav2vec2 work, enables to learn a discretization of input speech in order to produce discriminative speech discrete targets, while the MLM loss enables to learn contextual speech representations with a masked prediction task on discrete speech units, inspired by HuBERT work. However, contrarily to HuBERT, discrete targets are jointly learned and only a single training iteration is needed.

Finally, Chiu et al. (2022) presented a simple yet effective approach called BEST-RQ. BEST-RQ uses a masked prediction learning task where the discrete targets are provided by a random-projection quantizer: a randomly initialized matrix projects the input speech and a randomly initialized codebook is used to find the nearest neighbor index as target. This approach was successfully scaled in the

proprietary USM (Zhang et al. 2023) project using 12 Million hours of training data and covering more than 300 languages.

2.2.3 Joint speech/text contextual representations and pre-trained models

Joint speech-text pre-training has been addressed recently by several works to better initialize speech recognition or speech translation systems. In that context, Y.-A. Chung et al. (2020) jointly pre-trained speech and text encoders with un-paired speech or text data as well as paired speech-text data introducing cross-modal token-level and sequence-level alignment losses. Tang et al. (2022) introduced a unified speech-text pre-training method based on multitask self-supervised and supervised learning in an encoder-decoder architecture. Another encoder-decoder joint pre-training approach is called SpeechT5 (Ao et al. 2021), which leveraged only unlabeled speech and text data, and that, among other things, tried to unify speech and text representations with a shared vector quantization codebook for both modalities in a joint pre-training task.

Another line of work is the SLAM (Bapna et al. 2021) and mSLAM (Bapna et al. 2022) pre-trained models. SLAM introduced a multimodal joint pre-training method for speech and text for English, using a combination of traditional self-supervised losses on unlabeled data and Translation Language Modeling (TLM) loss (as introduced in XLM (Conneau and Lample 2019) but for cross-modal learning) as well as a Speech Text Matching (STM) loss on ASR training data. mSLAM (Bapna et al. 2022) extended the SLAM model to massively multilingual joint speech/text pre-training with the same objective functions (but the STM loss is replaced by a CTC loss and text is handled at the character level). After fine-tuning, mSLAM improved the previous state-of-the-art (defined by an XLS-R finetuned model (Babu et al. 2021)) on speech-to-text translation as well as other speech processing tasks.

Finally, following Wav2vec2, Baevski et al. (2022) extended speech pre-training to the multimodal setting with data2vec. Data2vec learned contextual representations of speech, text and images with the same objective for all modalities. It used masked prediction learning with a self-distillation approach where the teacher is an exponentially decaying average of the student.

2.3 From monolingual to multilingual fixed-size sentence representations

In this part, we further develop representation learning of sentences with fixed-size representations of text and speech utterances. As opposed to variable-length sentence representations, which involve the concatenation of contextual subword representations, sentence embeddings are single vectors representing whole sentences. We also introduce bitext mining as well as evaluation strategies for multilingual sentence embedding spaces.

2.3.1 Monolingual text sentence embeddings

Pre-trained encoder models like BERT achieve strong results on semantic similarity tasks when inputting a concatenation of two compared sentences to the model. However, finding most similar sentence pairs in a big set of sentences is a really computationally expensive operation if one follows such method (Reimers and Gurevych 2019). In order to efficiently compare sentences at scale (in terms of semantic similarity), fixed-size sentence representations were introduced, where a simple similarity metric computation in the embedding space directly provides an estimation of the similarity between sentences. The performance of such semantic similarity estimation heavily relies on the organization of sentences in the embedding space. We review in this section the main methods that were explored to build good monolingual sentence embedding spaces.

First, Skip-thought (Kiros et al. 2015) is well-known to have extended the Word2vec Skip-gram approach from words to sentences where a RNN-based encoder-decoder learns to predict neighbouring sentences. Other methods used labeled data to learn sentence embeddings. Among them, InferSent (Conneau et al. 2017) is a supervised method which trains a siamese Bi-LSTM model on Stanford Natural Language Inference (SNLI) (Bowman et al. 2015) and MultiGenre Natural Language Inference (multiNLI) (Williams et al. 2017) labeled data to build sentence embeddings. Universal Sentence Encoder (USE) (Cer et al. 2018) embeddings used a transformer model trained on both unsupervised learning tasks and supervised SNLI task.

While BERT was pre-trained with a CLS token to compute a global sentence representation, using this representation or mean-pooling of contextual representations (without fine-tuning) as a sentence embedding gives poor results on semantic similarity tasks (Reimers and Gurevych 2019), even worse that sentence embeddings obtained by averaging GloVe word embeddings

(Pennington et al. 2014). In that context, Reimers and Gurevych (2019) introduced sentence-BERT or SBERT, a sentence embedding encoder based on BERT and fine-tuned using a siamese network architecture that outperformed previous state-of-the-art sentence embeddings.

Then, notably, SimCSE (Gao et al. 2021) presented a simple unsupervised constrative approach based on dropout as augmentation for positive pairs which performed on par with previous supervised methods. They also introduced a supervised approach that provided even better results. Some unsupervised approaches were presented and outperformed simCSE, like DiffCSE (Chuang et al. 2022). Other contrastive methods, like DeCLUTR (Giorgi et al. 2020), took different nearby text spans from the same document as positive pairs.

As mentioned before, computing a mean-pooling of BERT contextual representations performs poorly on semantic textual similarity, when not fine-tuned. Some works (B. Li et al. 2020; Su et al. 2021) showed that BERT pre-training introduces a non-smooth anisotropic semantic space of sentences that may explain the poor performances on semantic similarity. This anisostropy of BERT sentence representations was addressed by BERT-flow (B. Li et al. 2020) and (Su et al. 2021) in order to produce better sentence embeddings for semantic similarity estimation.

Other methods, like Sentence-T5 (Ni et al. 2021) explored how to create sentence embeddings from encoder-decoder models like T5. Relatedly, TSDAE (K. Wang et al. 2021) explored sentence representations built out of an encoder-decoder model with a sequential denoising auto-encoding objective. Finally, data augmentation has also been used to improve fixed-size sentence representations like Augmented SBERT (Thakur et al. 2021). Some more recent methods leveraged LLMs to produce synthetic training data for sentence embedding models (J. Zhang et al. 2023; L. Wang et al. 2023).

More generic fixed-size text embeddings, beyond sentences only, were also explored like the OpenAI text embeddings (Neelakantan et al. 2022) scaling model size and using neighbouring texts as positive pairs, or the E5 embeddings (L. Wang et al. 2022) using weakly-supervised contrastive learning. These generic text embeddings appear to perform really well on semantic textual similarity for sentences.

2.3.2 Multilingual text sentence embeddings

Monolingual sentence embeddings have proven to be useful to efficiently compare sentences, in terms of semantic meaning, directly in the embedding space based on cosine similarity calculation. The interesting semantic property of

such embedding spaces is that paraphrases are closely encoded in the sentence embedding space. However, one may be interested in comparing sentences from different languages. Building a multilingual sentence embedding space could lead to an embedding space where not only paraphrases are close but also translations. We illustrate these monolingual and cross-lingual semantic properties of multilingual sentence embedding spaces in Figure 2.2. Moreover, the increasing number of labeled data in MT could be leveraged as good "cross-lingual paraphrases" which include inherent word re-ordering and paraphrasing properties of translations. In this section, we present the main approaches to learn such multilingual sentence embedding spaces. We will develop the evaluations and applications of these multilingual embedding space in the next section.

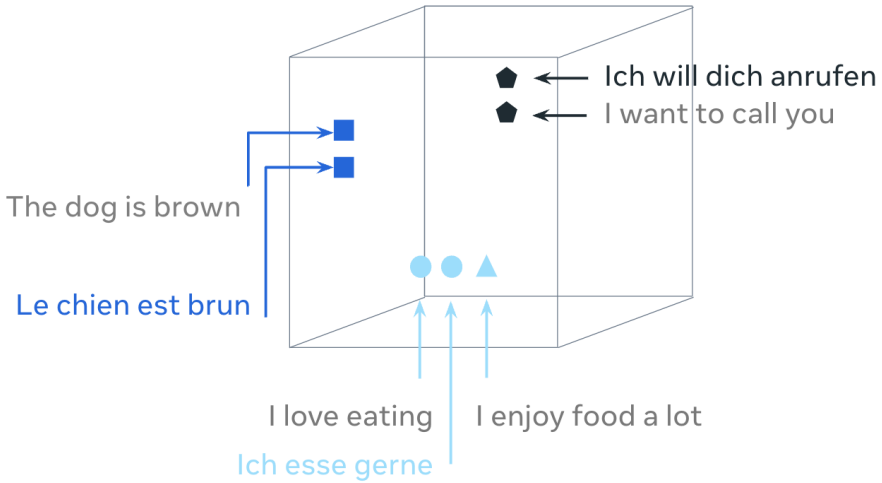

Figure 2.2. – **Illustration of multilingual sentence embeddings.** Text sentences from multiple languages can be encoded as vectors in a multilingual sentence embedding space. Paraphrases and translations are supposed to be close in the embedding space. *Adapted from Meta blog on LASER sentence embedding space.*

Multilingual vector representations for text started with cross-lingual word embeddings (Ruder et al. 2019; Gouws et al. 2015; Luong et al. 2015) which enabled to obtain sentence-level vector representations of text using a weighted sum of the word embeddings (Klementiev et al. 2012; Dufter et al. 2018). Sentence-level representations are also intrinsically learned in sequence-to-sequence RNN models and early works on these architectures (Sutskever et al. 2014) noticed that MT sequence-to-sequence models produce representations of similar sentences that are close in the embedding space. In that context, some works proposed and explored multilingual sentence representations for a few languages (Schwenk and Douze 2017; K. Yu et al. 2018; Schwenk 2018).

In parallel and subsequent works, massively multilingual representations were studied for word embeddings (Ammar et al. 2016), and then for contextual representations of words with XLM (Conneau and Lample 2019) and concurrently for sentence embeddings for 93 languages with LASER (Artetxe and Schwenk 2019b). Following LASER, LaBSE (Feng et al. 2020) was presented as another powerful massively multilingual embedding space that follows a dual encoder approach. These multilingual sentence embedding spaces were proven to be useful for massively multilingual bitext mining, as presented in the next section.

After this brief historical overview, we can summarize the main different methods to train such multilingual sentence embeddings.

First, sentence embedding spaces may be learned in a neural machine translation framework (España-Bonet et al. 2017; Schwenk and Douze 2017; Kvapilíková et al. 2020). In that context, the LASER (Artetxe and Schwenk 2019b) embedding space scaled sentence embedding training to the massively multilingual setting, improving previous state-of-the art results. It used an encoder-decoder LSTM model architecture, that builds a fixed-size representation of sentences by max-pooling encoder outputs. At each decoding time-step, the decoder takes as input the input sentence embedding, a language id token, as well as the previous predicted token. It is trained with a cross-entropy loss on bitext training data with English and Spanish as target languages (hence the language id token for the decoder, to either decode into English or Spanish).

Second, contrastive learning has been widely explored to build multilingual sentence embedding spaces with good semantic properties (Guo et al. 2018; Y. Yang et al. 2019a), explicitly aligning representations of translations in the embedding space while pushing apart some "negative" instances to avoid collapse (i.e. avoid to predict the same embedding for every input). Multitask learning on monolingual data can also be added to the training (Chidambaram et al. 2018; Z. Yang et al. 2020), and this line of work includes the m-USE embeddings (Y. Yang et al. 2019b). LaBSE (Feng et al. 2020) scaled language coverage for contrastive learning and leveraged a pre-trained encoder, presenting a state-of-the-art sentence embedding model for more than 100 languages.

Third, teacher-student training was introduced to extend a (possibly monolingual) pre-existing sentence embedding space to new languages. The existing embedding space is used as teacher to train student encoders for new languages. Bitext training data is used for this kind of training, where the sentence in the new language is encoded with a trained encoder, while its translation in another supported language is encoded with the pre-existing encoder as target. This approach was first introduced by (Reimers and Gurevych 2020) to extend a monolingual embedding space from SBERT (Reimers and Gurevych 2019) to new languages. This method was also used to present LASER3 (Heffernan et al.

2022), which improved and extended LASER for low-resource languages and outperformed LaBSE on several low-resource languages. Following LASER3, LASER3-CO (Tan et al. 2022) was introduced and also followed a distillation approach but based on contrastive learning.

Finally, some methods explored creation of language-agnostic sentence embeddings by isolating or removing language-specific information (Wieting et al. 2019; Tiyajamorn et al. 2021).

2.3.3 Application and evaluation of multilingual sentence embeddings

One main application of multilingual sentence embeddings is cross-lingual similarity search of sentences from monolingual corpora. This task, which can be summarized as finding parallel texts in a source and target language from raw data, is commonly called bitext mining.

2.3.3.1 Bitext mining

There has been a large body of research focusing on bitext mining. The first methods that have been explored are following hierarchical or local mining, where sentence comparisons are only done between sentences from automatically-matched parallel documents. For example, Resnik and Smith (2003) located web pages as potential parallel translations based on an improved version of STRAND algorithm (Resnik 1999) that leverages document structure. Another approach (Fung and Cheung 2004) used an iterative bootstrapping approach for document matching and then parallel sentence extraction. Munteanu and Marcu (2005) and Utiyama and Isahara (2003) performed cross-lingual alignment of news articles in order to then find parallel translations. Bilingual document alignment has been extensively explored, especially in the (Buck and Koehn 2016) shared task. Among the different methods, some were based on n-gram matching, machine translation models or word translation lexicons.

A well-known initiative of hierarchical mining for all European languages is the European ParaCrawl project (Bañón et al. 2020), where parallel data was collected, mining primarily all European languages against English text. This project identifies parallel sentences with different sentence alignments methods. The first one is called Hunalign (Varga et al. 2007) and based on a bilingual dictionary. The second one is called Bleualign (Sennrich and Volk 2010), which first translates non-English sentences into English before matching them with

English sentences with a variant of BLEU score. The last method is called Vecalign (Thompson and Koehn 2019) and based on LASER sentence embeddings.

This Vecalign method was following previous work for cross-lingual similarity estimation based on multilingual embeddings, where sentences are compared using cosine similarity between sentence embeddings. Early works, including (España-Bonet et al. 2017; Guo et al. 2018; Hassan et al. 2018; Y. Yang et al. 2019a), used bilingual sentence embeddings.

The efficient computation of similarity estimation between sentences using sentence embeddings enabled global mining, which compares all possible sentence pairs in large monolingual corpora for several languages. Moreover, the increasing language coverage, from bilingual to massively multilingual sentence embeddings, enabled to create a full matrix of alignments for any language pair. Indeed, massively multilingual embedding spaces allow for the similarity estimation between sentences from any two languages handled (Artetxe and Schwenk 2019b; Feng et al. 2020). In this context, LASER was first used to mine parallel sentences from Wikipedia for 1,620 language pairs with WikiMatrix (Schwenk et al. 2019).

More recently, Schwenk et al. (2021) extended global mining to CommonCrawl monolingual corpora, to introduce CCMatrix, mining billions of sentences from the web with LASER embeddings. CommonCrawl raw text data corresponds to partial snapshots of the internet totalling terabytes of text data extracted from web pages in many languages. The authors used the open-source FAISS library (Johnson et al. 2019) in order to optimize similarity search at scale, where dimensionality reduction and data compression using product quantization (Jegou et al. 2010) were applied on LASER sentence embeddings. CCMatrix was successfully used to train state-of-the-art massively multilingual machine translation models like M2M100 (Fan et al. 2021) and Deepnet (H. Wang et al. 2022). Global mining was recently scaled to 200 languages with the newly trained LASER3 encoders and the mined data was successfully used to train the NLLB state-of-the-art machine translation model (NLLB Team et al. 2022). Finally, another recent bitext mining project is Samanantar (Ramesh et al. 2022), which provides a large publicly available parallel corpus for 11 indic languages, mined using the LaBSE sentence embedding space.

2.3.3.2 Evaluation of sentence embeddings for bitext mining

In order to iterate and improve multilingual sentence embeddings with the goal to perform better bitext mining, one has to come up with an efficient evaluation framework. Indeed, the simplest evaluation idea would be to perform bitext mining for each new sentence embedding space variant and then train MT

systems on the mined bitext data. The performance of the MT model on traditional translation test sets gives a quantitative evaluation of the usefulness of the mined data to train MT systems. Such thorough evaluation comes at the price of an important computational cost, as one has to perform large-scale mining and MT training for each sentence embedding variant that needs to be evaluated.

To overcome this issue, some proxy to bitext mining performance was introduced. In that way, cross-lingual similarity search, also sometimes called `xsim`, is commonly used as an evaluation for multilingual sentence embeddings (Artetxe and Schwenk 2019b; Feng et al. 2020). Given a translation test set $(s_i, t_i)_{1,\dots,N}$ with s_i being a source sentence in a non-English language and t_i its corresponding target English translation, we encode the source and target sentences in the sentence embedding space. For each source sentence encoding s_i^E, we then search the closest target sentence embedding t_j^E, counting an error if it is not the expected one. An error rate on the test set is then reported. Commonly used test sets for cross-lingual similarity search include BUCC (Zweigenbaum et al. 2017), Tatoeba, [1] FLoRes (Goyal et al. 2022) and FLoRes-200 (NLLB Team et al. 2022). The latter is interesting as it is a n-way parallel test set for 200 languages and enables evaluation for a large number of low-resource languages. However, there are only 1k sentences for each language, making the cross-lingual similarity search task easy and error rates for mid- to high-resource languages quickly saturate to 0%.

To overcome this issue, the more challenging xsim++ task was introduced to evaluate more subtle improvements in bitext mining (M. Chen et al. 2023). It augments the FLoRes English target sentences with hard-to-distinguish negative examples for cross-lingual similarity search. These challenging negatives are automatically created with transformations of the original English target sentences, such as causality alternation, entity replacement, and number replacement.

2.3.4 Fixed-size representations for speech utterances

Historically, fixed-size speech representations have been mainly studied at the word level for different specific tasks, ranging from spoken term detection, speech pattern discovery, to speech segmentation into words. Different approaches have been studied to extract a fixed-size representation from speech input. Holzenberger et al. (2018) introduced a method to extract a fixed-size vector from speech inputs using Gaussian downsampling, without the need of any training. Holzenberger et al. (2018) and Y. Chung et al. (2016) introduced an

1. https://tatoeba.org/en/

auto-encoder approach based on recurrent neural networks to extract a fixed-size vector between the encoder and the decoder. Some other works (Settle and Livescu 2016; Riad et al. 2018; Thiolliere et al. 2015) trained Siamese networks with a contrastive loss to build a fixed-size representation. Audhkhasi et al. (2017) studied a keyword search task, building fixed-size representations for audio and words before inputting these representations to a third neural network.

At the sentence level, text-audio sentiment analysis often intrinsically introduces a fixed-size cross-modal representation before classifying the input, as in (K. Yang et al. 2020; Tsai et al. 2019). However, such works did not focus on speech/text alignments, but rather took advantage of information coming from both modalities. Even though being more an utterance-level pre-training method for speech, speech Sim-CLR (Jiang et al. 2020) used pooled representations to perform contrastive learning on augmented speech inputs. Khurana et al. (2020) focused on speech representation learning with speech translation data and a contrastive loss at the sentence level. The model was first evaluated on a retrieval task but not used for large-scale speech translation mining, the speech encoder was rather used for a phone recognition task. Harwath et al. (2018), Merkx et al. (2019), Ilharco et al. (2019), Harwath et al. (2019), and Monfort et al. (2021) built joint speech/visual embedding spaces at the sentence level and were evaluated with a retrieval task.

Following our work on building a multilingual speech/text sentence embedding space (Duquenne et al. 2021) (presented in Chapter 3), Khurana et al. (2022) applied the same teacher-student strategy using LaBSE multilingual text encoder instead of LASER. We may also cite notable more recent works introducing joint representations of audio with other modalities at the sentence level. For example, MuLan (Q. Huang et al. 2022) used contrastive learning on pooled representations of music recordings and text to learn an joint representation of music audio and text while SpeechCLIP (Shih et al. 2023) bridged speech and text representations through images and evaluated zero-shot speech-text retrieval.

2.4 Multilingual and multimodal communication tasks

Multilingual and multimodal communication has been widely studied in machine learning, from text-based machine translation to recent research tackling the translation of speech, like the European live translator (ELITR) project (Bojar et al. 2020) and the recent Seamless Communication project (Seamless Communication et al. 2023b).

In this thesis, we deal with several multimodal communication tasks either as evaluations tasks for our methods which automatically collect speech translation data, or as tasks performed directly with our sentence representations. We first present text-to-text MT and speech recognition tasks, before introducing S2TT and S2ST tasks and concluding with expressive speech generation.

2.4.1 Machine Translation

Machine Translation is one of the most studied machine learning fields and has been particularly popularized with online services like Google Translate. After phrased-based machine translation (Koehn et al. 2003), neural machine translation appeared to be a breakthrough in the field, in particular with the introduction of sequence-to-sequence models (Kalchbrenner and Blunsom 2013; Cho et al. 2014; Sutskever et al. 2014). However, first RNN-based sequence-to-sequence models showed limitations when dealing with long sentence inputs as the decoder takes as input a single vector computed by the encoder: Cho et al. (2014) showed the important decrease in performance of encoder-decoder models for MT with increasing sentence length. To overcome this problem, an attention mechanism between the encoder and decoder in sequence-to-sequence model was introduced (Bahdanau et al. 2014), allowing the decoder to attend to different locations of the input contextual representations during the decoding process and significantly improved performance for long sentences.

The attention mechanism was extended to self-attention, as well as improved cross-attention, and led to another breakthrough in sequence-to-sequence modeling for text, with the Transformer architecture presented in Section 2.1.2. The transformer architecture is used for almost every MT system nowadays.

In this section, we review some challenges of machine translation, like domain generalization or low-resource languages handling, which have been addressed from both data and modeling perspectives. Finally, zero-shot transfer in machine translation between languages has also been widely studied and will be explored in this thesis.

Machine translation training data State-of-the-art MT models are relying on labeled data, also called bitexts, composed of source sentences in one language paired with target sentences in another language. The amount of available labeled translation data is heavily dependant on the language direction. The amount of labeled data decreases quickly for low-resource languages or language directions involving two non-English languages. There have been several efforts to gather parallel sentences in different languages. Some labeled data come from

international organizations like the European Parliament (Koehn 2005) or the United Nations (Ziemski et al. 2016) but are limited to the political domain. Other initiatives provide open-source translations of public texts like OPUS (Tiedemann 2012) or OpenSubtitles (Lison and Tiedemann 2016).

To scale the domain and language coverage of bitext data for MT training, some methods to automatically create or collect parallel text were explored. A first method is bitext mining, as presented in the previous section, and which was used in many state-of-the-art systems like M2M100, DeepNet or NLLB.

Another important approach is called back-translation, which leverages monolingual corpora in the target language to automatically create synthetic bitext data. Existing labeled data is used to train a MT model in the reverse language direction. This model is used to translate monolingual data from the target language to the source language. This synthetic data is used as additional source-to-target bitext data to train a new MT system. Source sentences have been generated by a model and are therefore imperfect and noisy, but have shown not to hurt model performance, but rather significantly improve results when carefully designed. Back-translation was extensively studied at scale by Edunov et al. (2018), and heavily used in projects like NLLB.

Finally, monolingual data can also be leveraged as pre-training data like in mBART (Liu et al. 2020a), where an encoder-decoder model is trained on several languages with a denoising auto-encoding objective.

Massively multilingual modeling Besides bitext mining, data augmentation and pre-training techniques, translation of low-resource languages may be significantly improved with cross-lingual transfer learning between languages in massively multilingual systems (Zoph et al. 2016; Nguyen and Chiang 2017). Similar languages may benefit each other in these massively multilingual settings, as demonstrated in several works, e.g. (Arivazhagan et al. 2019b; Fan et al. 2021). In that context, multilingual models were shown to lead to better translation performance compared to bilingual models. The development of bitext mining at scale for many language pairs also enabled the development of such massively multilingual systems like M2M100, and led to state-of-the-art systems like DeepNet and then NLLB.

As an increased language coverage often coincides with an increased amount of training data, scaling in terms of model size has also been an active area of research: DeepNet scaled MT model size to 200 layers and NLLB used mixture-of-experts techniques to introduce a 54B parameter model.

Zero-shot cross-lingual transfer in machine translation In machine translation, cross-lingual transfer to improve low-resource language directions has been

widely studied. As stated in the previous paragraph, one way to encourage cross-lingual transfer is building a massively multilingual translation system as highlighted in (Fan et al. 2021). Some other works such as (M. Zhang et al. 2022) make an efficient use of MT data involving a pivot language thanks to weight freezing strategies to force representations to be close to the pivot language representations. One extreme scenario of cross-lingual transfer learning is called zero-shot transfer, where one learns to translate one language and directly apply the decoding process to an unseen language. Several methods have been tried to improve zero-shot transfer. Arivazhagan et al. (2019a) and Pham et al. (2019) added language similarity regularization on pooled representations of encoders outputs as an auxiliary loss to a MT objective in order to improve zero-shot transfer. Liao et al. (2021), Vázquez et al. (2018), and Lu et al. (2018) introduced shared weights between language-specific encoders and decoders, commonly called an interlingua that captures language-independent semantic information. Finally, Escolano et al. (2020a), Escolano et al. (2021a), and Escolano et al. (2020b) focused on incremental learning of language-specific encoders and decoders using cross-entropy loss, alternately freezing parts of the model to ensure a shared representation between languages.

Evaluations While there exists many evaluation metrics for MT (Papineni et al. 2002; T. Zhang et al. 2019; Rei et al. 2020; Sellam et al. 2020; Banerjee and Lavie 2005; Popović 2015), we only develop here the few metrics used as evaluations in this thesis.

As stated in Chapter 1, several possible translations exist and are commonly accepted for a given source sentence. This makes the evaluation of MT models difficult as paraphrases of some annotated references may also be good translations. Human evaluation is therefore a good practice to assess the quality of an MT system, but remains an expensive solution that is not scalable to evaluate and compare different models or model variants. In that context, automatic metrics based on labeled references from open-sourced test sets were introduced. The most commonly used automatic metric is called BiLingual Evaluation Understudy (BLEU) (Papineni et al. 2002), a corpus-level metric that computes the geometric mean of n-gram matching precisions between references and generated translations.

However, it has been shown that the BLEU metric correlates poorly with human judgements (Callison-Burch et al. 2006). More recently, Bert-score (T. Zhang et al. 2019) was shown to better correlate with human judgments, creating soft-alignments between contextual representations of references and generated translations and then providing semantic similarity scores based on these alignments.

Finally, COMET (Rei et al. 2020) proposed a neural evaluation metric with improved correlation with human judgments. It finetuned cross-lingual pretrained models on a regression task on labeled quality assessments to mimic human judgements based on source, hypothesis and reference sentences as inputs.

2.4.2 From speech recognition to direct Speech-to-Text Translation

Automatic Speech Recognition Speech recognition has been on of the main task of the speech processing field over the last years. Initial works on ASR used Hidden Markov Model (HMM) (Baum and Petrie 1966; Baker 1975), later coupled with Gaussian Mixture Model (GMM) to approximate probability distribution over vocal states. Neural networks were then integrated with DNN-HMM models (Bourlard and Morgan 1993; Hinton et al. 2012) closing the gap and then outperforming GMM-HMM methods (Mohamed et al. 2009; Dahl et al. 2011). However, neural networks were only a subpart of more complex pipelines. Later, end-to-end methods for speech recognition with RNN architectures were introduced (Graves and Jaitly 2014; Maas et al. 2015) with the Connectionist Temporal Classification (CTC) loss, considering all possible speech-text alignments while optimizing likelihood. Chan et al. (2016) and Bahdanau et al. (2016) studied encoder-decoder approaches for speech recognition and the introduction of the Transformer architecture was quickly used as alternative to RNN-based sequence-to-sequence models for speech recognition (L. Dong et al. 2018). Finally, self-supervised speech pre-training improved the state of the art (Babu et al. 2021; Bapna et al. 2022). and several state-of-the-art models based on encoder-decoder architectures for a large number of languages were introduced (Radford et al. 2023; Seamless Communication et al. 2023a).

Standard evaluation of ASR models is Word Error Rate (WER) computation which accounts for the number of insertion, deletion and replacement in generated transcripts.

Direct speech-to-text translation The parallel development of research on MT and ASR, enabled to introduce cascaded S2TT systems, which use independently trained ASR and MT models to first transcribe audio and then translate it (Stentiford and Steer 1988). Such an approach can leverage state-of-the-art models from both disciplines and benefit from the important data collection efforts for MT and ASR. However, such methods are subject to error propagation in the cascaded pipeline and domain mismatch as models are trained independently. Indeed, MT training data is often coming from domains which are different from

conversational data, and the MT models are not trained on source sentences containing ASR errors.

To overcome these issues, end-to-end S2TT models were explored (Bérard et al. 2016; Bérard et al. 2018) where a model directly translates speech in a source language into text in a target language. These end-to-end models are also called direct S2TT models or more simply S2TT models in opposition to cascaded speech-to-text translation. S2TT has recently been a growing field in machine learning accompanied by the development of S2TT training datasets like Must-C (Di Gangi et al. 2019) which provides more than 385 hours of English speech translated in 8 languages, CoVoST2 (C. Wang et al. 2021b), which provides speech translations in 15 English-to-X (eng-X) directions and 22 X-to-English (X-eng) directions and EuroparlST (Iranzo-Sánchez et al. 2020) which provides 30 different speech translation directions from 6 European languages.

Direct S2TT is also motivated by the compute efficiency of end-to-end methods compared to cascaded systems and recently the translation performance of these end-to-end models is approaching the performance of cascaded systems (Seamless Communication et al. 2023a).

Several techniques like self-supervised pre-training (Babu et al. 2021; X. Li et al. 2020), speech mining (Duquenne et al. 2021), pseudo-labeling (Pino et al. 2020), which is using pre-trained cascaded systems to create synthetic Speech-to-Text (S2T) data, helped obtain these significant improvements. In this context, during the past year, several state-of-the-art S2TT models were introduced. Universal Speech Model (USM) (Zhang et al. 2023), while pre-trained on more than 300 languages, reports S2TT results on 21 CoVoST2 languages. Whisper (Radford et al. 2023) uses weakly-supervised pre-training on 680k hours of speech and can perform both ASR and S2TT into English improving the previous state of the art. The AudioPaLM (Rubenstein et al. 2023) speech-text language model, which combines PaLM text model (Anil et al. 2023) with audioLM modeling (Borsos et al. 2023), can perform multiple speech processing tasks, including S2TT for many languages with new state-of-the-art results. Finally, Seamless Communication et al. (2023a), in which we participated for speech mining efforts, present massively multilingual S2TT results in a framework that unifies MT, ASR, S2TT and S2ST for many languages and defines the current state of the art.

The evaluation of S2TT models is commonly done with BLEU score on S2TT test sets.

Zero-shot transfer in Speech Translation To bridge the gap with cascaded systems, some work integrated MT data into the training, with the goal of cross-modal transfer from the text-to-text translation task to the speech-to-text translation task. Several works added MT data in S2TT training, using an auxiliary

loss to bridge the modality gap, like adversarial (Alinejad and Sarkar 2020), and distance (Q. Dong et al. 2021; Liu et al. 2020b) regularization. (Xu et al. 2021; X. Li et al. 2020) used adaptor modules to address the length mismatch between audio and text representations. Several works studied how to efficiently perform zero-shot cross-modal transfer from text to speech in the frame of direct speech translation. Following (Escolano et al. 2020a; Escolano et al. 2021a; Escolano et al. 2020b) presented previously for text, Escolano et al. (2021b) learned a speech encoder compatible with decoders trained on text only, freezing the text decoder during training and using cross-entropy on the output of the decoder. Other works such as (Dinh et al. 2022; Dinh 2021) studied zero-shot speech translation employing a cross-modal similarity regularization as an auxiliary loss. However, they obtained low zero-shot results possibly due to the mismatch in the encoder output lengths between speech and text.

2.4.3 Direct Speech-to-Speech translation

Generating speech as output a of machine learning model has first been explored through TTS synthesis. Recent neural methods like (N. Li et al. 2019; Ren et al. 2020) rely on phoneme form of the input text and are trained to predict target waveforms or mel-spectrograms. A spectrogram vocoder can be trained to obtain a speech waveform. TTS was recently scaled to more than 1000 languages with MMS (Pratap et al. 2023).

Following research on TTS, S2ST started from cascaded systems consisting of successive ASR, MT and TTS synthesis (Nakamura et al. 2006; Do et al. 2015). The reliance on intermediate text outputs poses limitations on cascaded models to support efficient inference and unwritten languages. Given these challenges, there has been a recent surge of research interest in direct approaches to speech translation without the need of texts. Jia et al. (2019b) presented Translatotron, the first direct S2ST model that directly predict spectrograms of output speech. This first model, lagging behind cascaded systems, was later improved with a two-pass decoding method to present Translatotron2 (Jia et al. 2022).

Another line of research on S2ST uses discrete units of speech. Tjandra et al. (2019) first introduced a encoder-decoder model which translates discrete units from source language to discrete units from a target language. Lee et al. (2022a) presented an S2ST model architecture that uses HuBERT units (from k-means clustering of some intermediate layer representations as extracted for HuBERT pre-training, see Section 2.2.2) as targets of Speech-to-unit (S2U) model. Similarly to spectrogram-based vocoders, a unit-based vocoder is introduced based on Hi-Fi GAN training to produce speech waveforms from discrete units (Polyak et al. 2021).

Despite this progress on direct S2ST, it is faced with the challenge of data scarcity of aligned speech with speech in different languages and has often been using synthetic speech to overcome this issue (Post et al. 2013).

S2ST systems are commonly evaluated with ASR-BLEU metric, which relies on pre-trained open-sourced ASR models to automatically transcribe the generated speech before computing BLEU score with text references on a S2TT test set. This evalutation metric is imperfect as it heavily depends on the quality of the ASR model used.

2.4.4 From controllable text-to-speech to expressive speech generation and translation

Expressivity of speech is central in human communication, as a message is often not only conveyed by content but also other expressive speech characteristics.

Text-to-speech synthesis has recently made great progress in generating natural-sounding speech (Kim et al. 2021) and many research works are focusing on controllable methods for output vocal style. Some work learned TTS models conditioned on some specific speech properties like pitch and energy (Ren et al. 2020). Others used pre-trained speaker style embeddings (Jia et al. 2018; Casanova et al. 2022) to condition the speech synthesis. Finally, another line of work learns low-dimensional residual embeddings or style tokens to extract the required output speech properties (Wang et al. 2018; Akuzawa et al. 2018; Hsu et al. 2018). These representations may represent the required vocal style of the output speech, and relatedly other works like Prosody2Vec learned disentangled prosody representations using speech reconstruction (Qu et al. 2023).

More recently, high-fidelity neural audio codecs like Soundstream (Zeghidour et al. 2021) and Encodec (Défossez et al. 2022) were introduced. These neural audio codecs are learned with an encoder-decoder model and provide residual quantized acoustic representations of speech from coarse-to-fine learned codebooks (sets of learned discrete units). Some recent works like Vall-E (Wang et al. 2023a) or Spear-TTS (Kharitonov et al. 2023) address TTS as a language modeling task on these new acoustic tokens. Such systems are able to preserve the acoustic environment and vocal style in output speech using prompting. Finally, leveraging representations from neural audio codec, and training diffusion models, NaturalSpeech2 (Shen et al. 2023) can perform singing synthesis. TTS across languages is even made possible in a zero-shot way in Voicebox (Le et al. 2023) with flow-matching.

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, followed by vocal style-preserving Soundstream (Zeghidour et al. 2021) units. Such audio language modeling techniques can be extended to perform speech processing tasks such as TTS, like in AudioPaLM (Rubenstein et al. 2023).

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. Translatotron 2 (Jia et al. 2022) was trained on vocal style aligned speech generated with a vocal style preserving TTS model. AudioPaLM model also provides vocal style preservation for S2ST with audio prompting. Relatedly, Polyvoice (Q. Dong et al. 2023) used two language models: one for translation and one for speech synthesis. Similarly in (Wang et al. 2023b), speech-to-speech translation with vocal style preservation leveraged acoustic Soundstream tokens.

CHAPTER 3

EMBEDDING SPEECH/TEXT SENTENCES AND MULTILINGUAL SPEECH MINING

Chapter abstract

As presented in Chapter 2, several multilingual sentence embedding spaces were developed to address cross-lingual similarity search as well as semantic comparison of sentences from different languages. However, these methods were limited to the text modality. In this chapter, we present an approach to encode a speech signal into a fixed-size representation which minimizes the cosine loss with the existing massively multilingual LASER text embedding space, to introduce a multilingual speech/text sentence embedding space. Sentences are close in this embedding space, independently of their language and modality, either text or audio. Using a similarity metric in this multimodal embedding space, we introduce speech mining, where we automatically align sentences from different languages and modalities as potential translations (either speech-to-text or speech-to-speech translations). To evaluate the automatically mined speech translation corpora, we train neural speech translation systems and demonstrate that adding mined data to the training can significantly improve BLEU score performance of speech translation systems.

The work in this section has led to two publications: [1]

- *Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk (2021). "Multimodal and multilingual embeddings for large-scale speech mining". In:* Advances in Neural Information Processing Systems *34*
- *Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswami, Changhan Wang, Juan Pino, Benoit Sagot, and Holger Schwenk (July 2023a). "SpeechMatrix: A Large-Scale Mined Corpus of Multilingual Speech-to-Speech Translations". In:* Proceedings of the 61st Annual Meeting of the Association for

1. This chapter is adapted from these two publications

Computational Linguistics (Volume 1: Long Papers). *Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 16251–16269.* URL: https://aclanthology.org/2023.acl-long.899

Contents

3.1 Introduction

While self-supervised pre-training methods have a growing importance in Natural Language Processing (NLP) (see Section 2.2) and while there is promising research on unsupervised Machine Translation (MT) (Lample et al. 2018; Artetxe et al. 2017), labeled data is still required to achieve best performance in MT. Labeled data scarcity for language directions involving low-resource languages may lead to strongly imbalanced training data in multilingual systems and poor translation performance. To overcome this issue and scale machine translation to hundred languages (Fan et al. 2021), bitext mining was introduced to automatically collect parallel data (see Section 2.3.3).

Back in 2021, the available amount of Speech-to-Text Translation (S2TT) training data was limited. Speech-to-Speech Translation (S2ST) training data was scarcer, even for some high resource languages. This is still true today for mid-resource and low-resource languages.

To overcome this training data scarcity in speech translation, we explore in this chapter speech mining as an extension of bitext mining to the speech modality. In order to perform speech mining, we introduce a multilingual speech/text sentence embedding space, with the required semantic properties: two sentences with similar meaning are closely encoded in the embedding space, independently of their language or their modality (either speech or text).

Based on this multilingual and multimodal sentence embedding space, we present and perform speech mining on several raw text and audio corpora,

and automatically align speech-text pairs and speech-speech pairs in different languages. Finally, we train S2TT and S2ST translation systems in order to validate the quality and the usefulness of the mined data.

This chapter is divided into two parts. The first part introduces a multilingual speech/text sentence embedding space based on LASER and presents speech mining. The second part scales speech-to-speech mining to build SpeechMatrix a corpus of mined Speech-to-Speech (S2S) translations for 136 spoken language pairs.

3.2 Multimodal and multilingual embeddings for speech mining

In this part, we introduce how we build a multilingual speech/text sentence embedding space with the desired semantic properties illustrated in Figure 3.1.

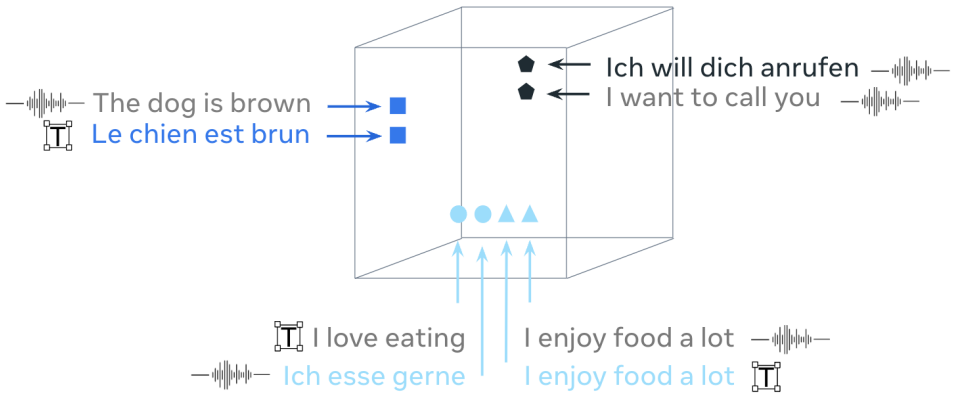

Figure 3.1. – **Illustration of a multilingual and multimodal sentence embedding space.** Paraphrases, transcriptions, spoken and written translations are closely encoded in the sentence embedding space. *Extended the original figure from Meta blog on the LASER sentence embedding space.*

3.2.1 Speech encoder training

Training a multimodal audio/text fixed-size embedding space could be motivated by research on training Siamese networks with a contrastive loss, e.g. (Feng et al. 2020). Instead of two text encoders, one would use one speech encoder and one text encoder. However, since both encoders are trained from scratch, such a procedure would probably require a large amount of labeled multimodal

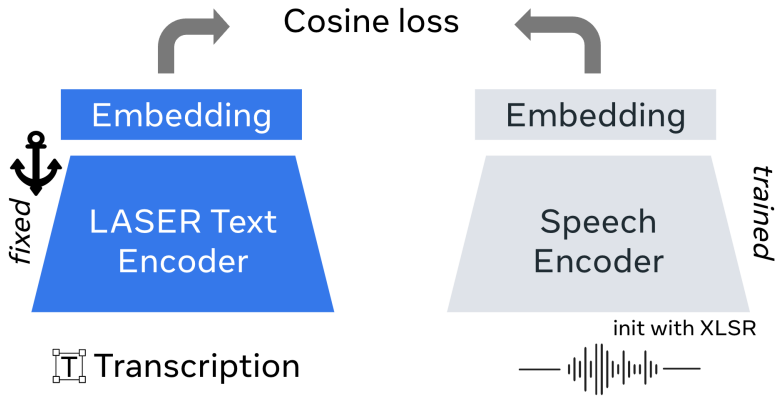

Figure 3.2. – **Architecture of the proposed teacher-student approach.** We train a speech encoder to minimize the cosine loss between speech sentence embeddings and the existing LASER text sentence embeddings.

training data. Instead, we apply a teacher-student training framework (which was already explored for text, see Section 2.3.2): we use an existing text encoder as teacher and train an audio encoder to minimize the cosine loss between the two encoder outputs. This architecture is summarized in Figure 3.2. It can be trained with two types of labeled data:

- **Speech transcriptions**: both encoders use input in the same language, but differ in the modality;
- **Speech-to-text translations**: we can also minimize the cosine loss of the speech embedding with respect to its written translation in one of the languages supported by the text encoder.

Concretely, we use the multilingual LASER sentence encoder which is freely available[2] and which was successfully used in large-scale bitext mining approaches (cf. Section 2.3.3). A thorough comparison with other multilingual sentence encoders is left for future research, in particular LaBSE (Feng et al. 2020). The LASER encoder is fixed during teacher-student training.

Our speech encoder is based on the Wav2vec2 architecture and its weights are initialized with the XLS-R model:[3] we use the XLS-R frozen pre-trained feature encoder and fine-tune the weights of XLS-R transformer encoder to obtain our fixed-size audio representation. During fine-tuning, the feature encoder representations are masked with a strategy similar to SpecAugment (Park et al. 2019) as introduced in Wav2vec2 paper.

2. https://github.com/facebookresearch/LASER
3. https://github.com/pytorch/fairseq/tree/master/examples/wav2vec

We explored several methods to get a fixed-size speech representation, like max- or mean-pooling of the encoder outputs. Best performance is obtained using the output of the transformer encoder corresponding to a particular Beginning-Of-Sentence (BOS) vector. This BOS vector is simply a vector filled with ones (1.0), added at the beginning of the feature sequence in Wav2vec2 architecture. This method is inspired by BERT (Devlin et al. 2019) sentence representations for text, that are often extracted using a Classification (CLS) token at the beginning of the input sentence.

3.2.1.1 Encoder evaluation

In order to compare different versions of our speech encoder, an evaluation framework is needed. The ultimate goal is to mine speech translations and to show improvements when training S2TT or S2ST systems. However, this is computationally rather expensive and we only apply it for some selected encoders (see Section 3.2.2). To evaluate a standalone speech encoder, we propose to use cross-modal similarity search, extending the `xsim` evaluation metric presented in Section 2.3.3 for cross-lingual similarity search. Given a multimodal test set $(a_i, t_i)_{1,\dots,N}$ with a_i being the audio file and t_i its corresponding text, we encode the speech and texts. We normalize the texts by removing quotes encapsulating whole sentences and lower casing them. For each speech encoding a_i^E, we then search the closest text embedding t_j^E, counting an error if it is not the expected one. We adopt the margin based similarity proposed by Artetxe and Schwenk (2019a) which was reported to outperform simple cosine similarity. The margin $\mathrm{sim}(x, y)$ between two embeddings x and y is defined as the difference between the cosine similarity between x and y, and the average cosine similarity of its nearest neighbors in both directions:

$$\mathrm{sim}(x, y) = \cos(x, y) - \left(\sum_{z \in \mathrm{NN}_k(x)} \frac{\cos(x, z)}{2k} + \sum_{z \in \mathrm{NN}_k(y)} \frac{\cos(y, z)}{2k} \right) \tag{3.1}$$

where $NN_k(x)$ are the nearest neighbors of x. We use the Dev and Test set of the CoVoST2 corpus (C. Wang et al. 2021b) which statistics are summarized in Table 3.1.

3.2.1.2 Single multilingual speech encoder

All our speech encoders are trained and evaluated on the CoVoST2 dataset (see Table 3.1) released under CC0 license. CoVoST2 is a large-scale multilingual speech translation corpus based on Common Voice (Ardila et al. 2020). In this work, we focus on five spoken languages: English (eng), German (deu),

Table 3.1. – **Statistics of CoVoST2 S2TT corpus.** Number of hours of speech data from CoVoST2 used to train and evaluate the speech encoders.

	eng			deu			spa			fra			rus		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Audio [hours]	430	26	25	184	21	22	113	22	23	264	22	23	18	10	11
#sentences	289k	16k	16k	128k	14k	14k	79k	13k	13k	207k	15k	15k	12k	6k	6k

French (fra), Spanish (spa) and Russian (rus). For each audio input language, we explore different textual training targets, namely the transcriptions encoded with LASER and the English translation encoded with LASER. We use the German translations as a teacher for English speech data. We call them respectively, the *"transcription teacher"* and the *"translation teacher"*. We also train on both, i.e. using transcriptions and translations as teachers.

In this section, we first train one multilingual speech encoder for all five languages. To handle unbalanced training data between languages, speech sentences are sampled according to a multinomial distribution with probabilities $\{q_i\}_{i=1,..,N}$ (as done for multilingual text in (Conneau and Lample 2019)):

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \quad \text{with} \quad p_j = \frac{n_j}{\sum_{k=1}^{N} n_k} \tag{3.2}$$

In the following experiments, we use $\alpha = 0.2$. For all training methods, we take the checkpoint with the lowest validation loss. The learning rate to fine-tune XLS-R transformer is set to 10^{-4}, and training was performed on 24 Tesla V100 Graphics Processing Unit (GPU)s.

Our multimodal similarity search results are summarized in Table 3.2. In the top block of results, we first report the multimodal similarity error rates between the speech embeddings and the text embeddings of the human transcriptions, for the three training methods. The error is below 1% for German, French and Spanish when the training and evaluation criterion are the same (row A.1). The performance on Russian is significantly worse: about 25% error rate, probably due to the small amount of training data. Not surprisingly, the error rates are higher when using the embeddings of translations into English as targets (row A.2), but also when using both (row A.3).

We then switch to similarity search of speech against the translation into English (block B). These results are relevant to our use case of speech mining (see Section 3.2.2). Overall, the error rates are about twice as high. Surprisingly, performance is slightly better when using transcriptions as the teacher (row B.1) than translations (row B.2), although this corresponds to the evaluation criteria.

Table 3.2. – **Similarity search results for a multilingual speech encoder.** Error rates for the different training methods with a multilingual speech encoder on CoVoST2 test set.

	Teacher mode	eng	deu	fra	spa	rus
A) Search audio against transcriptions						
A.1	Transcriptions	2.70	1.03	0.79	0.57	25.63
A.2	Translations	3.25	1.93	1.40	0.89	28.32
A.3	Both	3.01	1.21	0.91	0.64	36.19
B) Search audio against translations (eng)						
B.1	Transcriptions	-	3.58	2.31	1.79	30.46
B.2	Translations	-	4.06	2.57	1.88	31.65
B.3	Both	-	3.36	2.05	1.66	40.54
C) Search transcriptions against translations (eng)						
	n/a	-	1.96	0.97	1.00	1.05

Best results are obtained when using both (row B.3). Finally, as lower bound, we calculate the similarity error between the speech transcriptions and their translations (block C). Both source and target are sentences, i.e. no audio encoder is used. The error rates are about 1% for French, Spanish and Russian, and 2% for German. Compared to these numbers, the performance of our multilingual speech encoder (row B.3) seems to be very good (with exception of Russian).

3.2.1.3 Separate speech encoders per language

We now switch to separate speech encoders, trained with both translations and transcriptions as teachers. In Table 3.3, we observe a huge improvement for the Russian speech encoder: the error rate against the English translations went down from more than 30% to 6.9%.

Table 3.3. – **Similarity search results for monolingual speech encoders.** Error rates for separate speech encoders trained with transcription + translation teachers on CoVoST2 test set.

	eng	deu	fra	spa	rus
Search audio against transcriptions	2.72	1.44	0.84	0.57	3.46
Search audio against translations (en)	-	3.73	1.86	1.78	6.86
Search transcriptions against translations (eng)	-	1.96	0.97	1.00	1.05

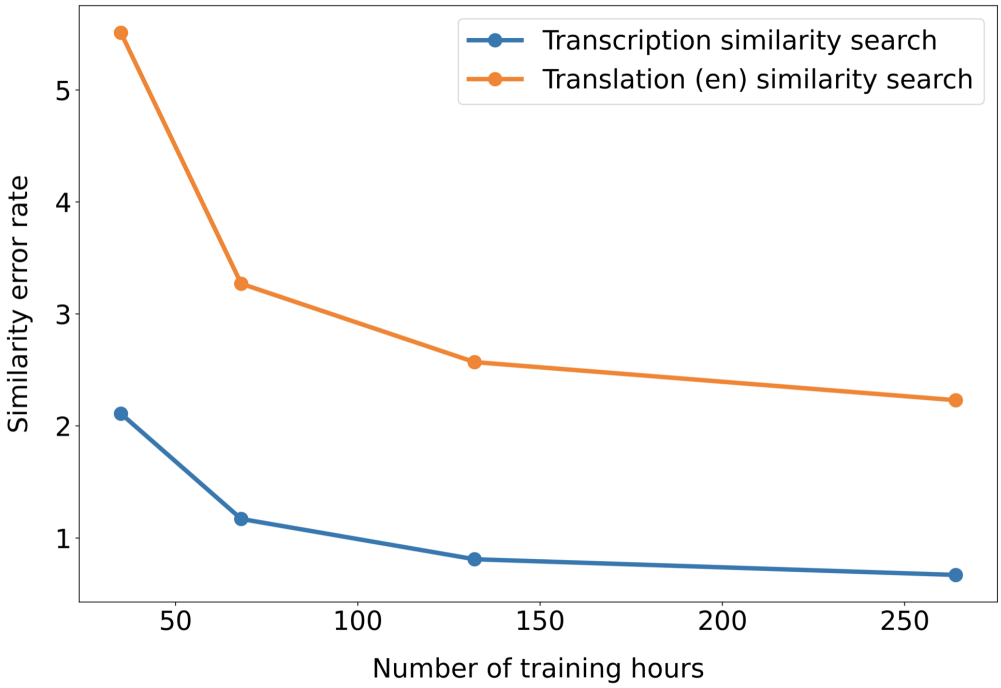

Figure 3.3. – **Similarity search error rates vs. training data size.** We evaluate similarity search error rates for French speech encoders trained with a transcription teacher for different training data sizes.

The error rate for French decreased from 2.05% to 1.86%, but those for Spanish and German are slightly higher than with the multilingual speech encoder. Our experimental evidence seems to indicate that low-resource languages, e.g. Russian in our study, which have no similarity with other trained languages are better handled by an individual speech encoder.

We also studied the quality of the speech encoder in function of the training data size (transcriptions only, Fig 3.3). As expected, more data gives better performance, but the curve flattens out quickly and good performance is already achieved for relatively small amounts of training data.

3.2.2 Speech mining

We now apply the encoder to mine unlabeled raw audio against huge collections of texts. We could mine either for transcriptions of the audio input, i.e. texts in the same language, or translations into another language. The encoders are trained to map sentences with similar meaning close in the embedding space, independently from the language and the modality. This can include paraphrases, that is, sentences which express the same meaning but with a different wording.

Therefore, we argue that adding automatically mined speech transcriptions is unlikely to improve a speech recognition system, since it requires training data which are exact word-by-word transcriptions of the speech. Speech-to-text translations systems however can benefit of paraphrased output, instead of strict word-by-word translations. We therefore focus on mining speech translations.

3.2.2.1 Speech translation mining

We used Librivox as our set of unlabeled speech data. Librivox is a repository of open domain audio books in different languages. [4] We focus on German, French, Spanish and English audio books. Data statistics are reported in Table 3.4. The very limited amount of Russian raw speech data in Librivox prevented us to perform speech translation mining for this language.

Table 3.4. – **Librivox data statistics.** Number of audio books and number of hours of raw speech data from Librivox for our languages of focus.

	deu	fra	spa	eng
#audio books	633	257	343	13 292
#hours	3 529	1 535	1 770	73 511

As English texts, we use five snapshots from Common Crawl as processed in CCNet (Wenzek et al. 2019). The texts come in paragraphs which we segment into sentences and deduplicate them. This yielded about 15 billion sentences. We also mined the English audio against six text languages, chosen to cover various linguistic features, namely Arabic (arb), French (fra), Spanish (spa), Russian (rus), Turkish (tur) and Vietnamese (vie). We used the same 32 Common Crawl snapshots as in (Schwenk et al. 2021). The amount of sentences varies from 786 million (Arabic) to 5684 million (French). The same procedure could be also applied to all 80 languages supported by the LASER encoder.

Audio segmentation is a key element to obtain high recall in speech mining since texts in Common Crawl are usually sentences. Librivox audio books are separated into different chapters, but speech data is not segmented into sentences. Voice Activity Detection (VAD) is commonly used to segment audio, as it was done to generate LibriLight (Kahn et al. 2019) or Multilingual LibriSpeech (MLS) (Pratap et al. 2020). However, those audio segments are not guaranteed to be real sentences. On one hand, it cannot be excluded that (multiple) silences appear within a sentence. And on the other hand, several sentences may follow each other without any silence in between them.

4. https://librivox.org/api/

In this part, we propose to first generate multiple plausible speech segmentations, and let the mining algorithm decide which ones are best aligned with the existing texts. Eventually, we filter the mined speech/text alignments to

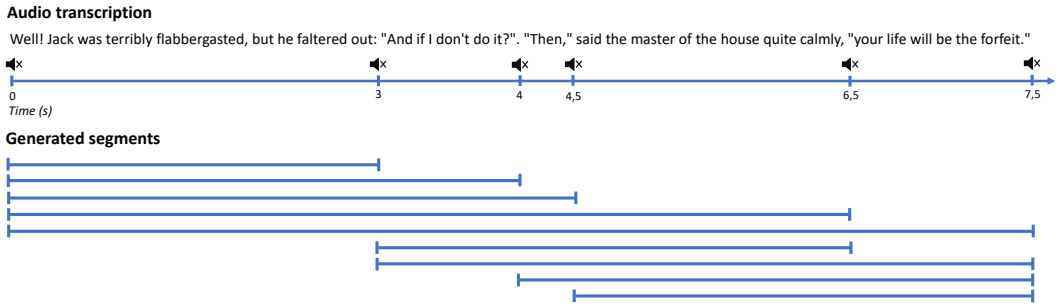

Figure 3.4. – **Example of generated segments by our segmentation.** The transcription is: *Well Jack was terribly flabbergasted,* | *but he faltered out:* | *"And if I don't do it?".* | *"Then,"* | *said the master of the house quite calmly,* | *"your life will be forfeit.".*

exclude overlapping segments. For each long audio input, theoretically containing several sentences, we run VAD with Flashlight [5] pre-trained models. This generates several detected silences in the audio. Based on these detected silences, we segment the audio into several parts under the following rules: a segment boundary is defined by two silence timestamps, a segment should be at least 3 seconds long and at most 20 seconds long. An example of this segmentation procedure is given in Figure 3.4.

Mining algorithm. Once all audio segments and texts are encoded, we can apply mining procedures which were successfully applied to large-scale text-to-text mining, in particular CCMatrix (Schwenk et al. 2021). It has been observed that an absolute threshold on the cosine distance is globally inconsistent (Guo et al. 2018). Therefore, we apply a margin criterion as for similarity search evaluation (see Section 3.2.1). We use the $k = 16$ nearest neighbors to calculate the average distance for both directions. To make mining efficient at this scale, in particular when searching in fifteen billion English sentences, a compact representation and fast search is needed. The open-source FAISS library [6] for fast index search was used for this (Johnson et al. 2019), as in several other large-scale text and image mining projects.

Post-processing. The mining algorithm can align several sentences to the same speech segment, and vice-versa. We remove all these duplicates keeping those with highest alignment score. Our segmentation algorithm generates multiple candidates for each speech segment (see Figure 3.4 above). In Table 3.5 we report

5. https://github.com/flashlight/wav2letter/
6. https://github.com/facebookresearch/faiss/wiki/Faiss-indexes

the total size of aligned speech segments (row "Sum"), and counting overlapping sub-segments only once (row "Union").

Table 3.5. – **Speech-to-text mined data statistics.** Number of hours of speech for Speech-to-Text (S2T) mined data, either counting the sum and union of durations or the post-processed duration.

	deu-eng	fra-eng	spa-eng	eng-spa	eng-fra	eng-rus	eng-arb	eng-tur	eng-vie
Sum	2 247	933	1 391	8088	8 327	3 959	1 718	1 851	1 527
Union	1 296	630	798	6825	7 080	3 529	1 606	1 721	1 437
Post-processed	1 074	543	668	6 289	6 544	3 330	1 549	1 656	1 390

We further post-process the mined data, in order to get speech segments without overlaps. We sort alignments by decreasing order of similarity scores. Then, following this decreasing order, we successively select alignments involving speech segments if they are not overlapping with the previously selected segments. In particular, this post-processing ensures that similar speech segments (for example a speech segment corresponding to a full sentence, and another segment corresponding to the same sentence with an additional word at the end), are not selected twice, and that only the best matching pair is selected. It should be pointed out that we were able to align a significant percentage of the available speech, e.g. 1074 hours out of 3529 hours of raw German speech (see Tables 3.4 and 3.5).

3.2.3 Speech-to-text translation with mined data

Finally, we train S2TT systems on the mined speech data. The evaluation of these S2TT trained models will assess the usefulness of the mined data to improve S2TT models as well as the quality of the speech-text alignments.

Evaluation of mined X-eng data. We use the well established CoVoST2 task, following the train-test splits in (C. Wang et al. 2021b). See Table 3.1 for corpus statistics.

The 2021 best performing S2TT approach, named LNA (X. Li et al. 2020) builds on extensively pretrained models: a Wav2vec2 speech encoder (Baevski et al. 2020) and a mBART model (Liu et al. 2020a) as the text decoder. mBART is first pre-trained on monolingual text data from 100 Common Crawl snapshots, and then trained on the parallel texts from OPUS (Tiedemann 2012). X. Li et al. (2020) jointly trained an LNA S2TT system on multiple languages to enhance performance via cross-lingual transfer. Another strong multilingual S2TT system, E2E S2T, was proposed in (C. Wang et al. 2021b) for evaluation on the CoVoST2 dataset. It has an encoder-decoder architecture trained end-to-end. We also report the results

of a cascaded model (Iranzo-Sánchez et al. 2020): the audio is first transcribed as texts, and then translated into the target language with a machine translation model.

We follow exactly the procedure of the LNA approach, but train separate models for each language pair to independently evaluate the quality of each mined speech/text corpus. We tune layer norm and multi-head attention parameters on the train set in each language direction, while other model parameters are frozen during the fine-tuning stage. The BLEU scores on the CoVoST2 test set are reported in Table 3.6. Our baseline bilingual LNA model is on-par with the best performing bilingual models reported in the literature in 2021. In this table, mined data is used and selected with a threshold of $t = 1.07$, following an ablation study on this threshold selection.

Indeed, one hyperparameter in our speech translation mining process is the threshold on the alignment scores. Mined speech-text pairs are kept and considered as translations if their alignment scores are greater than or equal to the threshold. Speech translation models are trained on the combination of CoVoST2 train set and mined data at different thresholds. We report the performance of each model on the dev set of CoVoST2 in Figure 3.5, and find the optimal value for the threshold.

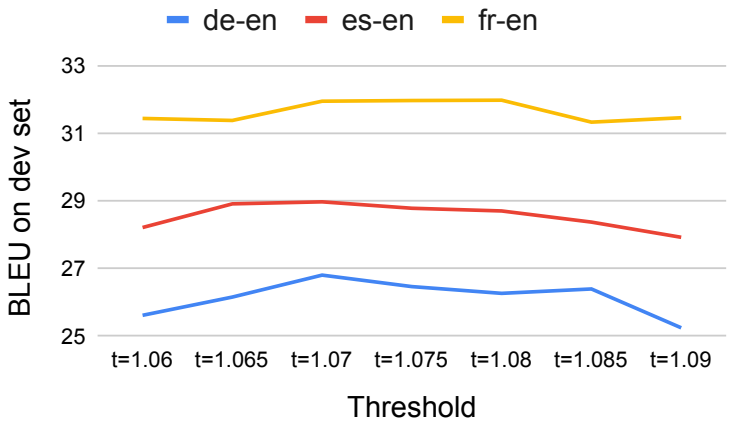

Figure 3.5. – **Speech-to-text translation evaluation of mined data at different thresholds.** BLEU on dev set achieved by S2TT models trained on CoVoST2 train set + mined data at different thresholds.

Based on Figure 3.5, the optimal threshold is $t = 1.07$ for deu-eng, spa-eng and fra-eng directions. As we decrease the threshold, more mined data is added to the train set improving model performance. When decreasing the threshold below $t = 1.07$, the data quality decreases too: despite the larger training data size, the

translation performance decreases due to more noise. The threshold is $t = 1.07$ is used for all experiments with mined data.

Table 3.6. – **Speech-to-text evaluation of X-eng mined data.** BLEU scores of S2TT on CoVoST2 test set.

Approach	Train	deu-eng	spa-eng	fra-eng
Bilingual models:				
Cascaded S2T	CoVoST2	23.2	31.1	29.1
E2E S2T	CoVoST2	17.1	23.0	26.3
LNA (ours)	CoVoST2	24.4	29.9	30.7
LNA (ours)	CoVoST2 + **mined**	**26.4**	**31.6**	**32.0**
Multilingual models:				
E2E S2T	CoVoST2	18.9	28.0	27.0
LNA	CoVoST2	28.2	35.2	35.0

In Table 3.6, we notice that adding the mined data brings significant improvements of more than 1.3 BLEU in average for all language pairs. Please note that our mined data is generic and not selected to match the domain of the CoVoST2 task. The results of the multilingual models are not directly comparable with ours since they benefit from knowledge transfer across languages. We provide them here for the completeness of empirical results.

Evaluation of mined eng-X data. We further evaluate the quality mined data in eng-X directions (for a threshold of 1.07, not particularly optimized for eng-X directions) using Must-C dataset (Di Gangi et al. 2019) and S2T Transformer (C. Wang et al. 2020b), considering the established baseline results of S2T Transformer on Must-C.

S2T Transformer used in this work has 6 encoder layers and 6 decoder layers with 4 attention heads. The feedforward dimensions are 2048 and 256. Following the empirical setup in (C. Wang et al. 2020b), S2T Transformer is first trained on Must-C Automatic Speech Recognition (ASR) data in order to initialize its encoder parameters. Then the model is trained with Must-C speech translation data in a given language direction, which serves as a baseline in our experiments.

We augment the train set of speech translations with the mined data. With the encoder initialized with ASR training, S2T Transformer is trained on the combination of Must-C and mined data for the task of speech translation for 200k steps and finetuned on Must-C data only for 100k steps. Table 3.7 reports BLEU scores of models trained with and without mined data in six eng-X language directions.

Table 3.7. – **Speech-to-text evaluation of eng-X mined data.** BLEU scores of S2TT on the Must-C test set.

Train data	eng-spa	eng-fra	eng-rur	eng-arb	eng-tur	eng-vie
MuST-C	27.2	32.9	15.3	12.3	9.7	21.4
MuST-C + Mined (t=1.07)	28.7	34.4	16.1	12.8	10.5	21.8

As is shown in Table 3.7, mined data brings improvements in the BLEU score of 1.5, 1.5, 0.8, 0.5, 0.8 and 0.4, in speech translation from English to Spanish, French, Russian, Arabic, Turkish and Vietnamese respectively. The performance improvements again demonstrate that mined speech-to-text data is of good quality and useful for model training.

3.2.4 Speech-to-speech translation with mined data

The LASER teacher text encoder and all student speech encoder are mutually compatible. This enables us to perform speech-to-speech mining directly in the embedding space without the need to transcribe or translate. We use the same speech embeddings from Librivox as in Section 3.2.2.1 and mine for all pairs of German, Spanish, French and English speech.

Table 3.8. – **Speech-to-speech mined data statistics.** Numbers of hours for source and target languages of S2S mined data, either counting the sum and union of durations or the post-processed duration.

	spa-deu	fra-deu	fra-spa	eng-spa	deu-eng	eng-fra
Sum	64 / 65	52 / 59	235 / 259	732 / 936	557 / 821	1 049 / 1 210
Union	45 / 47	37 / 43	121 / 133	488 / 486	373 / 421	562 / 518
Post-processed	40 / 41	33 / 38	101 / 111	425 / 442	324 / 363	470 / 447

The amount of automatically mined speech-to-speech alignments are given in Table 3.8, using the same post-processing as for speech-to-text mining (applied on the source speech). Overall, we provide a speech-to-speech corpus of 1393 hours in six language pairs. This should be put into context with the 2021 best practice in S2ST, which was mostly based on the Fisher Spanish-to-English speech corpus (Post et al. 2013) of 160 hours of source speech, the target speech being artificially created by speech synthesis (Jia et al. 2019b).

We first provide an initial human analysis of these alignments highlighting the quality of the S2S alignments. We randomly sampled one hundred S2S alignments

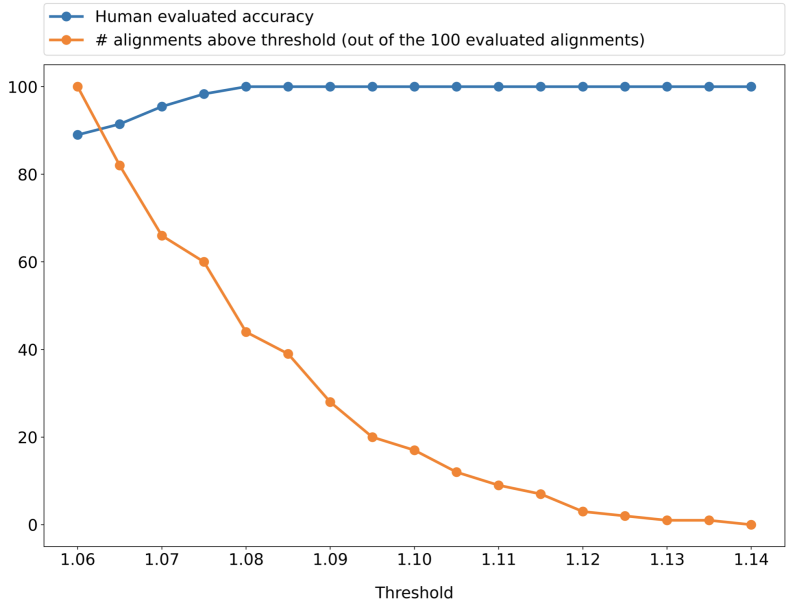

Figure 3.6. – **Human evaluation of S2S mined data.** Human evaluation of 100, randomly sampled, S2S alignments for the fra-spa pair.

for the French-Spanish pair with alignment scores above $t = 1.06$. We manually checked the alignments and reported the accuracy we obtained in Figure 3.6. A threshold of $t = 1.06$ corresponds to the one hundred evaluated alignments. We also report the results when removing the alignments with a score below a given threshold (varying from 1.06 to 1.14). A higher threshold gives a better accuracy but less alignments.

To better evaluate the quality of the mined data, we did a research collaboration with the Universal Speech Translation team from Meta to train S2ST systems on the mined S2S data. The method is introduced in (Lee et al. 2022b), where the HuBERT normalized units are extracted from the target speech and used as targets for a Speech-to-Unit (S2U) model. The speech normalization procedure helps handling real speech as target, which is exactly the case when dealing with S2S mined data. The mined data was added to an existing training set of real S2ST data from VoxPopuli labeled set (C. Wang et al. 2021a). We focused on Spanish-to-English and French-to-English translation directions where VoxPopuli training data totals approximately 500h of source speech for each direction. We report the ASR-BLEU results on CoVoST2 test set of S2ST systems trained on VoxPopuli training data only and VoxPopuli training data combined with the mined data. We notice in Table 3.9, that adding mined speech-to-speech translation data to the training significantly boost the performance of the S2ST system, with more than 50% ASR-BLEU gain on CoVoST2.

Table 3.9. – **Speech-to-speech evaluation of mined data.** ASR-BLEU scores of S2ST on CoVoST2 test set.

	spa-eng	fra-eng
Speech-to-speech translation		
trained on Voxpopuli	9.2	9.6
trained on Voxpopuli + mined data	15.1	15.9

3.3 SpeechMatrix: scaling speech-to-speech translation mining

In the first part, we presented different ablations on the training of a multilingual speech/text sentence embedding space and introduced speech mining. Speech translation evaluations highlighted the quality and usefulness of the automatically aligned data for a few languages. In this part, we develop the work we have done to scale S2S mining to 17 spoken languages from the European Parliament: Czech (ces), German (deu), English (eng), Estonian (est), Finnish (fin), French (fra), Croatian (hrv), Hungarian (hun), Italian (ita), Lithuanian (lit), Dutch (nld), Polish (pol), Portuguese (por), Romanian (ron), Slovak (slk) and Slovenian (slv). In this work, our contribution mainly lies on the speech mining part while the S2ST model training on the mined data is a research collaboration with the Universal Speech Translation team from Meta.

3.3.1 Speech-to-speech mining

We follow the speech-to-speech mining approach presented in Section 3.2.2.1, but train new speech encoders covering our 17 languages of focus. We evaluate these new encoders and performed mining on raw speech recordings from the European Parliament.

3.3.1.1 Speech encoders

We follow the teacher-student approach introduced in Section 3.2.1 and train speech encoders with the supervision of the multilingual LASER text encoder. Contrarily to the work presented in the first part, we scale all aspects of speech encoder training to boost the end performance of these models. Both transcriptions and written translation of the audio utterances are encoded with LASER text encoder as target vectors for speech encoder training. Speech encoders are initialized with the 2B-parameter XLS-R model (Babu et al. 2021), which was

pre-trained on nearly half a million hours of publicly available audios in 128 languages. The fixed-size representation for speech is obtained with max pooling of the encoder outputs which appeared to work better compared to other pooling methods on these large XLS-R variants.

We use various publicly available ASR data sets which cover our languages to train the speech encoders, including CoVoST2 (C. Wang et al. 2020a; C. Wang et al. 2021b), Common Voice (Ardila et al. 2020), EuroparlST (Ardila et al. 2020), mTedx (Salesky et al. 2021), Must-C (Di Gangi et al. 2019) and VoxPopuli (C. Wang et al. 2021a), as well as speech translation data from the foreign languages into English and from English into German.

We remove training samples which transcription or written translation consists of multiple sentences, as LASER has been trained on single sentences only. For better training efficiency, we train speech encoders for each language family instead of each language, which has proven to work well for text (Heffernan et al. 2022). The language grouping is provided in Table 3.10.

Family	Languages
Romance	spa, fra, ita, por, ron
Slavic (+Baltic)	ces, pol, slk, slv, hrv, lit
Germanic	eng, deu, nld
Uralic	fin, est, hun

Table 3.10. – **Language family grouping.** Language family groups used to train speech encoders and HuBERT models. Lithuanian was our only Baltic language. In order to avoid training it alone, we added it to the Slavic language family.

To better handle imbalanced training data, we sample the training data from different languages with the approach presented in Section 3.2.1. For English (eng), Slovenian (slv), Lithuanian (lit) and Dutch (nld), we also trained separate monolingual speech encoders that had lower valid cosine losses compared to multilingual encoders, and these four monolingual encoders were used for mining.

3.3.1.2 Evaluation of speech encoders

We evaluated similarity search of audios against transcriptions on VoxPopuli ASR test set in Table 3.11, which is our target domain as we plan to mine unlabeled speech from VoxPopuli (see Section 3.3.1.3).

We also provide results for similarity search of audios against written translations or transcriptions on CoVoST2 test set for all languages from CoVoST2

Sim Search	ces	deu	eng	spa	est	fin	fra	hrv	hun	ita	lit	nld	pol	por	ron	slk	slv
# test instances	1k	1.7k	1.5k	1.4k	47	0.4k	1.5k	0.3k	1k	1k	39	1k	1.6k	—	1.3k	0.6k	0.3k
Error rates	0.6	1.0	0.2	0.7	0.0	0.7	0.5	0.3	1.1	4.9	0.0	0.8	0.9	—	0.9	0.7	3.1

Table 3.11. – **Similarity search results on VoxPopuli ASR.** Error rates (in %) of audio against transcriptions on VoxPopuli ASR test set.

covered by our speech encoders in Table 3.12, in order to evaluate cross-modal similarity search. We also report text-to-text similarity search (last line in Table 3.12) using the LASER text encoder as lower bound for the speech translation similarity search error rates, since we use gold transcriptions to search against written translations. We report error rates (in %) that are percentage of audio utterances incorrectly matched with text transcripts from the same test set. We note that error rates are very low for all languages (below 5% and around 1 or 2% for most languages), which is an initial validation of good-quality speech encoders before the large-scale mining.

	deu	eng	spa	est	fra	ita	nld	por	slv
# test sentences	14k	16k	13k	2k	15k	9k	2k	4k	0.4k
Audio									
vs. transcriptions	1.4	2.9	0.4	0.1	0.5	0.5	1.0	1.1	1.7
vs. en translations	3.3	—	1.3	1.0	1.5	1.7	4.4	1.9	4.4
Text transcription									
vs. en translations	2.0	—	1.0	0.1	1.0	1.3	2.4	0.7	0.8

Table 3.12. – **Similarity search results on CoVoST2.** Error rates (in %) on CoVoST2 test set for SpeechMatrix speech encoders.

We also compare, in Table 3.13, similarity search results of audio against written translations with the ones obtained by speech encoders trained in the first part, and we notice that our new speech encoders have lower error rates compared to encoders from Section 3.2.1.

Audio vs. en translations	deu	spa	fra
Encoders from Section 3.2.1	3.36	1.66	2.05
SpeechMatrix encoders	**3.27**	**1.26**	**1.55**

Table 3.13. – **Similarity search results compared to previous work.** Error rates (in %) on CoVoST2 test set.

3.3.1.3 Large-scale speech mining

We used VoxPopuli (C. Wang et al. 2021a) as our source of unlabeled speech for our 17 languages of focus. We downloaded the full unsegmented parliament session recordings from VoxPopuli.[7] We present in Table 3.14 the number of hours of unlabeled speech for each language, which range from 8k hours to 24k hours depending on the language.

We follow the same global speech mining approach as described in Section 3.2.2 and compare all segments in the spoken source language with all segments in the spoken target language. Similarity scores are calculated in both directions using the margin as described in Equation 3.1, considering $k = 16$ neighbors. Segments are considered to be parallel if the margin score exceeds a threshold, we use 1.06 if not specified otherwise.

Similarly to Librivox recordings, the VoxPopuli recordings have a rather long duration, e.g. one hour and a half on average for English. We apply VAD using Silero-VAD (Silero-Team 2021)[8] which supports over 100 languages and we follow the "over segmentation" approach outlined in Section 3.2.1. Initial experiments suggested that segments shorter than 1 second or longer than 20 seconds are unlikely to be aligned and therefore were excluded. After mining, the resulting speech alignments may have overlap as we over-segment the unlabeled speech. We follow the post-processing method presented in Section 3.2.2.1 to remove overlaps between mined speech segments on the source speech side but relax it to allow for some overlap between mined speech segments: for two audio segments that overlap on the source side, if the overlap represents more than 20% of the first segment and of the second segment, we discard the alignment with the lowest mining score. We did an ablation study on different thresholds of overlap ratio for one low-resource, one mid-resource and one high-resource direction and found that 20% was the best overlap threshold in all settings.

We report the statistics of the mined S2S pairs in Table 3.14, with a mining score threshold of 1.06. We call this corpus of S2S mined data SpeechMatrix. The mined data totals 418k hours of parallel speech with an average of 1,537 hours of source speech in all translation directions. While some high resource languages like English, Spanish or French can reach up to 5k hours of aligned speech with other spoken languages; lower resource languages such as Estonian and Lithuanian obtain much fewer alignments, with only a few hours of aligned speech for Lithuanian. We also perform mining of the source speech in sixteen languages against more than twenty billion English sentences from Common Crawl. This yielded speech-text alignments between 827 and $3,966$ hours (c.f. the

7. https://github.com/facebookresearch/voxpopuli
8. https://github.com/snakers4/silero-vad

last column of Table 3.14). The evaluation of speech-text alignments is not done in this study.

	Speech targets																	Text
Src/Tgt	**ces**	**deu**	**eng**	**spa**	**est**	**fin**	**fra**	**hrv**	**hun**	**ita**	**lit**	**nld**	**pol**	**por**	**ron**	**slk**	**slv**	**eng**
cs	-	2381	3208	2290	952	1312	2476	726	1396	2410	84	2377	2516	1867	1190	2146	452	2528
de	2386	-	4734	3113	901	1477	3536	498	1871	3476	41	3384	2632	2250	1281	1646	361	3073
en	3172	4676	-	4715	1585	2169	5178	824	2266	4897	82	4422	3583	3572	2258	2306	586	-
es	2240	3041	4708	-	862	1373	4446	528	1599	4418	47	3067	2646	3484	1857	1603	308	3966
et	943	892	1593	877	-	1201	934	265	1119	1019	39	1055	949	721	419	780	196	1578
fi	1296	1463	2180	1393	1197	-	1449	306	1473	1599	47	1654	1350	1128	621	977	260	1969
fr	2424	3457	5171	4455	923	1435	-	560	1711	4618	50	3273	2822	3384	1991	1657	326	3966
hr	736	507	854	553	273	317	588	-	328	615	24	546	660	433	277	586	136	1311
hu	1417	1897	2346	1672	1140	1507	1787	328	-	1855	68	1839	1566	1315	808	1064	311	2301
it	2404	3460	4948	4500	1028	1614	4700	607	1823	-	103	3414	2848	3421	1995	1656	474	2891
lt	78	38	79	46	37	44	48	21	61	95	-	77	80	35	18	64	6	827
nl	2322	3305	4396	3066	1040	1633	3269	521	1768	3355	80	-	2459	2399	1352	1646	458	2708
pl	2530	2646	3662	2735	967	1378	2913	656	1554	2883	88	2540	-	2121	1301	1892	431	2871
pt	1849	2224	3606	3525	722	1131	3421	421	1279	3403	37	2436	2087	-	1579	1358	247	3540
ro	1187	1275	2290	1894	423	627	2024	271	789	1996	19	1384	1288	1592	-	870	125	2784
sk	2127	1628	2329	1631	781	982	1685	574	1038	1650	69	1676	1869	1361	867	-	370	2090
sl	436	350	579	307	192	254	324	128	295	461	6	454	413	241	121	359	-	1267
# hours of unlabeled speech																		
	18.7k	23.2k	24.1k	21.4k	10.6k	14.2k	22.8k	8.1k	17.7k	21.9k	14.4k	19.0k	21.2k	17.5k	17.9k	12.1k	11.3k	

Table 3.14. – **Mined data statistics of SpeechMatrix.** Duration statistics (hours of source speech) of speech-to-speech alignments for each pair of 17 languages (for mining threshold of 1.06). The last column provides statistics for alignments of source speech against 21.5 billion sentences of English texts. The last row provides duration of raw speech from VoxPopuli used for mining.

3.3.2 Experiments & results

To evaluate the quality of the mined data, we initiated a collaboration with the Universal Speech Translation team from Meta to train S2ST models on SpeechMatrix data and report the translation performance.

3.3.2.1 Massively multilingual S2ST evaluation data

Besides the S2S mined data which will be used as the train set, we leverage labeled public speech datasets as the evaluation sets. We need to gather massively multilingual test sets for our 272 language directions in order to evaluate S2ST systems trained on our mined data. In our experiments, we derive test sets from three public corpora, evaluating translation models trained on mined data across different domains:

- EuroparlST (EPST) (Iranzo-Sánchez et al. 2020). It is a multilingual speech-to-text translation corpus built on recordings of debates from the European Parliament, containing 72 translation directions in 9 languages. [9]

- VoxPopuli (VP) (C. Wang et al. 2021a) S2S data, as part of VoxPopuli release, provides aligned source and target speech together with source transcriptions. We prepare the S2T data with target speech and source transcription as our test set. To ensure that there is no overlap between the mined data and VoxPopuli test sets, we remove speech from mined alignments which are from the same session as test samples. In order to keep as much mined data as possible, we use VoxPopuli test set only when a language direction is not covered by EPST considering their domain similarity. Moreover, similarity scores are provided to indicate the quality of VoxPopuli samples. To choose high-quality data, we sort all sessions in the VoxPopuli S2S data in a decreasing order of the average similarity score of their samples. We keep adding samples from highly ranked sessions to the test set until the test size reaches 1000.

- FLEURS (Conneau et al. 2023). Built upon N-way text translations from FLoRes (Goyal et al. 2022), FLEURS provides speech for aligned texts and creates S2S data covering all mined directions. We take its source speech and target texts as the test data. For this project, in the case where multiple utterances correspond to one piece of source text, we generate one test pair for each source utterance respectively. FLEURS texts are from English Wikipedia, which is a different domain from VoxPopuli and EPST.

Valid sets are prepared for S2ST modeling using VoxPopuli and FLEURS data in a similar way as test sets. For VoxPopuli, we extract a valid set of about 1000 samples by adding data from highly scored sessions which are not in the test set. FLEURS valid set is derived from its valid samples.

3.3.2.2 Experimental setup

As discussed in Chapter 2, recent progress in speech-to-speech modeling suggests to discretize the target speech waveform into a unit sequence, relieving models from the complexity of predicting continuous values. We borrow the idea of training S2U model, where units are pre-generated from target speech with a pre-trained HuBERT model (see Section 2.2.2). During S2U training, models are periodically evaluated on the valid set of speech-to-unit samples, and the best checkpoint with the lowest valid loss is saved for model inference.

9. eng, fra, deu, ita, spa, por, pol, ron and nld

When it comes to inference, speech could be synthesized from the predicted units with a vocoder, as the output of the S2ST pipeline. It is then transcribed into texts by an off-the-shelf ASR model for evaluation purposes. A BLEU score is calculated by comparing the automatic transcriptions against the ground truth target texts, which serves as the quantitative metric of mined data quality. This score, called ASR-BLEU score, is not a perfect metric for data quality, as it is unavoidably affected by the quality of ASR models. Next we discuss each module of the pipeline.

Speech-to-Unit. The S2U model takes the source speech and predicts a sequence of target units. It typically has an encoder-decoder architecture, where the encoder consists of convolutional and Transformer encoder layers, and the decoder is a Transformer decoder. We have experimented with different model variants, and discuss bilingual and multilingual training in Section 3.3.3 and Section 3.3.4.

HuBERT. HuBERT is used to extract speech features of audio frames, which are then grouped into k-means clusters. The continuous features are thus mapped to corresponding clusters. In this way, speech could be discretized into unit sequence where units are basically indices of clusters. We reuse the same HuBERT model and k-means clusters for English, Spanish and French as in (Lee et al. 2022b) for a fair comparison with existing results. We also train multilingual HuBERT models to cover other languages in SpeechMatrix, and HuBERT training details can be found in Appendix A.1.1.

Vocoder. Unit-based HiFi-GAN vocoders are trained to synthesize speech from unit sequence (Polyak et al. 2021). In our experiments, vocoders are separately trained from S2U model. We train vocoders on three datasets:

- CSS10 (Park and Mulc 2019). It is a single-speaker corpus which we use to train vocoders in German, Finnish, Hungarian and Dutch.
- VoxPopuli (C. Wang et al. 2021a). Given its ASR data with speaker id, we sort speakers based on their speech duration, and keep adding the top speakers until the speech is more than 20 hours.
- Common Voice (Ardila et al. 2020). Portuguese and Estonian are not covered by the two corpora above, and thus we turn to Common Voice. Again, we select top speakers and prepare 12-hour and 10-hour speech for the vocoder training in Portuguese and Estonian respectively.

Data preprocessing and training details are included in Appendix A.1.3.

ASR. In order to compute ASR-BLEU scores, we use off-the-shelf ASR models to transcribe the speech generated by vocoders. Details about the ASR models and their benchmark results of word error rates are provided in Appendix A.1.2.

3.3.3 Bilingual speech-to-speech translation baselines

In this part, we discuss the bilingual S2ST models trained in each of the 272 language directions in SpeechMatrix. The Textless model architecture is used for bilingual translation in our experiments (Lee et al. 2022a). A Textless model consists of a speech encoder, Transformer encoder and decoder.

Training. For a given direction, we extract units for source and target speech with their corresponding HuBERT models. Taking source speech, the model is trained to predict target unit sequence with cross-entropy loss as well as source unit reconstruction as an auxiliary task.

For the training efficiency of extensive S2ST experiments, we use a subset of mined data as the train set. Mined samples are selected if their alignment scores are above a preset threshold. We performed an analysis of the threshold selection in Appendix A.1.4.

Comparison with existing results. Since we adopt the same model as previous work (Lee et al. 2022a) with the only difference lying in the train set, it is straightforward to compare with existing results. Table 3.15 shows the results of S2ST models which are trained on our SpeechMatrix mined data compared to VoxPopuli S2ST data in each of four language directions: spa-eng, fra-eng, eng-spa and eng-fra. The threshold of mined data is set as 1.09 for these four directions, yielding an average of 1,436-hour train set. Compared with 480-hour labeled speech from VoxPopuli, SpeechMatrix achieves an an average improvement of 5.4 BLEU, indicating the good quality and usefulness of the mined data.

Train set		spa-eng	fra-eng	eng-spa	eng-fra
VoxPopuli S2S	Hours	532	523	415	451
	BLEU	13.1	15.4	16.4	15.8
SpeechMatrix ($t = 1.09$)	Hours	1,353	1,507	1,366	1,518
	BLEU	20.4	20.7	21.9	19.3

Table 3.15. – **Speech-to-speech evaluation of models trained on SpeechMatrix compared to previous work.** BLEU scores on EPST test sets by S2ST models with different training data.

3.3.3.1 Large-scale bilingual evaluation

A large-scale evaluation is performed covering 272 mined languages directions, and bilingual models are trained for each direction to establish baseline results in S2ST.

	ces	deu	eng	spa	est	fin	fra	hrv	hun	ita	lit	nld	pol	por	ron	slk	slv
cs	-	12.9	22.7	16.7	-	0.6	21.1	4.4	0.5	10.2	0.1	6.1	8.5	-	4.3	16.9	3.0
de	7.3	-	16.3	11.7	-	1.2	10.7	4.5	0.6	3.8	0.1	10.4	3.5	7.1	5.2	3.0	4.1
en	8.2	10.1	-	21.9	-	1.9	19.2	8.4	1.1	11.5	0.3	15.1	8.2	11.8	7.6	5.7	5.5
es	5.2	6.1	20.4	-	-	1.3	16.3	3.6	0.7	11.1	0.1	8.0	3.9	13.3	5.2	2.2	2.2
et	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
fi	3.0	9.0	19.7	11.4	-	-	14.1	1.5	0.0	5.8	0.1	6.6	4.5	-	4.4	1.7	1.6
fr	5.4	6.3	20.7	18.4	-	0.8	-	5.4	0.7	10.2	0.1	8.4	4.8	13.4	5.6	1.6	1.5
hr	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
hu	2.6	7.3	15.3	9.5	-	0.7	13.8	1.9	-	6.3	0.1	3.0	1.6	-	2.4	0.9	1.2
it	6.4	4.9	18.9	19.6	-	0.4	15.3	5.2	0.7	-	0.1	6.5	3.6	12.4	3.7	2.1	2.8
lt	0.2	0.0	3.1	0.8	-	0.0	0.7	0.1	0.0	0.6	-	0.7	0.1	-	0.0	0.0	0.1
nl	3.5	8.1	18.0	13.2	-	0.5	13.0	3.3	0.4	5.2	0.1	-	3.4	6.7	4.1	1.7	2.1
pl	7.2	2.8	4.9	6.3	-	1.0	5.5	4.5	0.5	5.8	0.2	1.6	-	6.1	3.2	4.7	2.4
pt	-	4.7	21.2	23.2	-	-	18.1	-	-	4.4	-	5.0	3.6	-	4.4	-	-
ro	4.6	6.5	22.6	20.1	-	0.8	18.6	2.4	0.4	8.7	0.1	3.5	4.6	10.3	-	2.3	0.7
sk	28.2	10.7	21.4	15.5	-	1.0	19.2	5.0	0.5	4.7	0.1	4.2	5.3	-	4.4	-	3.6
sl	4.0	11.1	19.5	8.6	-	0.8	13.2	4.8	0.4	6.0	0.1	4.5	6.7	-	1.1	1.7	-

Table 3.16. – **Mined data evaluation on EPST/VP.** BLEU scores of bilingual S2ST models on EPST/VP test sets. EPST score is underscored.

Table 3.16 summarizes performance of bilingual S2ST models on two test sets. In each direction, Table 3.16 reports ASR-BLEU scores in the European Parliament domain, either EPST or VoxPopuli set. Indeed, EPST, while being a well-known test set, only covers a subset of language directions. EPST ASR-BLEU is underlined to be distinguished from VoxPopuli ASR-BLEU. Results for FLEURS test data from the Wikipedia domain can be found in Appendix in Table A.3.

Bilingual results. Empirically we find that translations into high-resource languages such as eng, spa and fra outperform those into low-resource languages such as lit and slv based on the amount training data of these languages in Table 3.14.

It is also found that translation results are not symmetric for some language pairs, for example, ron-eng has a ASR-BLEU of 22.6 while eng-ron ASR-BLEU is only 7.6 on EPST. Besides different complexity levels of target languages and test sets, such asymmetry also results from the dependency of ASR-BLEU scores on the speech synthesis quality of the vocoder and transcription quality of the ASR model. For languages whose vocoder and ASR models are not good, they are likely to obtain low ASR-BLEU scores. In this case, Romanian vocoder and ASR are not as strong as English models as reflected by its higher word error rate in speech resynthesis as reported in Appendix A.1.3.

3.3.4 Multilingual speech-to-speech translation

Multilingual modeling has been explored in tasks of language understanding and machine translation, demonstrating knowledge transfer among languages. However, back in 2022, there were only few studies on multilingual S2ST on real speech, partially due to the lack of multilingual speech-to-speech resources. With the massively multilingual data we have mined, we are able to explore multilingual S2ST training.

In this work, we focus on many-to-English translation, studying the translation from 6 Slavic languages to English in Section 3.3.4.1 and the translation from all 16 languages in SpeechMatrix to English in Section 3.3.4.2. [10] We present here multilingual models used in our experiments:

- **Textless model**. The same model with 70M parameters that we use for bilingual evaluation is reused in the multilingual experiments. Given diverse multilingual data, we increase the model size for larger model capacity, trying multilingual models with 70M and 260M parameters.

- **XM Transformer**. Inspired by findings which showed that cross-modal pre-training is beneficial for speech translation (Popuri et al. 2022), we apply XM Transformer to multilingual training, whose encoder is initialized from pre-trained XLS-R model with 1B parameters (Babu et al. 2021) and decoder is initialized from a unit decoder pre-trained in an mBART style (Popuri et al. 2022). With multilingual speech-to-unit data, the model is further fine-tuned to minimize the cross-entropy loss in target unit prediction.

- **XM Transformer with Sparsity**. Sparse modeling, in particular Mixture-of-Experts (MoE), has been widely studied in multilingual machine translation. MoE increases the number of parameters without sacrificing computation efficiency. We explored GShard which is a sparse scaling technique proposed in (Lepikhin et al. 2021). A learnable gating function routes input tokens to different experts (NLLB Team et al. 2022). We apply GShard architecture on the decoder of XM Transformer, and expert weights are all initialized with the pretrained unit mBART.

10. English-to-many or many-to-many translations were not explored in this study from 2022, but in recent large-scale systems like (Seamless Communication et al. 2023a) which also heavily relies on mined S2ST data.

3.3.4.1 Slavic-to-English translation

The six Slavic (+Baltic) languages include Czech (ces), Croatian (hrv), Polish (pol), Slovak (slk), Slovenian (slv) and Lituanian (lit) (Lithuanian is also included as our only Baltic language). In the multilingual setting, all mined data from each these languages into English are combined (without upsampling) as the train set.

	Bilingual				Multilingual			
	EPST/VP	FLEURS	EPST/VP	FLEURS	EPST/VP	FLEURS	EPST/VP	FLEURS
Textless	70M		260M		70M		260M	
Avg.	14.3	5.1	16.8	6.5	14.1	2.5	22.4	11.2
XM	Dense(1.2B)				Dense (1.2B)		GShard (4.3B)	
Avg.	18.1	10.1			26.0	15.2	27.0	15.5

Table 3.17. – **Speech-to-speech evaluation of Slavic-to-English models.** Average ASR-BLEU of Slavic-to-English models in EPST/VP and FLEURS domains.

We summarize ASR-BLEU scores of different models averaged over six Slavic-to-English directions in Table 3.17. As is shown, Textless model benefits from the parameter increase to 260M, and multilingual training further brings ASR-BLEU gains of 5.6 and 4.7 in EPST/VP and FLEURS. We tried larger models than 260M but didn't see more gains.

Comparing against bilingual Textless model (70M), bilingual XM Transformer achieves +3.8 ASR-BLEU in EPST/VP and +5.0 ASR-BLEU in FLEURS. Multilingual training further improves dense XM Transformer by 7.9 and 5.1 ASR-BLEU. GShard with 64 experts brings +1.0 ASR-BLEU over dense XM Transformer to EPST/VP, and +0.3 ASR-BLEU to FLEURS. Overall the best Slavic-to-English translation is achieved by XM Transformer with GShard trained in the multilingual setting. This demonstrates that multilinguality, pre-training and model sparsity are helpful for speech-to-speech translation modeling.

3.3.4.2 All-to-English translation

We extend the multilingual focus by switching from the Slavic language family to all languages in SpeechMatrix. We adopt the best models obtained for Slavic-to-English translation, i.e. multilingual XM Transformer with both dense and sparse architectures.

Results. Compared with XM Transformer (1.2B) dense model, MoE-GShard64 (4.3B) with the same forward computation time, brings gains of +0.9 and +0.2 ASR-BLEU to EPST/VP and FLEURS respectively. Similar to our findings in Slavic-to-

	Dense (1.2B)		GShard (4.3B)	
	EPST/VP	FLEURS	EPST/VP	FLEURS
ces	29.9	18.7	30.9	18.2
deu	18.8	19.0	19.3	20.3
spa	22.8	15.2	23.3	15.9
est	-	16.7	-	16.7
fin	26.8	14.1	28.2	14.0
fra	23.5	18.3	24.1	18.9
hrv	-	16.6	-	16.8
hun	20.2	12.0	21.3	12.5
ita	36.3	16.2	37.8	14.9
lit	21.9	9.8	23.8	10.3
nld	21.4	16.4	22.1	17.3
pol	21.2	12.4	21.3	13.4
por	23.8	21.8	24.2	22.3
ron	25.1	19.7	25.0	19.8
slk	30.8	19.6	32.2	18.2
slv	28.3	13.7	29.9	13.7
avg	25.1	16.3	26.0	16.5

Table 3.18. – **Speech-to-speech evaluation of All-to-English multilingual models.** ASR-BLEU of All-to-English multilingual models across FLEURS and EPST/VP domains (for EPST/VP column, underlined scores are on EPST data, and others on VoxPopuli data).

English setting, increasing the capacity with sparse modeling benefits in-domain (EPST/VP) more than out-of-domain FLEURS test set.

Given sparse architecture of XM Transformer with GShard, all-to-English model shows +0.6 and -0.4 ASR-BLEU difference compared to the Slavic-to-English model on EPST/VP and FLEURS respectively, averaged over Slavic languages. Multilingual sparse model benefits from the additional in-domain data in other languages when evaluated in EPST/VP domain, while sees performance degradation in out-of-domain data.

3.4 Conclusion

In this chapter, we have applied a teacher-student approach to extend the existing LASER multilingual sentence embedding space to the speech modality. The speech encoders leverage pretrained multilingual speech representations of the XLS-R model. We have explored several training procedures and compared

multilingual with language specific speech encoders, based on multilingual and multimodal similarity search error rates. We have empirically shown that this embedding space is suitable for large-scale speech-to-text and speech-to-speech mining. The quality of the mined speech-to-text and speech-to-speech alignments was evaluated by training speech-to-text translation systems for the well-established CoVoST2 and Must-C tasks as well as speech-to-speech translation systems for real data, evaluated on CoVoST2.

Based on these promising results, we scaled speech-to-speech mining to introduce a large-scale multilingual speech-to-speech corpus mined from VoxPopuli, called SpeechMatrix. In 2022, it was the largest resource of speech alignments with a coverage of 17 languages. We performed an extensive evaluation of the mined parallel speech, showing the good quality of the speech alignments. Multilingual speech-to-speech models can be efficiently trained on this corpus and we suggested different methods, such as sparse scaling using Mixture-of-Experts, to further boost translation performance in the multilingual setting.

Further scaling of speech mining will be presented in Chapter 5 using a new sentence embedding space called SONAR.

Chapter 4

DECODING SENTENCE EMBEDDINGS AND ZERO-SHOT CROSS-MODAL MACHINE TRANSLATION

Chapter abstract

In the previous chapter, we demonstrated that a joint audio/text fixed-size representation space enables the comparison of sentences in terms of semantic similarity. However, it is still unclear how much information can be decoded from these representations. In this chapter, we explore the decoding of multilingual and multimodal sentence embeddings and how to perform zero-shot cross-modal machine translation in this framework. Multilingual speech and text sentences are encoded in a joint fixed-size representation space. Then, we compare different approaches to decode these multimodal and multilingual fixed-size representations, enabling zero-shot translation between languages and modalities. In this framework, models can be trained without the need of end-to-end cross-modal labeled translation data. Despite a fixed-size representation, we achieve very competitive results on several text and speech translation tasks. We also introduce the first results for zero-shot direct speech-to-speech and text-to-speech translation. Finally, we explore how this type of approach can be further improved with multilingual training. The work in this section has led to the writing of two publications:[1]

- *Paul-Ambroise Duquenne, Hongyu Gong, Benoit Sagot, and Holger Schwenk (Dec. 2022). "T-Modules: Translation Modules for Zero-Shot Cross-Modal Machine Translation". In:* Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. *Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5794–5806.* URL*: https://aclanthology.org/2022.emnlp-main.391*

1. This chapter is adapted from these two publications

- *Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot (2023c). "Modular Speech-to-Text Translation for Zero-Shot Cross-Modal Transfer". In:* Proc. INTERSPEECH 2023, *pp. 32–36*

Contents

4.1 Introduction

In this chapter, we explore how much information can be decoded from sentence embeddings, training several new decoders. We also analyze how such decoders may be used to perform zero-shot cross-modal machine translation, when decoding multilingual and multimodal sentence embeddings.

We notice that while sentence embeddings are fixed-size compact representations, they are also good candidates for compatible representations between speech and text across languages. The modality gap, first highlighted by the length mismatch between audio and text sequences, remains a key challenge for efficient cross-modal transfer in speech translation (Q. Dong et al. 2021; Liu et al. 2020b; Alinejad and Sarkar 2020; Ye et al. 2022; H. Zhang et al. 2022). Overcoming this modality gap could enable to leverage Machine Translation (MT) text labeled data to benefit speech translation tasks.

In the first part of this chapter, we compare different variants of teacher-student training to learn better multilingual speech and text representations based on LASER. This relates to Chapter 3 where we trained speech student encoders. But in this part, we additionally train student encoders for text in multiple languages. Then, we learn decoder models to decode these representations into text and speech in different languages, which enables cross-modal machine translation. We analyze the zero-shot cross-lingual and zero-shot cross-modal translation results, combining at inference time, independently trained encoders and decoders. This architecture is called T-Modules, which stands for Translation Modules.

In the second part, we investigate how such a modular architecture can lead to better results when multilingual training is used. We investigate the impact of multilingual text encoders, an English text decoder trained on multilingual

embedding inputs and multilingual speech encoders, either combining all languages at hand or only those in the same language family.

Throughout this chapter, we carry out experiments on English (eng), German (deu), French (fra), Spanish (spa), Catalan (cat), Turkish (tur), Japanese (jpn) and Mongolian (mon), which were chosen for their linguistic variety as well as covering both low- and high-resource settings.

4.2 T-Modules: translation modules for zero-shot cross-modal machine translation

4.2.1 Exploring training strategies

The purpose of this part is to explore how a common fixed-size representation for multilingual speech and multilingual text, such as the one presented in Chapter 3, can be efficiently decoded in text and speech in different languages. We investigate language-specific encoders and decoders compatible with a common fixed-size representation. Plugging one encoder with one decoder from different modalities and/or different languages enables performing zero-shot cross-modal translation.

To this end, we first study how to efficiently decode fixed-size LASER sentence representation for text. Second, we study how to improve similarity for sentence embeddings between languages, compared to the original LASER space. After an ablation study on the Japanese-English text translation direction, we extend the best training strategy to several other languages and a new modality, speech.

4.2.1.1 Better decoding of sentence embeddings

Motivations First, we studied how multilingual sentence embeddings can be efficiently decoded. We focused on LASER, like in Chapter 3, as it originally has a decoder, and we studied how we can improve the decoding of sentence embeddings. As an initial experiment, we evaluated auto-encoding of English sentences from FLoRes (Goyal et al. 2022) in Figure 4.1 left, with the original LASER encoder and decoder, bucketing sentences by length, and reporting BLEU scores.

The LASER encoder handles several languages: decoding these multilingual embeddings enables to translate the input sentence into English with the original LASER decoder. We report the BLEU scores for the different sentence lengths in Figure 4.1 right for the German-English translation direction from FLoRes. We

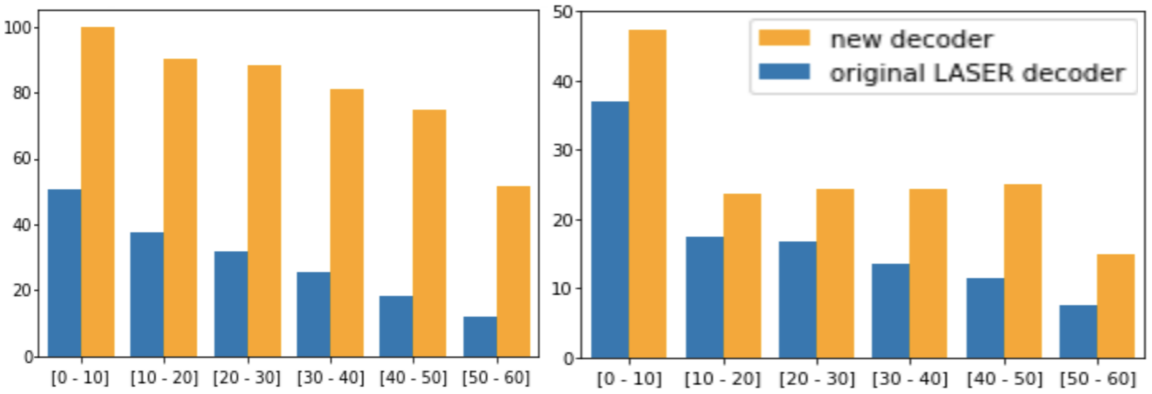

Figure 4.1. – **Decoding evaluations of the original LASER decoder and a newly trained transformer decoder.** BLEU vs. sentence length on FLoRes devtest. English auto-encoding (left), German-to-English translation (right).

notice that BLEU scores are low for both auto-encoding and translation tasks and decrease with the sentence length. The fixed-size representation seems to be a bottleneck for decoding tasks, especially for long sentences. However, the original LASER decoder is really shallow (one Long Short-Term Memory (LSTM) decoder layer), an interesting question is: can we improve decoding by training a new deeper decoder?

Training new decoders We chose to train a new decoder to decode LASER sentence embeddings, with a transformer architecture and 12 layers. To train this new decoder, we use an auto-encoding objective, feeding raw English sentences to the model: we use original LASER encoder, whose weights are not updated during training, and plug a new transformer decoder to decode the fixed-size sentence representation output by the LASER encoder (the decoder, using its cross-attention layers, attends on the sentence embedding output by the encoder only). We used 15B English sentences from CCNet (Wenzek et al. 2019) to train the decoder. We compare the new decoder with original LASER decoder on the auto-encoding task and the German-English translation task of FLoRes in Figure 4.1.

Results First, we notice an important boost on the auto-encoding task with the new decoder, with high BLEU scores even for sentences with more than 50 words. Second, training a new decoder with an auto-encoding objective improves the decoding of sentence embeddings from another language, German. The new decoder can be directly applied to German sentence embeddings because German embeddings are supposed to be close to their English translations encoded with LASER.

4.2.1.2 Making languages closer

Motivations To get an idea of the closeness of translations in the LASER space, we inspected the L2 squared distances of sentence embeddings in different languages to their English translations sentence embeddings in Figure 4.2. We noticed that high resource languages are closer in the LASER space to English, compared to low resource languages. Figure 4.2 also highlights that Japanese translations are more distant to English translations compared to German translations.

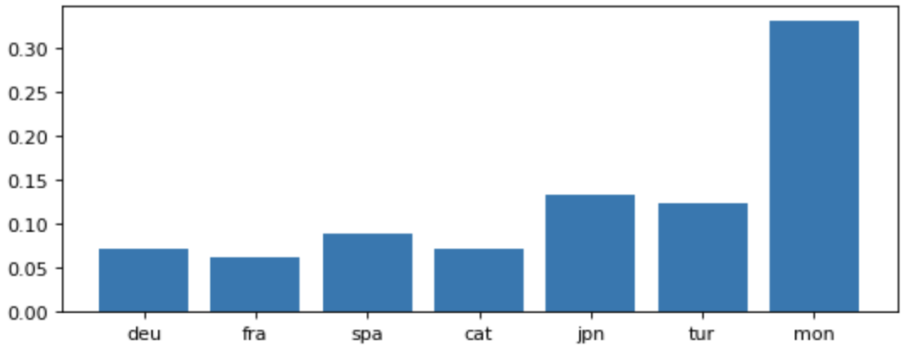

Figure 4.2. – **Distances between translations in LASER space.** L2 squared distances to English embeddings in LASER space for translations from FLoRes devtest

Therefore, we studied how our newly trained decoder is performing on a more distant language in LASER space, Japanese. We report the results of the jpn-eng translation task using the original decoder and the new decoder in Table 4.1. We notice that both decoders perform poorly on the jpn-eng translation tasks, but that the original LASER decoder leads to higher scores. An hypothesis is that the new decoder has over-fitted English embeddings leading to bad generalization on distant Japanese embeddings.

	jpn-eng
Original encoder + original decoder	6.9
Original encoder + new decoder	5.5
Student - max pooling + original decoder	12.2
Student - BOS pooling + new decoder	19.5
Student - max pooling + new decoder	22.5
Student - max pooling & CE + new decoder	22.6

Table 4.1. – **Results of initial decoding experiments.** spBLEU scores for jpn-eng on FLoRes devtest

Teacher-student training of text encoders To overcome this issue, we suggest to follow a method introduced by Reimers and Gurevych (2020), as presented in Section 2.3.2, where new encoders are trained to fit an existing sentence embedding space. Here, we are trying to make the Japanese translations closer to English embeddings in our 1024 dimensional space. The original LASER encoder is fixed during training to encode English translation, behaving as the teacher, while we train a new Japanese encoder as a student to fit English sentence embeddings. We use bitexts from CCMatrix for the jpn-eng pair to train the Japanese text student. Following (Reimers and Gurevych 2020), we minimize the Mean-Squared Error (MSE) loss (equivalent to L2 squared distance) between the generated Japanese sentence embedding and the target English sentence embedding.

The Japanese encoder is not trained from scratch, but we fine-tune XLM-R large. To extract the sentence embedding, we tested two methods: The classical output of the encoder corresponding to the Beginning-Of-Sentence (BOS) token, a method widely used for text classification ; or max-pooling of the encoder outputs, less common but LASER has been trained with such pooling method.

Finally, we tested another objective that is supposed to better match with our decoding task: we encode the Japanese sentence with the encoder being trained, decode the pooled sentence embedding with our new decoder which weights are not updated during training, and we compute the cross entropy loss of the output of the new decoder with the English target sentence. The training was unstable when using XLM-R weights as initialization. Therefore, instead of fine-tuning XLM-R, we fine-tune the encoder obtained from our previous method (trained with MSE loss), which leads to a stable training. We report all the results in Table 4.1. For text-to-text translation results, we use spBLEU of M2M-100 with the public checkpoint and script to evaluate on FLoRes, in order to compare with the supervised baselines.

Results In Table 4.1, we first notice that learning a new Japanese student significantly improves the results for the jpn-eng translation task. The best pooling method seems to be max-pooling, we suspect that this could be explained by the fact that LASER has been trained with max-pooling. The second step of fine-tuning with cross entropy loss does not improve the results for our jpn-eng translation task, despite of the significant decrease of cross entropy valid loss during this second step fine-tuning. This validates the use of a simple MSE loss which seems sufficient for future decoding purposes and is a lot cheaper in terms of computation compared to cross entropy loss. We conclude that learning a new Japanese student with max-pooling and MSE loss leads to the best results.

Using this new Japanese encoder, our new decoder significantly outperforms the original LASER encoder.

These experiments show that LASER sentence embeddings can be better decoded by training a new decoder on a large amount of raw text data. This new decoder can be used to decode sentence embeddings from other languages handled by LASER. However, translations are still more or less distant in the space, making them explicitly closer with a MSE loss objective significantly improves the results on a translation task. Therefore, we decide to extend this idea to other languages and a new modality, speech, to see if it can help performing cross-modal translation tasks.

4.2.2 Overall architecture

In this part, we present the overall architecture of our model.

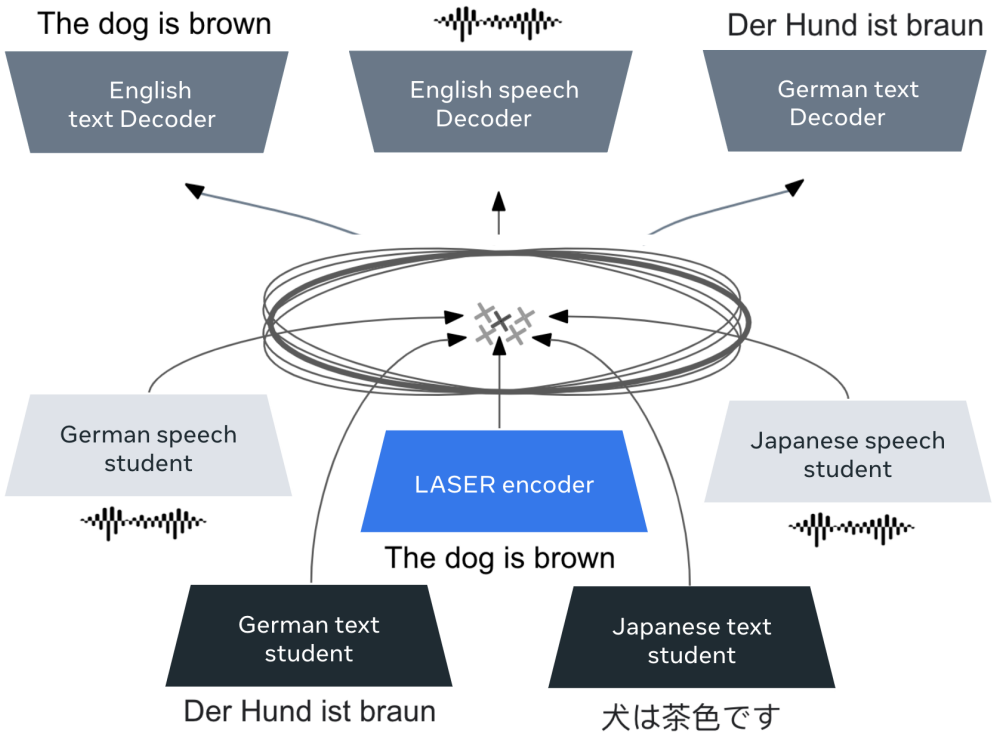

Figure 4.3. – **Summary of the T-Modules model architecture.** Independently trained encoders and decoders for speech and text in different languages which can be combined in a zero-shot way to perform cross-modal machine translation.

Text student encoders We now want to train several text students for different languages, in order to plug, at test time, these encoders to different decoders to perform translation tasks. We decide to use LASER **English** embeddings as our teacher for other text languages (the original multilinguality of LASER is not used here). This English space has proven to have good semantic properties: paraphrases are close in the embedding space, and makes it a good teacher for English translations. Moreover, most of MT data involve English translations that we will use to learn our text students. We focus on seven languages, namely, German, French, Spanish, Catalan, Japanese, Turkish, and Mongolian. We use CCMatrix bitexts to learn our text students, and bitexts mined with LASER3 (Heffernan et al. 2022) for Mongolian.

Text decoders We saw above that we can train a new English decoder with raw English data, using a fixed encoder and an auto-encoding objective. However, such an approach can lead to over-fitting to English sentence embeddings and bad generalization on other languages. We made languages closer together in our 1024 dimensional space thanks to our new student encoders but translations are not perfectly mapped to a real English sentence embedding in this continuous space. Therefore, we explore different methods to make the decoders robust locally in the sentence embedding space in order to generalize better on unseen languages.

First, we can improve our decoder training with an auto-encoding objective by adding synthetic noise in the sentence embedding space. We add noise to a sentence embedding by multiplying it by $1 + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \alpha)$. In our experiments, we took $\alpha = 0.25$, which leads to an empirical average L2 squared distance of approx. 0.05. between the noisy embedding and the original embedding.

Second, we tested another approach to make our decoder robust to translations in the sentence embedding space: we added bitexts from the deu-eng direction (chosen as it comes with an important amount of bitext data) to the training of the English decoder. We use bitexts from CCMatrix (Schwenk et al. 2021), and the English part of the bitexts for the auto-encoding loss in order to have a good balance between bitexts and raw data.

Finally, we trained decoders for five non-English languages to see how our approach behaves for other languages. All text decoders are 12-layers transformer decoders. With the hope that bitext data can help the decoder be robust to other unseen languages, we trained decoders for German, French, Spanish, Turkish and Mongolian and use eng-X bitexts, in addition to raw X data to train the decoders. For all decoder trainings, we use bitexts from CCMatrix (Schwenk et al. 2021), for the auto-encoding loss we use one side of the bitexts corresponding to the language that we are trying to decode. However, for Mongolian, we take all the

raw Mongolian text data from CCNet (Wenzek et al. 2019), to augment the training data size for this low-resource language.

Speech student encoders In Chapter 3, we showed that it is possible to learn speech students compatible with the LASER text space. The training of speech students is in fact similar to the one presented above for text. We use XLS-R large model variant (Babu et al. 2021) as in the SpeechMatrix project (cf. Chapter 3), which has more than two billion parameters and extracted the fixed-size representation for speech with max-pooling to follow what we have done for text students. We minimize the MSE loss between the output of the speech encoder and the transcription/translation encoded by one of our text encoders. Unlike the work we have done in Chapter 3, we did not use the original LASER encoder to encode text transcripts but our newly trained text students which are supposed to be close to the LASER English embeddings. As in Chapter 3, we can use either transcriptions or written translations as teachers for our speech student. We used CoVoST2 as our training data. Figure 4.4 summarizes the process to train a speech student with transcriptions only: First, we train a text student for the language we want to cover, we will use this encoder to encode transcriptions. Then, we train a speech student to fit text embeddings output by our text student.

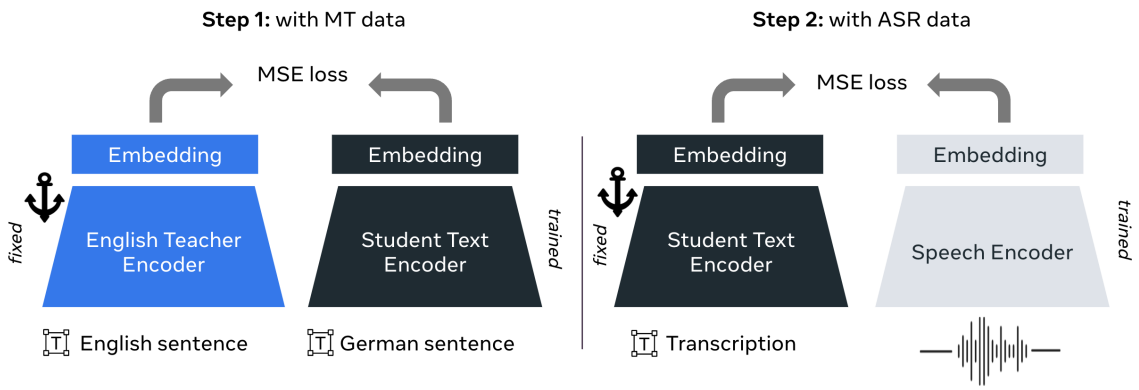

Figure 4.4. – **Incremental learning of a speech students.** As a first step, we train text students for non-English languages, using LASER English embeddings as targets. As a second step, we train speech students using the previously trained text encoders as teachers.

We trained independent speech student encoders for German, French, Turkish, Japanese and Mongolian spoken languages on the CoVoST2 training set. For Catalan and Spanish, we trained a single speech student encoder for both languages as they have high language similarity.

Speech decoders In this last part, we introduce a speech decoder in our framework, which can be learned with raw speech data. We focus on English speech decoding but it could be extended to other languages. To learn to decode English speech, we follow the work done by Lee et al. (2022b), who learn to decode HuBERT units. At test time, the generated units are transformed into speech using a vocoder.

One method is to follow the same approach presented for raw text data to learn an English decoder. The English speech encoder previously trained to fit LASER text space on CoVoST2 training set is used to encode raw speech, and its weights are not updated during training. We trained a unit decoder to decode sentence embeddings output by the speech encoder. The unit targets correspond to the ones of the input speech as we are trying to auto-encode speech. We follow the recipe of Lee et al. (2022b) to prepare target units as we are dealing with real speech data: we extract HuBERT units from input speech, normalize the units with their speech normalizer. This preparation of target data is done unsupervisedly and any raw speech data can be processed with this method. We summarize the speech decoder training in Figure 4.5. Another method is to leverage English speech recognition data where English text transcripts are encoded through LASER encoder which weights are fixed during training and a decoder predicts the sequence of units of the corresponding speech.

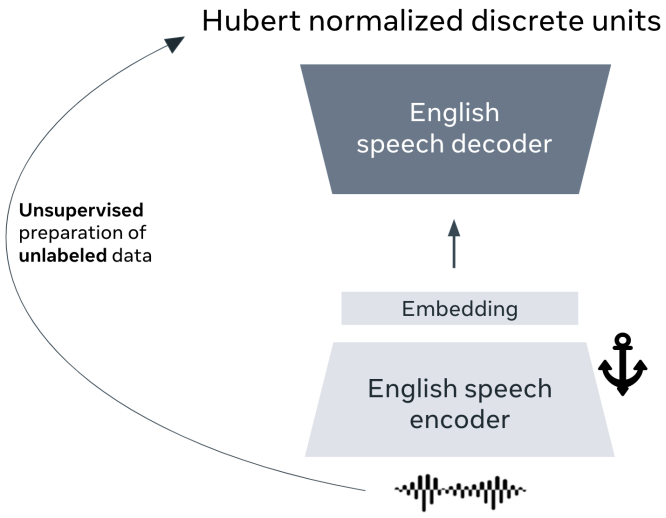

Figure 4.5. – **Speech decoder training.** We train an embedding-to-unit decoder in an unsupervised way. Raw speech is encoded in the sentence embedding space with a frozen speech encoder and the unit decoder is trained to recover the HuBERT units of the input speech.

Once the English speech decoder is trained, we can plug any text or speech encoder to perform direct text-to-speech or speech-to-speech translation in a

zero-shot way. This completes the overall T-Modules architecture, presented in Figure 4.3.

4.2.3 Results and discussion

Text-to-text translation We present the results for text-to-text translation for X-eng directions in Table 4.2 for the different decoder training methods on FLoRes devtest. *{en}_en decoder* corresponds to the decoder trained with an auto-encoding objective, *{en+noise}_en decoder* corresponds to the decoder trained with an auto-encoding objective and additional noise in the sentence embedding space, and *{en,de}_en decoder* corresponds to the decoder trained with a combination of deu-eng bitexts and English raw data. We compare our zero-shot text-to-text translation results with two supervised baselines: M2M-100 (Fan et al. 2021), a massively multilingual trained on many-to-many training data from different sources, with 24 encoder layers and 24 decoder layers; and DeepNet (H. Wang et al. 2022) a recent work trained on 1932 language directions from different sources with 100 encoder layers and 100 decoder layers, which was the best performing MT system at the time of the experiments. We put these results as a supervised reference but we recall that in our framework, we perform zero-shot text-to-text translation for most of the language pairs (all directions except deu-eng). Please note the cross-lingual transfer we obtain thanks to our training method: the English decoder has never seen Spanish embeddings before but is able to achieve competitive results compared to supervised baselines.

	deu	fra	spa	cat	jpn	tur	mon
This work - zero-shot except for deu-eng							
{en}_en decoder	40.7	41.9	30.4	36.7	22.5	32.8	13.0
{en+noise}_en decoder	39.5	40.6	29.4	35.8	23.7	33.2	16.4
{en,de}_en decoder	44.2	44.9	32.6	40.7	26.5	37.3	19.4
Previous works - supervised							
M2M-100 (12B - 48 layers) (Fan et al. 2021)	44.7	45.5	31.1	42.5	26.1	36.9	20.9
Deepnet (3.2B - 200 layers) (H. Wang et al. 2022)	48.0	49.9	35.2	46.2	32.7	44.2	23.9

Table 4.2. – **Zero-shot X-eng text-to-text translation.** spBLEU on FLoRes devtest for text-to-text X-eng translation using different English decoders compared to supervised baselines.

In Table 4.2, we see that adding synthetic noise to the sentence embeddings helps translating low resource languages unseen by the decoder. However, it slightly decreases the performance on high resource languages. Moreover, natural noise from deu-eng translations leads to even better results for both high and low resource languages, getting closer to the state-of-the-art MT results which have

been obtained with end-to-end training. Finally, we trained decoders for German, French, Spanish, Turkish and Mongolian in order to be able to translate from any of our languages to any other. We present the results in Table 4.3.

	eng	deu	fra	spa	cat	jpn	tur	mon
Translation into German								
This work - zero-shot expect for eng-deu								
{en,de}_de decoder	39.1	—	32.6	24.6	29.2	20.9	27.9	12.8
Previous works - supervised								
M2M-100 (12B - 48 layers) (Fan et al. 2021)	42.1	—	34.5	27.1	30.9	21.4	28.4	15.9
Deepnet (3.2B - 200 layers) (H. Wang et al. 2022)	46.0	—	36.2	29.2	32.5	24.7	31.9	21.7
Translation into Spanish								
This work - zero-shot expect for eng-spa								
{en,es}_es decoder	29.1	25.9	26.8	—	26.3	18.6	22.8	12.2
Previous works - supervised								
M2M-100 (12B - 48 layers) (Fan et al. 2021)	30.3	27.2	28.2	—	26.6	19.4	24.0	14.9
Deepnet (3.2B - 200 layers) (H. Wang et al. 2022)	32.2	28.3	28.8	—	26.9	21.5	25.9	18.8
Translation into French								
This work - zero-shot expect for eng-fra								
{en,fr}_fr decoder	49.1	38.3	—	31.2	37.6	25.3	33.4	16.6
Previous works - supervised								
M2M-100 (12B - 48 layers) (Fan et al. 2021)	51.4	42	—	32.8	39.7	26.6	35.1	20.8
Deepnet (3.2B - 200 layers) (H. Wang et al. 2022)	54.7	43.4	—	35.2	41.6	29.9	38.2	26.6
Translation into Turkish								
This work - zero-shot expect for eng-tur								
{en,tr}_tr decoder	31.2	27.1	26.4	21.5	24.2	19.1	—	13.7
Previous works - supervised								
M2M-100 (12B - 48 layers) (Fan et al. 2021)	32.8	26.9	26.6	22.3	24.3	18.6	—	16.1
Deepnet (3.2B - 200 layers) (H. Wang et al. 2022)	39.5	32.0	31.6	26.2	28.2	23.2	—	21.0
Translation into Mongolian								
This work - zero-shot expect for eng-mon								
{en,mn}_mn decoder	15.7	15.8	15.2	13.6	15.2	13.5	15.4	—
Previous works - supervised								
M2M-100 (12B - 48 layers) (Fan et al. 2021)	12.0	10.7	10.9	9.2	10.8	9.3	11.0	—
Deepnet (3.2B - 200 layers) (H. Wang et al. 2022)	18.3	16.8	16.2	15.0	15.8	13.7	15.9	—

Table 4.3. – **Zero-shot text-to-text translation for non-English decoders.** spBLEU on FLoRes devtest for text-to-text translation for deu, spa, fra, tur and mon decoders

Similar to what we noticed with our English decoder, we obtain excellent zero-shot cross-lingual transfer: the German decoder has never seen Japanese embeddings before and Japanese has never been aligned to German. However, the jpn-deu results are competitive compared to state-of-the-art translation models trained in an end-to-end way with much more data.

Speech-to-text translation Then, we tried to plug the decoders trained on text data to our speech encoders in order to perform zero-shot speech-to-text translation. We report direct Speech-to-Text Translation (S2TT) results in Table 4.4

for speech encoders trained with transcriptions as teachers. We have put several baselines for direct S2TT: two supervised baselines based on finetuning XLS-R (Babu et al. 2021) or mSLAM (Bapna et al. 2022) with S2TT data, which were the best performing S2TT systems at the time when our experiments were done. We also put the results on zero-shot cross-modal transfer from text to speech with the mSLAM pre-trained multimodal encoder, which is not working in this zero-shot setting.

	deu	fra	spa	cat	tur	jpn	mon
Speech training hours in CoVoST 2	184h	264h	113h	136h	4h	2h	3h
This work - zero-shot							
{en}_en decoder	27.3	32.2	34.0	24.7	7.4	3.3	0.1
{en+noise}_en decoder	29.2	33.3	35.3	27.3	10.1	5.2	0.3
{en,de}_en decoder	33.0	35.7	37.1	30.2	11.2	6.1	1.0
Previous work - zero-shot							
mSLAM (Bapna et al. 2022) cross-modal zero-shot	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Previous works - supervised							
XLSR (2B) (Babu et al. 2021)	33.6	37.6	39.2	33.8	16.7	3.5	1.6
mSLAM (2B) (Bapna et al. 2022)	35.9	39.0	41.0	35.4	24.2	3.3	0.8

Table 4.4. – **Zero-shot speech-to-text translation on CoVoST2.** BLEU on CoVoST2 test set for zero-shot speech-to-text translation (X-eng) compared to zero-shot and supervised previous work.

In our framework, the deu-eng S2TT direction benefits from cross-modal transfer while all other directions benefit from both cross-modal and cross-lingual transfer as the decoder has been trained on text and has only seen English and German embeddings. In this zero-shot cross-modal setting, we notice that the results are really competitive compared to supervised baselines trained end-to-end. Moreover, the supervised baselines use S2TT data, whereas our approach does not need S2TT data but only transcriptions. Except for Turkish, which has a really different morphological structure compared to English, S2TT results are close to their supervised counterpart trained with XLS-R. An interesting direction is jpn-eng, as we have a large amount of jpn-eng MT data but a really small amount of speech transcription data. For this task, we nearly doubled the BLEU score compared to supervised baselines without the need of Speech-to-Text (S2T) data.

We tested the different possible teachers for speech encoder training, namely transcription teacher (already presented), translation teacher, and both transcription and translation teachers. When using translation teacher, we use English text as the written translations from CoVoST2. We focus on two language directions, deu-eng (high resource) and jpn-eng (low resource). Results are shown in Table 4.5. We notice that a translation teacher is better if using the {en}_en decoder, which was expected as the decoder was trained on English embeddings.

Teacher mode:	Transcript.		Translation		Both	
	deu	jpn	deu	jpn	deu	jpn
{en}_en	27.3	3.3	27.9	3.5	28.1	3.1
{en+noise}_en	29.2	5.2	28.8	4.4	30.2	5.2
{en,de}_en	33.0	6.1	30.6	4.6	33.6	5.4

Table 4.5. – **Ablation on different teachers.** spBLEU on CoVoST2 test set for different teachers and decoders for zero-shot speech-to-text translation.

However, when using a decoder trained on noisy embeddings or with additional bitexts, results are better for speech encoders trained with transcription teacher rather than translation teacher. It may come from the fact that there exists a one-to-one mapping between transcriptions and audios, but not for audio and written translation (there can be several possible translations). For our high resource direction deu-eng, the best results are achieved when using both transcriptions and translations as teacher, reaching same performance level as with the end-to-end S2TT training of XLS-R.

Finally, we trained an English speech student with transcriptions on the Must-C training set (the TED talks domain, previous speech encoders were trained on CoVoST2) and compare our approach with the zero-shot approach by Escolano et al. (2021b). We report the results in Table 4.6. We notice significant improvements in the BLEU score compared to the baseline model, which was State-Of-The-Art (SOTA) at the time these experiments were done, for zero-shot S2TT on the Must-C dataset.

	State of the art in 2022	Our models
eng-deu	6.77	23.78
eng-fra	10.85	32.71
eng-spa	6.75	27.43

Table 4.6. – **Zero-shot speech-to-text translation on Must-C.** BLEU on Must-C test set for zero-shot speech translation, compared to the state of the art for zero-shot approaches in 2022 by (Escolano et al. 2021b).

Translation of text/speech into speech As presented in Section 4.2.2, we trained English speech decoders with raw English speech only or English speech transcriptions. We present three training settings: one decoder trained on raw English speech data from CoVoST2 (∼400h), another trained on raw English speech data from both Common Voice (CV) (∼2,000h) and Multilingual LibriSpeech (MLS) (∼40,000h), and finally another trained on English speech

transcription data from both Common Voice and Multilingual Librispeech. At test time, we can now plug these English speech decoders to any text or speech encoder. We focused on spa-eng and fra-eng language directions that were previously covered for direct Speech-to-Speech Translation (S2ST) in 2022. We report the results on CoVoST2 test set in Table 4.7. We also present Text-to-Speech Translation (T2ST) results, plugging text encoders to our speech decoders.

	spa-eng	fra-eng
Zero-shot text-to-speech		
trained on raw speech from CoVoST2	10.0	9.5
trained on raw speech from MLS + CV	22.8	20.9
trained on *eng* ASR data from MLS + CV	24.4	23.5
Zero-shot speech-to-speech		
trained on raw speech from CoVoST	9.9	9.1
trained on raw speech from MLS + CV	21.3	19.8
trained on *eng* ASR data from MLS + CV	22.4	21.1

(a) This work: *zero-shot* results

	spa-eng	fra-eng
Supervised speech-to-speech translation		
trained on VP	9.2	9.6
trained on VP + mined data	15.1	15.9
Supervised speech-to-speech via text pivot		
trained on VP+EPST+CoVoST2	26.9	27.3

(b) Results from previous *supervised* models trained by Lee et al. (2022b) on real (non synthetic) data. The speech-to-speech via text pivot baseline relies on speech-to-text by C. Wang et al. (2021a).

Table 4.7. – **Zero-shot text-to-speech and speech-to-speech translation.** ASR-BLEU on CoVoST2 test set for text-to-speech and speech-to-speech translation compared to other speech-to-speech translation baselines.

We compute ASR-BLEU scores on CoVoST2 based on an open-sourced Automatic Speech Recognition (ASR) system for English. We compare these results to a supervised baseline (Lee et al. 2022b) trained on real S2ST data from VoxPopuli (C. Wang et al. 2021a) and mined data from Chapter 3. We also provide a strong supervised baseline (back in 2022) composed of a Speech-to-text translation model from (C. Wang et al. 2021a) that is trained on a significant amount of S2TT data from VoxPopuli, EuroparlST and CoVoST2, followed by a text-to-unit model.

In Table 4.7, we notice that our speech decoders achieve strong results for this zero-shot setting, even with a limited amount of raw speech data. Incorporating much more raw speech data in the training, significantly improves the results. Using textual representation as input helps in speech decoder training, leading to best results. To the best of our knowledge, these were the first results for zero-shot direct S2ST. Such method to train a speech decoder based on fixed-size multimodal and multilingual semantic representations is further explored in Chapter 5.

This last experiment again highlights the compatibility between representations for different languages and modalities. Our approach enables to efficiently leverage raw speech data for T2ST and S2ST tasks.

4.3 Multilingual training in the T-Modules architecture

In the previous part, we highlighted the modularity of the T-Modules architecture, learning separately compatible encoders and decoders. While it can be seen as a strength, as one does not need to retrain the whole system to add a new language to the framework, it can also be seen as a limitation as the number of modules increases linearly with the number of languages. Moreover, multilingual models are known to benefit from positive cross-lingual transfer which can boost translation performance. In this part, we first extend the T-Modules architecture to multilingual training for text and trained a new text decoder on several X-eng directions. Then, we combine this decoder with speech encoders and explore multilingual training of speech encoders. A summary of the explored architecture is summarized in Figure 4.6

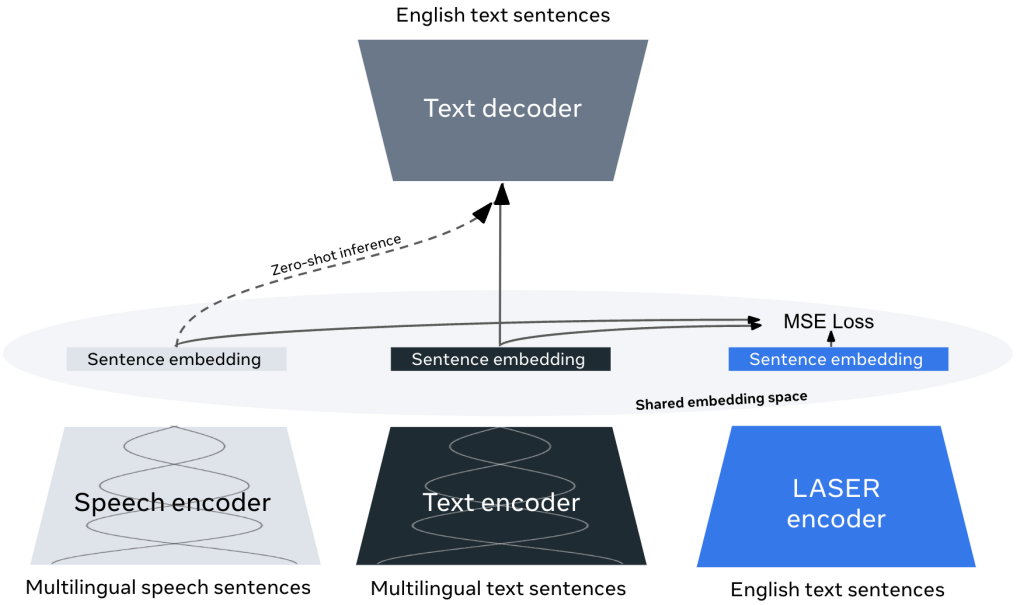

Figure 4.6. – **Summary of the multilingual T-Modules model architecture.** Multilingual speech and text encoders are trained in the T-Modules framework. A text decoder is trained with multilingual embedding inputs.

4.3.1 Multilingual training for text

We follow the training procedure introduced in Section 4.2, but train a multilingual text student. We focus on the same set of languages, namely English, German, French, Spanish, Catalan, Japanese, Turkish and Mongolian. The multilingual text encoder is initialized with XLM-R Large and finetuned with bitexts from CCMatrix. As in Section 4.2, we use additional bitext data mined with LASER3 for Mongolian. We train the multilingual text encoder to fit the English LASER space, minimizing the MSE loss between the outputs of our trained encoder and the LASER embeddings of the corresponding English translations. We will evaluate this new multilingual student encoder by combining it with the English decoder from Section 4.2 (named {en,de}_en decoder) trained to generate English translations from English and German sentence embeddings only, but that can be used on other languages at test time.

Based on this multilingual student encoder, we trained a new English decoder with bitexts involving all our languages of focus with English. Indeed, in Section 4.2, we showed that adding deu-eng bitexts to an English decoder training significantly improved the translation performance compared to a decoder trained only on monolingual English data. We extend this idea to a more multilingual setting and analyse the translation performance of this new decoder. For decoder training, we follow the same architecture as in Section 4.2, using a 12-layer transformer decoder. We name this new decoder: {en,de,fr,es,ca,ja,tr,mn}_en decoder.

Encoder	Decoder	deu	fra	spa	cat	jpn	tur	mon
Student multilingual	{en,de}_en	43.6	45.4	32.5	43.5	26.1	36.7	23.0
Student multilingual	{en,de,fr,es,ca,ja,tr,mn}_en	43.7	45.6	32.3	44.5	26.4	37.1	23.8
Baselines (previous works)								
Student monolingual Section 4.2	{en,de}_en	44.2	44.9	32.6	40.7	26.5	37.3	19.4
M2M-100	M2M100	44.7	45.5	31.1	42.5	26.1	36.9	20.9
DeepNet	Deepnet	48.0	49.9	35.2	46.2	32.7	44.2	23.9

Table 4.8. – **Text-to-text translation with a multilingual student.** BLEU on FLoRes devtest for text-to-text X-eng translation with a multilingual student. We compare our results to massively multilingual supervised models, M2M-100 (Fan et al. 2021) and DeepNet (H. Wang et al. 2022).

We present the MT scores on the FLoRes devtest set (Goyal et al. 2022) in Table 4.8, obtained by our new multilingual encoder combined with one or the other of our two English decoders ({en,de}_en and {en,de,fr,es,ca,ja,tr,mn}_en), and compare them to the scores obtained in Section 4.2. First, decoding the sentence embeddings produced by our multilingual student with the {en,de}_en decoder shows significant boosts in BLEU score for cat-eng and mon-eng translation

tasks, with respectively $+2.8$ and $+3.6$ BLEU gains compared to the work from Section 4.2 (line 1 and 3 in Table 4.8). For other language directions, we notice a performance degradation of -0.24 BLEU on average. Using the {en,de,fr,es,ca,ja,tr,mn}_en decoder instead significantly improves the translation results compared to decoding with the {en,de}_en decoder, with an additional gain of $+1.0$ and $+0.8$ for Catalan and Mongolian respectively (line 2 in Table 4.8). All other directions are also improved except for es-en. We hypothesise that Catalan, which has much less training data compared to French or Spanish, may benefit from cross-lingual transfer from those languages, and that Mongolian, which is our lowest resource language, may benefit from the larger training data size to avoid over-fitting.

In order to better analyse the cross-lingual transfer happening thanks to this multilingual training, we evaluate the translation of languages unseen during student and decoder training. Indeed, XLM-R Large was pre-trained on a larger set of languages. Aligning a subset of languages with LASER English embeddings may transfer to other languages in an unsupervised way thanks to cross-lingual pre-trained representations. Therefore, we encode unseen languages with different student encoders to analyse how unsupervised cross-lingual transfer occurs. We chose 4 different languages, Portuguese (por) and Italian (ita), which are Romance languages close to French, Spanish and Catalan; but also Dutch (nld), a Germanic language close to German; and Indonesian (ind) which is not similar to any language used for student and decoder training. We evaluate BLEU scores on the FLoRes devtest set in Table 4.9, using either the {en,de}_en or {en,de,fr,es,ca,ja,tr,mn}_en decoder.

Encoder	Decoder	por	ita	nld	ind
Student deu	{en,de}_en	35.2	22.7	**24.2**	**20.3**
Student cat	{en,de}_en	36.6	24.6	18.0	15.4
Student spa	{en,de}_en	**40.9**	**24.9**	19.4	16.0
Student multilingual	{en,de}_en	42.0	28.2	25.1	22.0
Student multilingual	{en,de,fr,es,ca,ja,tr,mn}_en	**43.3**	**28.6**	**25.7**	**24.2**

Table 4.9. – **Unsupervised text-to-text translation.** BLEU on FLoRes devtest for **unsupervised** text-to-text X-eng translation.

As expected, with monolingual student encoders, Portuguese and Italian are better translated using the Catalan or Spanish student encoders, while Dutch is better translated using the German student encoder. This highlights the impact of language similarity between the training and unseen languages. Remarkably, Indonesian works better with using the German student encoder than with the Spanish and Catalan ones. Also, Spanish student works better than the Catalan student on all unseen languages which may come from the fact that the Spanish

student was trained on much more training data. Moreover, we notice that our new multilingual student encoder outperforms all monolingual encoders by a high margin, thanks to cross-lingual transfer and larger training data size. Finally, plugging our new {en,de,fr,es,ca,ja,tr,mn}_en decoder further improves the results. This shows that multilingual training for text may help for translating low-resource and unseen languages in the T-Modules architecture.

4.3.2 Multilingual training for speech

Using our new {en,de,fr,es,ca,ja,tr,mn}_en decoder, we explore multilingual training of speech student encoders for either all languages or grouping languages by family, finetuning XLS-R 2B. In our languages of focus, we analyse the Romance family composed of French, Spanish and Catalan. Based on best results with students trained with a transcription teacher, we train a speech student with both transcription and translation as teachers which has previously shown best results in Section 4.2. We present results in Table 4.10.

Encoder	Decoder	deu	fra	spa	cat	tur	jpn	mon
Our models with no cross-modal supervision and full cross-lingual supervision								
Monoling. XLS-R student (transcription)	{en,de,fr,es,ca,ja,tr,mn}_en	33.0	37.3	37.7	29.4	12.0	6.1	0.8
Multiling. XLS-R student (transcription)	{en,de,fr,es,ca,ja,tr,mn}_en	30.9	35.3	37.5	31.9	16.5	0.9	0.7
Romance XLS-R student (transcription)	{en,de,fr,es,ca,ja,tr,mn}_en	—	37.0	38.4	33.2	—	—	—
Romance XLS-R student (transcr. + transl.)	{en,de,fr,es,ca,ja,tr,mn}_en	—	38.3	39.7	34.8	—	—	—
Our models with no cross-modal supervision and only partial cross-lingual supervision								
Monoling. XLS-R student (transcription)	{en,de}_en	33.0	35.7	36.3	27.9	11.2	6.1	1.0
Previous work: models with cross-modal and cross-lingual supervision								
XLS-R	finetuned mBART	33.6	37.6	39.2	33.8	16.7	3.5	1.6
mSLAM	mSLAM decoder	35.9	39.0	41.0	35.4	24.2	3.3	0.8
Whisper Large	Whisper Large	34.3	34.4	38.0	30.3	26.7	24.2	0.3

Table 4.10. – **Zero-shot speech-to-text translation.** BLEU on CoVoST2 test set for speech-to-text X-eng translation using a decoder trained on text for several X-eng directions. We compare our results to supervised baselines XLS-R (Babu et al. 2021), mSLAM (Bapna et al. 2022) and Whisper Large (Radford et al. 2023). Note that the latter was trained on significantly more speech data.

Then, we conduct the same analysis we have done for text on unseen languages for speech in Table 4.11. Indeed, XLS-R was pre-trained on a larger set of languages and we want to study the cross-lingual transfer that can occur when performing unsupervised S2TT on languages unseen during student training. Similarly to our findings on text, the German speech student encoder works better on Dutch, while Catalan, Spanish and Romance student encoders work

Encoder	Decoder	por	ita	nld	ind
Deu transcription student	{en,de,fr,es,ca,ja,tr,mn}_en	0.4	1.8	**3.6**	0.2
Spa transcription student	{en,de,fr,es,ca,ja,tr,mn}_en	**7.3**	**10.3**	0.8	0.2
Cat transcription student	{en,de,fr,es,ca,ja,tr,mn}_en	3.8	7.1	0.3	0.2
Romance transcription student	{en,de,fr,es,ca,ja,tr,mn}_en	**8.9**	**13.7**	1.2	0.2
Multilingual transcription student	{en,de,fr,es,ca,ja,tr,mn}_en	7.3	13.4	**5.1**	**0.4**

Table 4.11. – **Unsupervised speech-to-text translation.** BLEU on CoVoST2 test set for **unsupervised** speech-to-text X-eng translation.

better on Portuguese and Italian. The Spanish student encoder works better than the Catalan one on these languages, due to larger training data size. Indonesian is not working in this unsupervised setting, because it has no similarity with the languages used for training. Moreover, our findings regarding trained languages hold for unseen languages: the Romance encoder works better on Portuguese and Italian than the fully multilingual student and the Catalan or Spanish student encoders. However, not surprisingly, the fully multilingual student encoder works better for Dutch than the Romance encoder or the monolingual German one. This highlights even more that smart multilingual training for speech, grouping languages by family, yields to best results.

4.4 Conclusion

In this chapter, we extended the analysis of common fixed-size representation for text and speech in different languages to perform zero-shot cross-modal translation. By imposing a fixed-size representation and aligning explicitly languages and modalities, we have overcome the sentence length mismatch between audio and text, and obtained multilingual and multimodal representations compatible with decoders trained on other languages and/or modalities in a zero-shot setting. We have explored independent text and speech encoders for multiple languages compatible with text decoders for multiple languages as well as an English speech decoder. To the best of our knowledge, this was the first work tackling zero-shot direct T2ST and S2ST.

Finally, while we initially focused on zero-shot cross-lingual and cross-modal transfer in Section 4.2, in a second part, we focused on zero-shot cross-modal transfer only. We showed that a modular architecture can outperform strong supervised baselines while being zero-shot cross-modal. In this chapter, we focused on a limited set of languages, but going multilingual significantly improved the results. Interestingly, on the speech side, we showed that going

fully multilingual hurts the translation performance. However, we found that a language-family-wide encoder produces the best results while being easily trainable in such a modular framework. This line of work is extended to a large number of languages through the introduction of a new sentence embedding space called SONAR, that we present in the next chapter.

CHAPTER 5

SONAR: UTTERANCE-LEVEL REPRESENTATIONS FOR MASSIVELY MULTILINGUAL SPEECH AND TEXT

Chapter abstract

Based on conclusions drawn from the first two chapters, we introduce in this final chapter SONAR (Sentence-level multimOdal and laNguage-Agnostic Representations), a new massively multilingual speech/text sentence embedding space. Our single text encoder, covering 200 languages, substantially outperforms existing sentence embeddings such as LASER3 and LabSE on the `xsim` *and* `xsim++` *multilingual similarity search tasks. Using the teacher-student approach presented in previous chapters, speech segments can be embedded in the same SONAR embedding space using language-specific speech encoders trained on speech transcription data. Our new speech encoders outperform the ones based on LASER on similarity search tasks. We also provide a text decoder for 200 languages, which enables us to perform text-to-text and speech-to-text machine translation, including zero-shot language and modality combinations. Our Machine Translation (MT) results are competitive compared to the state-of-the-art NLLB 1B model, despite the fixed-size bottleneck representation. Our zero-shot speech-to-text translation results compare favorably with strong supervised baselines such as Whisper. We complement the SONAR semantic representation with a modality-specific additional representation for speech. This representation, encoding non-semantic speech properties, is learned together with an expressive speech decoder, that enables zero-shot expressive speech-to-speech translation in the SONAR framework. 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. The work in this section has led to the publication of two papers:* [1]

1. This chapter is adapted from these two papers

- *Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot (2023d).* SONAR: Sentence-Level Multimodal and Language-Agnostic Representations. URL: https://arxiv.org/abs/2308.11466

- *Paul-Ambroise Duquenne, Kevin Heffernan, Alexandre Mourachko, Benoît Sagot, and Holger Schwenk (2023b).* SONAR EXPRESSIVE: Zero-shot Expressive Speech-to-Speech Translation

Contents

5.1 Introduction

In the two first chapters, we introduced multilingual speech/text sentence embeddings based on an existing text sentence embedding space LASER. While LASER has proven to have good semantic properties and to be well-suited for mining, the encoder-decoder model is not based on the Transformer architecture but on a Long Short-Term Memory (LSTM) architecture (with a shallow decoder). Moreover, projects like NLLB introduced new bitext training data both human labeled, back-translated and mined data for language directions involving 200 languages, which represent much more training data that what was used to train the LASER embedding space. Finally, we can draw conclusions from Chapter 3, which showed that a text sentence embedding space can be easily extended to the speech modality, and from Chapter 4 which showed that sentence embeddings can be efficiently decoded into text or speech, to build a new sentence embedding space.

In this context, we focus on developing an encoder-decoder model for multilingual text to build, as an initial step, a new sentence embedding space for text. Our motivation for using an encoder-decoder approach for the initial text-based training phase is twofold. First, a multilingual decoder is trained along the multilingual encoder, which opens possibilities such as cross-modal MT (cf. Chapter 4). Second, a pre-trained state-of-the-art MT encoder-decoder model can

be used to initialize the whole encoder-decoder architecture. In contrast to LASER, we study the effect of different training objective functions on the properties of the resulting embedding space. In a second step, we train speech student encoders using our multilingual text encoder as a teacher. We demonstrate the cross-modal similarity search and Speech-to-Text Translation (S2TT) capabilities of the resulting SONAR framework.

In Chapter 4, we introduced a speech decoder in the T-Modules framework using HuBERT semantic discrete units of speech. Recently more acoustic units were introduced (Défossez et al. 2022) and Text-to-Speech (TTS) models that keep ones voice by prompting were introduced (Wang et al. 2023a). Inspired by this line of work, we complement SONAR with a new speech properties embedding and an expressive speech decoder. We demonstrate that we can perform high-quality, expressivity preserving zero-shot speech-to-speech translation in such a framework.

5.2 A state-of-the-art massively multilingual speech/-text sentence embedding space

In this part, we introduce our new multilingual and multimodal sentence embedding space SONAR that follows a two-step training strategy inspired by the work presented in Chapter 3 and Chapter 4.

5.2.1 Multilingual sentence representations for text

Contrarily to LASER's bidirectional LSTM architecture (Artetxe and Schwenk 2019b), SONAR relies on a Transformer encoder-decoder architecture, initialized with pre-trained MT model weights. However, as opposed to standard sequence-to-sequence architectures for MT, the architecture we use to train SONAR on parallel text data goes through a single vector bottleneck that represents the full sentence and does not use token-level cross-attention. The fixed-size sentence representation is computed by pooling the token-level outputs of the encoder. Instead of doing cross-attention on a variable-length sequence of encoder outputs, the decoder only attends to this single vector at each decoding step.

Unlike LASER (Artetxe and Schwenk 2019b), we do not only train our encoder-decoder architecture using an MT objective only. We investigated several other objectives and combinations thereof and analyzed their effect on the sentence embedding space and the decoding performance of the resulting model. We

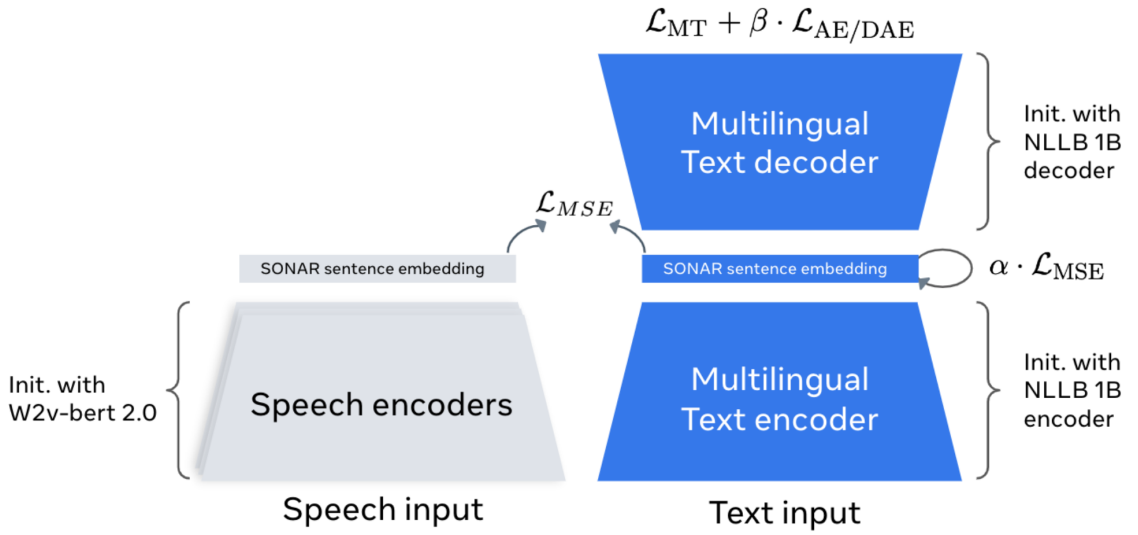

Figure 5.1. – **SONAR architecture.** An encoder-decoder with an intermediate fixed-size representation is trained with a combination of different objective functions. The resulting sentence embedding space is extended to the speech modality using teacher-student training.

introduce below the different objectives used to train our encoder-decoder architecture.

5.2.1.1 Objective functions

Translation objective Following the work of Artetxe and Schwenk (2019b), we used parallel data to train our encoder-decoder architecture with a translation objective. To better understand the motivation behind this objective, let us take this example: Given a triplet of translations x, y, z, where z is the English translation, decoding x and y into English may be easily achieved by the decoder if the sentence representation of these two input sentences are similar in the sentence embedding space. Training a encoder-decoder architecture on a translation objective may end up in this potential local minimum where translations are encoded closely to one another, so as to be decoded into the same target language sentence. However, there is no guarantee to converge to this local minimum. Nothing explicitly constrains a sentence in a language and its translation in another language to be encoded closely to one another. As a result, other local minima are possible, where translations are not encoded closely but still decoded into the same sentence for a given target language. To mitigate this, shallow decoders were used by Artetxe and Schwenk (2019b): a deeper decoder can more easily decode different points into the same sentence, whereas a shallower decoder

is more likely to need two vectors to be very similar whenever they must be decoded into the same sentence.

Auto-encoding and denoising auto-encoding objective Auto-encoders have been widely used to build representations. They have the advantage to encourage encoding fine-grained details of the input. However, this objective by itself is not likely to learn semantic representation of sentences. Moreover, this objective is much simpler to learn compared to a translation objective, which makes the combination of these two objectives difficult. To mitigate these issues, Liu et al. (2020a) introduce a *denoising* auto-encoding task, which has proven to be a good pre-training objective for translation tasks.

Mean-Squared Error (MSE) loss objective in the sentence embedding space Teacher-student approaches to multilingual sentence embedding space learning have shown that ensuring that translations of a same sentence are embedded close to one another in the sentence embedding space with an MSE loss works really well (Reimers and Gurevych 2020; Heffernan et al. 2022) and was again validated in Chapter 4. However, using this kind of loss without a frozen pre-existing teacher embedding space would lead to collapse (all inputs mapped to the same embedding), which is why contrastive learning methods were introduced to learn multilingual sentence embeddings from scratch (Feng et al. 2020). However, combining an MSE loss with a translation objective and/or a denoising auto-encoding objective could also prevent collapse from happening, as the model is forced to keep embeddings distinct to encode and decode different sentences.

Decoder finetuning In Chapter 4, we demonstrated that learning deep decoders for an existing sentence embedding space (in this case, LASER) can significantly improve translation and auto-encoding performance. While keeping the existing embedding space unchanged, such decoders greatly improve the decoding of sentence embeddings, therefore significantly improving auto-encoding and translation performance when combined with compatible encoders. This is of great interest for zero-shot (possibly cross-modal) translation, as shown in Chapter 4.

We introduce a decoder fine-tuning method called *random interpolation decoding*. Based on a trained encoder-decoder model with a bottleneck representation between the encoder and the decoder, we freeze the encoder weights and fine-tune the decoder weights only on a specific decoding task: Given a bitext x, y, we encode x and y with the frozen encoder, randomly draw z as a random interpolation of x and y embeddings, and learn to decode sentence embedding

z into y. This can be viewed as a continuous combination of translation and auto-encoding tasks.

5.2.2 Evaluations for text

To evaluate the semantic properties of the resulting sentence embedding space, we relied on a number of evaluation tasks on for the text modality:

xsim As presented in Chapter 2, cross-lingual similarity search, also called `xsim`,[2] evaluates the similarity between sentence embeddings across languages.

xsim++ More recently, `xsim++` was introduced as a more semantically challenging similarity search task (M. Chen et al. 2023),[2] as detailed in Chapter 2.

Translation tasks Multilingual embeddings are decoded into other target languages to perform MT. We report spBLEU (flores200) scores, in order to evaluate BLEU scores for low-resource languages as output, and COMET scores on the generated translations. Decoding sentence embeddings into other languages partially evaluates how much information is encoded in sentence embeddings, which is complementary to `xsim` and `xsim++` evaluations. However, please note that information may also be restored from the internal language modeling capabilities of the decoder, and not from the sentence embeddings themselves.

Auto-encoding task Similarly to translation tasks, we decode sentence embedding in the same language to perform auto-encoding and evaluate the content preservation of this operation.

All these evaluations for text were performed on FLoRes-200 devtest set,[3] which provides an N-way parallel corpus of translations in 200 languages.

5.2.3 Experiments on text

In this part, we first trained multilingual sentence embedding spaces using an encoder-decoder architecture on text and the objective functions presented above.

2. https://github.com/facebookresearch/LASER
3. https://github.com/facebookresearch/flores/tree/main/flores200

5.2.3.1 Training setup

We initialized our model with the NLLB 1B dense model (NLLB Team et al. 2022), that was trained on translation tasks with full cross-attention on variable length encoder outputs as it is commonly done for sequence-to-sequence MT model training. The model is composed of a 24 layers Transformer encoder and a 24 layers Transformer decoder and trained on a combination of human labeled data, back-translated data and mined data (NLLB Team et al. 2022). In order to build our fixed-size sentence representation, we added a pooling operation on the encoder outputs. The decoder only attends to this vector during training. As previously presented, several pooling methods are possible to train our new sentence embedding space: max-pooling as done in (Artetxe and Schwenk 2019b), mean-pooling as done in (Reimers and Gurevych 2019), or EOS-pooling which use the output representation of the End-Of-Sentence (EOS) special token appended at the end of sentences during NLLB training. Contrary to mean-pooling or EOS-pooling, max-pooling outputs a vector with a different range of values compared to NLLB training (due to the max operation), leading to worse results in our initial experiments. Since for EOS-pooling the training happened to be unstable during initial experiments, we focused on mean-pooling for the rest of our experiments. We trained our encoder-decoder model for 100k updates with same learning rate and batch size as NLLB training in the following experiments, unless explicitly specified. We used all bitext data used in the NLLB training, human labeled bitexts, back-translated data and mined data. This training dataset involves 200 target languages which contrasts with LASER training that only used English and Spanish as target languages. We ran an extensive study on the use of different training objectives, namely MT objective, Auto-Encoding (AE) objective, Denoising Auto-Encoding (DAE) objective and MSE loss in the sentence embedding space:

$$\mathcal{L} = \mathcal{L}_{\text{MT}} + \alpha \cdot \mathcal{L}_{\text{MSE}} + \beta \cdot \mathcal{L}_{\text{AE/DAE}} \tag{5.1}$$

We are using the same training data for auto-encoding and translation objectives, inputting the target sentences instead of the source sentences to perform auto-encoding of target sentences only. Incorporating more monolingual data in the training for the auto-encoding task is left to future work.

5.2.3.2 Initial experiment with translation objective only

We report the results of our experiments on text sentence embedding modeling in Table 5.1. Our first experiment, using only the translation objective for our encoder-decoder model with fixed-size intermediate representation, gives surprisingly good translation performance, given the bottleneck between the

encoder and the decoder. It yields -2 BLEU on X-eng direction and -3.8 BLEU on eng-X direction compared to NLLB 1B model with full-cross attention.

Method	X-eng↑	eng-X↑	AE↑	xsim↓	xsim++↓
$\mathcal{L}_{\text{MT}}$	33.2	21.1	28.6	1.3	19.6
$\mathcal{L}_{\text{MT}} + \mathcal{L}_{\text{AE}}$	17.6	18.6	94.6	15.9	65.7
$\mathcal{L}_{\text{MT}} + 0.1 \cdot \mathcal{L}_{\text{DAE}}$	31.6	20.9	41.6	2.6	26.2
$\mathcal{L}_{\text{MT}} + 0.1 \cdot \mathcal{L}_{\text{MSE}}$	31.7	20.2	27.2	1.3	14.3
SONAR sentence embedding space					
$\mathcal{L}_{\text{MT}} + 0.1 \cdot \mathcal{L}_{\text{MSE}} + 0.01 \cdot \mathcal{L}_{\text{DAE}}$	32.9	20.7	32.4	1.4	15.2
$\mathcal{L}_{\text{MT}} + 0.1 \cdot \mathcal{L}_{\text{MSE}} + 0.01 \cdot \mathcal{L}_{\text{DAE}}$ & fine-tuned dec.	32.7	21.6	41.7	1.4	15.2
MT topline					
NLLB 1B	35.2	24.9	39.0*	3.7*	49.6*
Similarity search baselines					
LaBSE	—	—	—	10.7	36.1
LASER3	—	—	—	5.1	36.4

Table 5.1. – **SONAR text evaluations.** Text evaluations on FLoRes-200 devtest set, averaged on the 200 languages supported by NLLB 1B: translation spBLEU for X-eng and eng-X directions, auto-encoding spBLEU, xsim and xsim++ similarity search results on X-eng pairs. Results with * are zero-shot evaluations of NLLB 1B model which was not trained to optimize these tasks.

We notice that auto-encoding evaluation (AE) significantly lags behind NLLB 1B model. This result may come from an inductive bias of the sequence-to-sequence architecture with full cross-attention, that could bias the model towards copying encoder inputs.

xsim and xsim++ results are significantly better compared to previous work, namely LaBSE and LASER3, on our 200 languages of focus, with approximately 45% relative reduction of xsim++ error rate compared to the baseline models. However, these supervised baselines were not trained for all the 200 languages of focus and are therefore evaluated in a zero-shot way for a number of them. We provide an evaluation on the intersecting languages of LASER3, LaBSE and the final SONAR embedding space in the following sections. Note that averaging NLLB 1B encoder outputs to perform similarity search already gives good xsim scores. This directly comes from the translation objective used during NLLB 1B training that encourages to encode multilingual sentences in similar ways for efficient cross-lingual transfer. However, the more difficult xsim++ evaluation remains challenging, in this zero-shot setting, for the original NLLB 1B model.

5.2.3.3 Experiments with auto-encoding objectives

Noticing the gap in the auto-encoding performance between the fixed-size bottleneck encoder-decoder model and NLLB 1B, we integrated an auto-encoding objective, hoping to close the gap with the NLLB 1B model. This model was only trained for 50k steps, as it converged quickly compared to other variants. We notice that auto-encoding task is easy to learn, even with a fixed-size bottleneck between the encoder and the decoder, almost reaching 95 BLEU in average on the 200 languages of NLLB. This shows that a lot of information can be efficiently stored in a fixed-size representation and that the bottleneck should not be seen as an hard limitation. While the translation performance on eng-X translation directions is not that much impacted, we see a big drop in translation performance for X-eng directions (-15,6 BLEU) compared to the fixed-size bottleneck encoder-decoder model trained only on a translation task. Moreover, we see a big drop in both `xsim` and `xsim++` evaluations showing that the model is not learning language-agnostic representations anymore, due to the auto-encoding objective that seems more easily optimized compared to the translation objective.

To mitigate the negative effects of the auto-encoding objective, while improving the auto-encoding performance at inference time, we switched to a denoising auto-encoding criterion, to avoid that the model overfits on the *copy* task. That would also make the task harder compared to simple auto-encoding and could be better combined with the non-trivial translation task. We also scaled down this denoising auto-encoding objective by a factor 0.1. This mostly mitigated the performance drops on translation tasks, while significantly boosting the auto-encoding task (+13 BLEU) compared to the translation-only model. However, the denoising auto-encoding criterion significantly affects the `xsim` and `xsim++` scores. This again shows that auto-encoding affects the organization of the sentence embedding space, learning distinct representations for different languages to optimize auto-encoding.

5.2.3.4 Experiments with cross-lingual similarity objective

Motivated by the recent distillation approaches to extend a sentence embedding space to new languages and the work presented in Chapter 4, explicitly aligning languages with an MSE criterion in the embedding space, we explored the use of an auxiliary MSE loss in the sentence embedding space. This is in addition to the translation loss, with the hope to explicitly make translations closer in the embedding space. In Table 5.1, we notice that this new constraint degrades the decoding performance of the encoder-decoder model for both translation and auto-encoding tasks. However, it significantly boosts the `xsim++` scores compared

to the encoder-decoder model trained only on a translation task, with -5.3 xsim++ error rate reduction.

5.2.3.5 Combining the objective functions to introduce the SONAR embedding space

Based on the conclusions of the previously trained models, we combined the translation loss, the auxiliary MSE loss and the denoising auto-encoding loss, to create the SONAR embedding space. In this run, the denoising auto-encoding loss is further downscaled, motivated by the high xsim++ score of the previously trained sentence embedding space trained on denoising auto-encoding. First, in the same tendency from previous training with (denoising) auto-encoding objective, we notice a slight degradation in xsim++ scores when adding the denoising auto-encoding in addition to the MSE loss. However, this degradation is only 0.9% which can be considered as acceptable. Initial experiments on larger scaling factors for the denoising auto-encoding criterion further increased, as expected, the xsim++ degradation, and we thus decided to stick with a 0.01 scaling factor for the denoising auto-encoding objective. On the other hand, for our new SONAR model, we see improvements on translation tasks compared to the model trained on MT and MSE loss. This may be due to efficient mitigation of collapse that could happen with MSE loss, thanks to the denoising auto-encoding objective. We also see big improvements in auto-encoding task (>+3.8 BLEU) compared to all models not trained with auto-encoding objectives. This variant seems to be the best setup in terms of sentence embedding space organization (following xsim and xsim++ scores) and decoding performance (following translation and auto-encoding evaluations). We also report the xsim and xsim++ results on the intersection of languages handled by LaBSE, LASER3 and SONAR in Table 5.2, and notice again that SONAR outperforms previous state-of-the-art sentence embedding spaces for multilingual similarity search.

	98 languages	
	xsim ↓	xsim++ ↓
SONAR	0.1	9.3
LASER3	1.1	27.5
LaBSE	1.5	15.4

Table 5.2. – **Multilingual similarity search results compared to previous work.** Comparison of similarity search results (error rates) on the intersection of languages handled by LaBSE, LASER3 and SONAR.

Finally, we tried to improve the decoding performances of our architecture, freezing the embedding space and our multilingual encoder, while fine-tuning

only the decoder. We used the *random interpolation decoding* method introduced in Section 5.2.1, where we compute a random interpolation of the source and target sentence embeddings and learn to decode the target sentence tokens. As the encoder is frozen, the `xsim` and `xsim++` scores won't change, but the decoding results will. With this decoder fine-tuning step, we notice similar translation results on the X-eng direction, while noticing a +0.9BLEU gain on the eng-X translation directions. More importantly, the auto-encoding performance is boosted by 9.3 BLEU with decoder fine-tuning method while the sentence embedding space is not affected. This finetuning step is trained for 50k additional steps.

We also evaluated the best performing models on translation tasks with COMET, which has proven to better correlate with human judgments compared to BLEU scores. We evaluated the two X-eng and eng-X directions involving the languages on which XLM-R was trained on, which are the languages supported by COMET (see Table 5.3). We see less that 1 point difference between our SONAR encoder-decoder model (with fine-tuned decoder) compared to NLLB 1B model for both eng-X and X-eng directions, showing the good quality of the translations.

Method	X-eng	eng-X
SONAR	85.9	83.4
SONAR & fine-tuned dec.	85.9	84.2
Topline		
NLLB 1B	86.5	85.2

Table 5.3. – **COMET text-to-text translation evaluation with SONAR.** Translation evaluations for X-eng and eng-X directions on FLoRes-200 devtest set: COMET scores averaged on 89 languages supported by both COMET and NLLB 1B models.

The NLLB 1B model still represents a topline, and to evaluate our SONAR framework against a more fair baseline involving a fixed-size sentence representation between the encoder and the decoder, we compared our results to the decoding of LASER embeddings, introduced in Chapter 4. As LASER3 encoders were trained with a cosine loss, the sentence embeddings cannot be efficiently decoded with T-Modules decoder from Chapter 4. This is why we trained new LASER3 encoders with MSE loss, which are really similar to text students Chapter 4, but added back-translated data from NLLB project in addition to the original training data of LASER3 encoders. These newly trained LASER3 $_{\text{MSE}}$ encoders can be combined with T-Modules decoder to perform X-eng translation. We report the results on 4 languages French, Spanish, Swahili and Russian in Table 5.4 and notice big improvements using SONAR on both X-eng translation task and `xsim++` evaluation.

	fra	spa	swh	rus
X-eng BLEU				
SONAR & fine-tuned dec.	46.1	34.5	42.4	37.1
LASER3 $_{\text{MSE}}$ & T-mod.	40.4	29.6	27.2	29.7
`xsim++`				
SONAR	4.8	7.9	7.1	6.5
LASER3 $_{\text{MSE}}$	7.6	12.6	15.2	12.4

Table 5.4. – **Comparison between SONAR and T-Modules.** Comparison to T-Modules framework based on LASER embedding space. spBLEU scores for X-eng translation directions on FLoRes-200 devtest set and `xsim++` for X-eng pairs on FLoRes-200 devtest set.

Please note that compared to the T-Modules work, we are able to encode and decode 200 languages with a single encoder and a single decoder. Finally, we provide an example of an English sentence auto-encoded with SONAR in Figure 5.2, to illustrate how it can preserve named-entities in long sentences.

English sentence from FLoRes:
Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.
Auto-encoding of the sentence with SONAR:
Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chairman of the clinical and scientific division of the Canadian Diabetes Association warned that the research is still in its early stages.

Figure 5.2. – **Example of a long sentence with named entities auto-encoded with SONAR.**

5.2.4 Multilingual sentence representations for speech

Based on the experiments and evaluations of multilingual sentence embedding spaces for text, we chose to rely only in this part on the text embedding space learned with translation, denoising auto-encoding and MSE objectives which seems to be a good trade-off between good semantic representation (`xsim` and `xsim++`) and good decoding performance (translation and auto-encoding). We follow a teacher-student approach to extend this space to the speech modality for several languages, as presented in the first two chapters of this thesis, using our newly trained text sentence embedding space as teacher. In this part, we only used transcriptions as targets, using written translations as targets is left for future work. As in previous work, we leveraged self-supervised pre-trained models, for

our speech encoders training, but this time using a W2v-BERT (Y.-A. Chung et al. 2021) pre-trained model as initialization.

5.2.5 Evaluations for speech

Similarly to the evaluations for our newly trained multilingual sentence embedding space for text, we report results for several evaluation tasks for the speech modality:

`xsim` for speech As introduced in Chapter 3, we calculate cross-modal similarity search. It follows the `xsim` evaluation presented above for text, but `xsim` is run on speech embeddings against English text translation embeddings.

`xsim++` for speech In addition to `xsim` computation for speech, we augment the English texts of FLEURS with challenging negative examples from the `xsim++` modified English sentences of FLoRes.

Zero-shot speech-to-text translation Following the work presented in Chapter 4, speech student encoders can be combined with text decoders at inference time. Since the speech encoder were trained on Automatic Speech Recognition (ASR) data only and the SONAR text decoder was only trained on text and has never seen speech embeddings during training, this corresponds to zero-shot S2TT. Similarly to text, it enables evaluating the content encoding in the speech embeddings. It also evaluates the compatibility between speech and text representations.

Zero-shot Automatic Speech Recognition: we also decode speech embeddings into the same language to perform speech recognition.

All these evaluations for speech were performed on FLEURS test set (Conneau et al. 2023), as it is a N-way parallel speech dataset in 102 languages (built on top of the text FLoRes benchmark).

5.2.6 Experiments on speech

We first performed an initial extensive study on five languages only: English (eng), Spanish (spa), French (fra), Russian (rus) and Swahili (swh). We then scale to 37 languages.

5.2.6.1 Experiments on 5 languages

We use a pre-trained W2v-BERT (Y.-A. Chung et al. 2021) 600 million parameter model to initialize the speech encoders and train them on Common Voice 12 ASR training set (Ardila et al. 2020). For our English speech encoder, we also used ASR training data from Must-C (Di Gangi et al. 2019), VoxPopuli (C. Wang et al. 2021a) and Librispeech (Panayotov et al. 2015).

We tested different pooling methods, namely mean-pooling, max-pooling and attention-pooling. Attention-pooling is performed with a three layer transformer decoder architecture with cross-attention on all the speech encoder outputs, in order to output a single vector as our speech sentence embedding (the decoder only outputs a single vector). Best results are achieved with attention-pooling (see details in Table A.1.5).

As a baseline, we compared our SONAR speech encoders to speech encoders trained with LASER as teacher (using our newly trained LASER3 $_{\text{MSE}}$ text encoders), with exact same training data and pre-trained W2v-BERT model, which makes them directly comparable. We report the `xsim` and `xsim++` cross-lingual and cross-modal results in Table 5.5 on FLEURS test set for foreign speech embeddings against English text embeddings.

	fra	**spa**	**swh**	**rus**
xsim				
SONAR	0.0	0.0	0.0	0.0
LASER3 $_{\text{MSE}}$	0.0	0.0	0.0	0.3
`xsim++`				
SONAR	12.3	13.9	22.8	24.6
LASER3 $_{\text{MSE}}$	17.5	24.9	40.7	42.1

Table 5.5. – **Multimodal and multilingual similarity search results.** Multilingual and multimodal similarity search evaluations on FLEURS test set: `xsim` and `xsim++` error rates on speech translation X-eng pairs.

Similarly to what M. Chen et al. (2023) noticed on FLoRes, `xsim` scores saturate to zero error rate on FLEURS test set, not providing useful insights on the multimodal sentence embedding space organization. Therefore, we also report `xsim++` scores for speech. We notice 41% `xsim++` relative reduction when switching from LASER as teacher to SONAR as teacher.

Following the work from Chapter 4, we decoded the speech sentence embeddings with our SONAR text decoder, performing zero-shot speech-to-text translation. Indeed, the text decoder has never seen speech sentence embeddings during training. Moreover, speech representations were only trained to match

their transcription representations but never translations. In Table 5.6, we report our zero-shot speech-to-text translation results on FLEURS test set for X-eng directions and compare it to the baseline trained on LASER space. We also report the state-of-the-art results (back at the time of these experiments) for speech-to-text translation, trained in a supervised way on significantly more training data. First, we notice large improvements in the BLEU scores compared to the LASER baseline on French, Spanish and Swahili, with an average 5.5 BLEU gain on these languages, while being slightly behind on Russian to English translation (-1.2 BLEU). This last result is surprising, as our SONAR speech encoder have much better `xsim++` score on Russian compared to the LASER speech encoder. Second, we notice that for our two high resource languages, namely French and Spanish, our zero-shot speech-to-text results are close to Whisper Large v1 supervised results, while being trained on much less training data. As for Swahili, our framework significantly outperforms Whisper models. We notice much better results for Russian-to-English for Whisper which was expected given the amount of training data and the supervised setting.

	fra	**spa**	**swh**	**rus**
Training hours				
SONAR/LASER ASR	0.8k	0.4k	0.3k	0.2k
Whisper ASR	10k	11k	0.01k	10k
Whisper S2TT	4k	7k	0.3k	8k
SONAR zero-shot S2TT				
SONAR	33.3	25.5	14.9	15.0
SONAR & fine-tuned decoder	33.4	24.8	15.6	14.6
Zero-shot S2TT baseline				
LASER3 MSE & T-Modules	30.7	22.9	3.7	16.2
Supervised S2TT toplines				
Whisper Large v1	33.8	27.0	5.2	30.2
Whisper Large v2	34.9	27.2	7.6	31.1

Table 5.6. – **Zero-shot speech-to-text translation.** spBLEU scores on FLEURS test set for zero-shot S2TT on X-eng directions.

Thanks to the compatibility across modalities and across languages, we decoded English, French, Spanish, Swahili and Russian speech sentence embeddings into the 200 text languages supported by our SONAR decoders. We report the zero-shot speech translation results using the fine-tuned SONAR decoder in Table 5.7. We notice that BLEU scores remain high for other languages than English, still in a zero-shot setting, highlighting again the compatibility between representations.

Finally, speech embeddings can be decoded into text in the same language, which can be seen as speech transcription. Since our model can often paraphrase

src\tgt	**eng**	**fra**	**spa**	**swh**	**rus**	**200 langs**
eng	69.7	44.3	26.9	27.8	29.8	17.7
fra	33.4	64.1	21.5	18.2	23.3	13.4
spa	24.8	25.1	58.9	16.0	16.8	11.7
swh	15.6	13.5	9.0	25.7	9.8	7.0
rus	14.6	17.3	11.0	10.4	35.0	8.0

Table 5.7. – **Massively multilingual zero-shot speech-to-text translation.** spBLEU scores on FLEURS test set for zero-shot S2TT on {eng,fra,spa,swh,rus}-X directions. Last column is the average spBLEU S2TT scores for decoding in the 200 languages supported by SONAR text decoder.

transcriptions, we report in Table 5.8 BLEU scores as well as bert-scores for this zero-shot transcription task. While being significantly behind on BLEU scores, which is expected as our model often paraphrases transcriptions, we see much less gap with Whisper transcriptions with the bert-score metric (which still advantages real transcriptions compared to paraphrases, but less than BLEU). Training data amount is also significantly different between the two setups, but it's interesting to notice that the gap in terms of bert-score remains reasonable.

	eng	**fra**	**spa**	**swh**	**rus**
Training hours					
SONAR/LASER ASR	4k	0.8k	0.4k	0.3k	0.2k
Whisper ASR	438k	10k	11k	0.01k	10k
Whisper S2TT	—	4k	7k	0.3k	8k
BLEU					
SONAR	64.7	54.3	50.0	17.7	29.1
SONAR & fine-tuned dec	69.7	64.1	58.9	25.7	35.0
Whisper v1	80.8	79.8	84.8	26.9	84.3
Whisper v2	81.3	82.0	85.3	36.0	85.3
Bert-score					
SONAR	0.948	0.926	0.923	0.808	0.853
SONAR & fine-tuned dec	0.951	0.939	0.936	0.831	0.870
Whisper v1	0.972	0.965	0.977	0.837	0.975
Whisper v2	0.972	0.969	0.979	0.865	0.978

Table 5.8. – **Speech recognition with SONAR.** Speech recognition spBLEU scores and Bert-scores on FLEURS test set.

5.2.6.2 Scaling to 37 languages

Through a research collaboration, the same recipe as described above extended the coverage of the speech encoders to 37 languages. These speech encoders were trained by linguistic language family, e.g. Romance or Indian languages, using speech transcriptions only, from public and licensed sources. Table 5.9 column "Train" gives statistics on the amount of training data.

As in Section 5.2.6.1, we evaluate the speech encoders by connecting them to the SONAR text decoder and calculate S2TT translation performance, as measured by spBLEU. Although our results are fully zero-shot speech translation, we achieve very competitive performance compared to the state-of-the-art model Whisper v2 large (Radford et al. 2023). The average on BLEU scores are slightly better for SONAR compared to Whisper, while being zero-shot speech translation. Our model performs less well on some high-resource languages like Mandarin Chinese, German or French, but outperforms Whisper for others like Spanish or Dutch and for several less common languages, like Swahili or Uzbek. Our modular approach seems to achieve particular good results on Indian languages: Bengali, Hindi, Kannada, Telugu, Tamil and Urdu.

In collaboration with the Seamless team at Meta, these encoders were used to mine S2TT and Speech-to-Speech Translation (S2ST) pairs for SeamlessM4T project (Seamless Communication et al. 2023a). Speech translation pairs were mined from 4 million hours of raw audio originating from a publicly available repository of crawled web data. Duration statistics of raw audio for each language can be found in Table 5.9. As for raw text data, it is originating from the same dataset used for NLLB bitext mining (NLLB Team et al. 2022). The amount of mined data is available in Table 5.9. A subset of this data was used to train the end model SeamlessM4T and ablation studies showed big improvements incorporating S2TT mined data (+2.7 BLEU for X-eng direction for instance) and slight improvements when using S2ST mined data. The reader is invited to read Seamless Communication et al. (2023a) for additional details.

5.2.7 Discussion

From all the experiments present in Section 5.2.3 and Section 5.2.6, we can draw a couple of high-level conclusions:

First, we have seen that the auto-encoding task can be greatly solved even with a fixed-size bottleneck between the encoder and the decoder, showing that a fixed-size representation should not be seen as a hard limitation, as a lot of information can be stored in a single vector. Then, similarly to Artetxe and

ISO	Language	Raw	Train	X-eng S2TT (↑BLEU)		Mined audio [h]		
		audio [h]	ASR [h]	Ours	Whisper	Sen2Txx	Sxx2Ten	Sxx2Sen
arb	MS Arabic	106755	822	30.9	26.9	1568	8072	776
ben	Bengali	7012	335	21.3	14.1	606	1345	263
cat	Catalan	43531	1738	37.7	36.9	1570	4411	354
ces	Czech	41318	181	32.0	30.3	1454	6905	602
cmn	Mandarin Chinese	79772	9320	18.6	20.8	5440	18760	1570
cym	Welsh	24161	99	14.5	13.4	–	4411	278
dan	Danish	34300	115	34.9	36.0	2499	6041	583
deu	German	490604	3329	36.2	38.8	91715	17634	1921
est	Estonian	12691	131	26.1	21.2	1022	3346	607
fin	Finish	32858	184	24.9	25.2	651	6086	526
fra	French	282179	2057	33.7	34.9	21523	17380	3337
hin	Hindi	15118	150	22.6	24.2	1041	2977	530
ind	Indonesian	11559	269	28.7	31.9	1938	2658	510
ita	Italian	79480	588	29.3	27.5	4378	6508	817
jpn	Japanese	75863	17319	20.2	20.8	1973	21287	1141
kan	Kannada	1451	114	21.4	13.1	–	232	198
kor	Korean	37854	316	17.1	24.2	–	8657	640
mlt	Maltese	2122	106	24.4	16.2	131	130	60
nld	Dutch	93933	1723	29.3	28.4	3720	6859	1210
pes	Western Persian	43788	386	24.4	20.9	–	7122	693
pol	Polish	53662	304	21.1	25.8	1324	9389	757
por	Portuguese	141931	269	38.3	41.4	4853	8696	928
ron	Romanian	18719	135	34.7	34.1	2770	2878	716
rus	Russian	103906	259	28.4	31.1	11296	13509	1252
slk	Slovak	16954	102	32.3	29.3	1267	3785	491
spa	Spanish	324086	1511	28.0	27.2	27778	17388	2727
swh	Swahili	18393	361	23.5	7.6	690	2620	484
tam	Tamil	100331	245	16.2	10.0	–	1664	867
tel	Telugu	3303	84	18.0	14.7	–	985	536
tgl	Tagalog	4497	108	14.6	26.8	–	633	266
tha	Thai	13421	195	16.9	17.8	2577	3563	542
tur	Turkish	23275	174	22.7	29.9	1417	6545	426
ukr	Ukrainian	6396	105	30.7	32.5	1220	1717	392
urd	Urdu	16882	185	19.7	18.1	773	3416	652
uzn	Uzbek	8105	115	20.0	6.6	475	1846	157
vie	Vietnamese	34336	194	19.1	21.9	1689	7692	868
Total/avr		2529741	43772	23.3	22.5	202796	239767	29161

Table 5.9. – **Statistics on speech encoders and amount of mined data.** Sen2Txx, Sxx2Ten, and SxxSen correspond to English speech paired with foreign text, foreign speech paired with English Text, and foreign Speech paired with English speech, respectively. Dashes are unmined directions. We provide the amount of raw audio data for mining and the amount of human-provided ASR transcripts to train the speech encoders. The speech encoders are evaluated for S2TT using spBLEU on the FLEURS test set. Our model performs zero-shot S2TT. Finally, the last three columns provide the amount of mined data. (Table and caption modified from (Seamless Communication et al. 2023a))

Schwenk (2019b), we noticed that a translation objective is well suited to build language-agnostic representations while making sure that the encoder model encodes enough information in the sentence embedding to be efficiently decoded (in another language). Adding an MSE loss in the training, which explicitly encourages to align languages in the sentence embedding space, leads to better language-agnostic representations. Moreover, denoising auto-encoding combined with MSE loss, can bring gains for decoding tasks, but too much of it affects the language-agnostic representations. Finally, the teacher-student approach to extend to the speech modality has once again proven to be effective and the mutual compatibility between speech and text multilingual embeddings is greatly highlighted by the fact that speech embeddings can be decoded in foreign text in a zero-shot way.

5.3 Speech properties embedding and expressive speech decoding

In this part, we aim at training a speech decoder model in the SONAR framework and complement the SONAR embeddings with an additional fixed-size representation for the non-semantic information of a speech signal. Following the modular training strategy presented in Chapter 4, 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). Compared to the work done in Chapter 4, we significantly scaled the amount of speech training data and used the new SONAR sentence embedding space. 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 the T-Modules work which uses HuBERT discrete units as target for the unit decoder, we used here 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-Generative Adversarial Network (GAN) vocoder has to be trained when using HuBERT units.

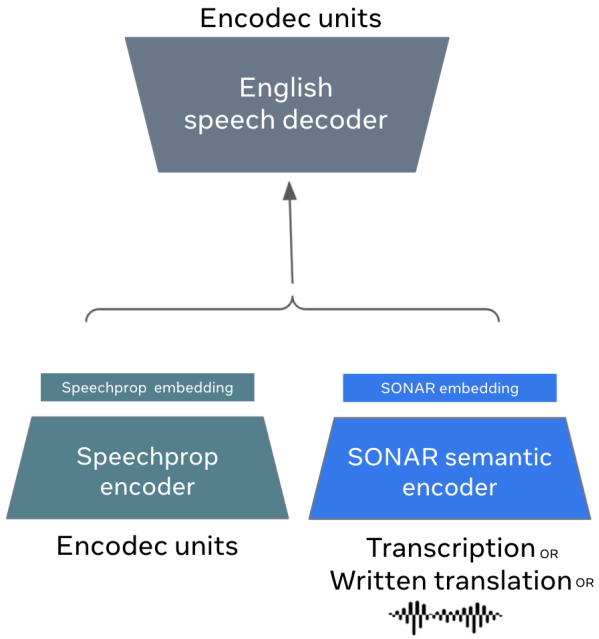

Figure 5.3. – **Model architecture for SONAR EXPRESSIVE.** An expressive speech decoder is trained to rely on both SONAR semantic embeddings and SpeechProp embeddings.

5.3.1 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 follow 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 5.3).

Training of new EnCodec model The original EnCodec model was trained on both speech and music with a 75Hz frame rate. In this part, a colleague from Meta introduced 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 three times shorter, improving memory usage during training. This new model 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). Natural multilingual speech data was used 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 5.3.3.2). In principle, our approach is generic and could be applied to any language. We use the SONAR English speech encoder introduced in Section 5.2 and train a new single speech encoder for the remaining five languages. We follow the recipe of Section 5.2 and train on public ASR data only. Table 5.10 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 5.10. – **Speech-to-text evaluation of our multilingual SONAR encoder.** Evaluation of our multilingual speech encoder on S2TT FLEURS test set (sacreBLEU scores).

5.3.2 Training setup

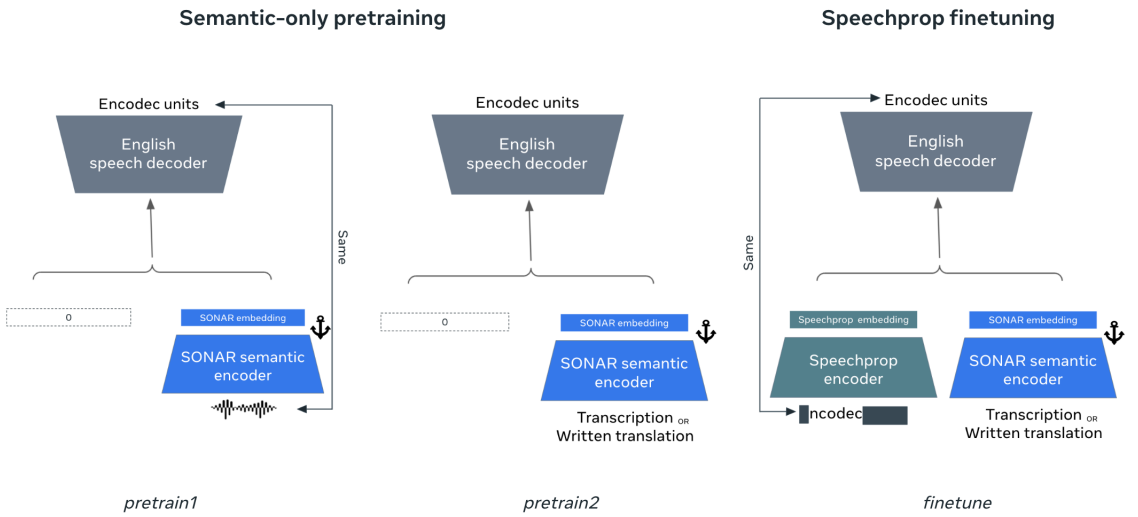

Figure 5.4. – **SONAR EXPRESSIVE multi-stage training.** The first stage of pre-training of the speech decoder uses only raw speech data and the speech decoder only relies on SONAR embeddings to predict output units. The second stage of pre-training uses S2TT data, and the speech decoder relies on multilingual SONAR text embeddings to predict output speech. The final stage is fine-tuning, introducing the SpeechProp encoder in the framework and using a random cropping strategy as regularization.

5.3.2.1 Multi-stage training

In order to train both the speech decoder and the SpeechProp encoder, we follow a multi-stage training strategy. 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. 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 unconditional setting, to

be used to compute classifier-free (CF) guidance (Gafni et al. 2022; Kreuk et al. 2022) 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, namely pre-train$_1$, pre-train$_2$ and fine-tune:

Pre-training with raw speech. (pre-train$_1$) 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. A first pre-training phase uses only raw monolingual speech data, following the training method introduced in T-Modules work. The raw speech is used in two different forms either sound frames or unsupervised Encodec units extracted from speech data. 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. (pre-train$_2$) A 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 by the NLLB system into our five 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 the T-Modules work for the 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 (fine-tune) Now that the speech decoders have 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.

5.3.3 Evaluation

5.3.3.1 Datasets

We evaluate our models on both the FLEURS (Conneau et al. 2023) test set as well as the new mExpresso benchmark dataset (Seamless Communication et al. 2023b).

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, bilingual 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 5.11.

	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 5.11. – **Data statistics per benchmark dataset.** Number of source hours for FLEURS and mExpresso.

5.3.3.2 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,[4] and then compare the transcriptions to the ground truth using sacreBLEU.[5]

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 (S. 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 (W.-C. 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.

4. large-v2 model.
5. 13a tokenizer.

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. [6] 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 Voice Activity Detection (VAD) (Silero-Team 2021). For both speech rate and pause alignment metrics, we used the open-sourced Rythmic Toolkit implementation (Seamless Communication et al. 2023b).

5.3.3.3 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).

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 5.12. – **Experiment on the use of classifier-free guidance for zero-shot S2ST.** ASR-BLEU performance with and without classifier-free guidance on FLEURS.

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 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.

5.3.3.4 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

6. For Mandarin, characters are treated as syllables.

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

Table 5.13. – **ASR-BLEU performance of SONAR EXPRESSIVE.** Speech decoding ASR-BLEU evaluation for TTS, Text-to-Speech Translation (T2ST) and S2ST tasks for the different training stages on FLEURS and mExpresso test sets.

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 5.13.

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 $\rightarrow$ 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 which are $\sim$3 seconds in average. This is to compare with TTS after the second stage of pre-training on mExpresso, which does not improve compared to the first stage of pre-training. Indeed, the average duration on mExpresso is 3.51 seconds (target-side) whereas the average duration on FLEURS is 9.78 seconds (target-side). After the second stage of pre-training, we get important performance boost on TTS ASR-BLEU on FLEURS, with more 10 ASR-BLEU gains (comparing "eng text" columns from pre-train$_1$ and pre-train$_2$).

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 automatic transcriptions. 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 models.

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 S2TT 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 (comparing "xxx text" columns from pre-train$_1$ and pre-train$_2$).

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 (comparing "xxx speech" columns from pre-train$_1$ and pre-train$_2$). 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 S2TT BLEU for cmn compared to 31.6 S2TT 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 style similarity. Results are shown in Table 5.14.

	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 5.14. – **Speaker style similarity performance in zero-shot S2ST.** Speaker style similarity evaluation of zero-shot S2ST with SONAR EXPRESSIVE for the different training stages on FLEURS and mExpresso test sets.

	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 5.15. – **Speech rate Spearman correlation in zero-shot S2ST.** Speech rate Spearman correlation evaluation of zero-shot S2ST with SONAR EXPRESSIVE for the different training stages on FLEURS and mExpresso test sets.

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 stage 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 5.15 and 5.16 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.

	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 5.16. – **Pause alignment results in zero-shot S2ST.** Pause alignment evaluation of zero-shot S2ST with SONAR EXPRESSIVE for the different training stages on FLEURS and mExpresso test sets.

Results from sentence-level prosodic similarity using the `AutoPCP` metric are shown in Table 5.17. As defined by the Prosody Consistency Protocol (cf. Section 5.3.3.2), a score of 1 corresponds 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.

5.3.3.5 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;

	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 5.17. – `AutoPCP` **results in zero-shot S2ST.** `AutoPCP` evaluation of zero-shot S2ST with SONAR EXPRESSIVE for the different training stages on FLEURS and mExpresso test sets.

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; Q. 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. Section 5.3.2.1). 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.

5.4 Conclusion

To conclude, we introduced a new multilingual and multimodal sentence embedding space called SONAR. We conducted an extensive study on objective functions to build our multilingual teacher sentence embedding space for text. We extended this new text sentence embedding space to the speech modality to introduce Sentence-level multimOdal and laNguage-Agnostic Representations (SONAR). We presented an extensive evaluation of our SONAR framework for both similarity search and decoding tasks.

Second, 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 S2TT 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.

CHAPTER 6

DISCUSSION

In this chapter, we provide a summary of the contributions made in this thesis and present insights into potential perspectives and future research directions.

6.1 Contributions

In this thesis, we introduced massively multilingual speech/text sentence embedding spaces. We demonstrated that a teacher-student approach is an effective way to transfer existing semantic properties to fixed-size representations for the speech modality. The increasing amount of Automatic Speech Recognition (ASR), Speech-to-Text Translation (S2TT) and Speech-to-Speech Translation (S2ST) training data in today's research may pave the way to multimodal and multilingual sentence embedding spaces learned from scratch.

Highly compatible representations between speech and text in different languages at the sentence level, where similar sentences are close in the embedding space independently of their language or their modality, allowed us to perform speech mining. We were able to automatically align Speech-to-Text (S2T) and Speech-to-Speech (S2S) pairs for several languages at scale, which were helpful to augment speech translation training sets and improved translation performances of end models. This line of research led to the creation of SpeechMatrix, totalling 418k hours of aligned S2S data for 136 language pairs, and used to train many-to-English S2ST models. It was extended in the Seamless project which scaled speech mining in 4 million hours of raw speech and was successfully used to train the state-of-the-art SeamlessM4T model.

Then, we analyzed how much we can decode from these fixed-size sentence representations, and discovered that it is possible to perform zero-shot cross-modal translation in such a framework. This line of research improved zero-shot S2TT and presented the first zero-shot S2ST results, leveraging big amounts of unlabeled speech data. These decoding experiments again highlighted the compatibility between speech and text representations across several languages.

Finally, based on these findings, we explored how to introduce a new massively multilingual speech/text sentence embedding space for better results in both multilingual and multimodal similarity search as well as decoding capabilities. Analysing the effect of several objective functions, we presented SONAR, a new sentence embedding space which outperforms previous multilingual and multimodal sentence embedding approaches. It covers 200 text languages as input and output and 37 spoken languages. It reduces `xsim++` error rates by 58% on these 200 languages compared to LASER3 or LaBSE, while providing competitive Machine Translation (MT) results when using the multilingual decoder. The compatibility between speech and text representations is highlighted with zero-shot speech decoding into text, introducing zero-shot S2TT that outperforms Whisper Large supervised results for a number of languages.

When dealing with the speech modality, not only the semantic content is conveyed, but also other non-semantic speech properties of the audio signal. We complement the SONAR semantic representations for speech with an additional embedding called SpeechProp embedding that encodes these expressive speech properties. Such line of work enabled to perform expressive S2ST from 5 languages into English based on the SONAR framework.

We demonstrated that SONAR can perform competitive text-to-text, speech-to-text, text-to-speech and speech-to-speech translation despite a fixed-size intermediate representation. This is interesting as cross-attention on encoder outputs has been used in all sequence-to-sequence models nowadays. While it is still providing best results, the good translation results of SONAR put into perspective the importance of such cross-attention.

The compatibility of multilingual speech/text representations of sentences could be further exploited for other Natural Language Processing (NLP) tasks to benefit from cross-lingual and cross-modal transfer. We discuss about these opportunities in the next section.

6.2 Perspectives

6.2.1 Multiple representation scales in text and speech

This thesis focused on sentence-level representations to solve multimodal NLP tasks. We may put this sentence-level focus into perspective with other scales of representation in NLP and Speech Processing.

Indeed, other scales of representation were explored in NLP, like words, which may have been the most studied scale of text representations, as presented

in Chapter 2. Tokenization techniques like BPE (Sennrich et al. 2016) and SentencePiece (Kudo and Richardson 2018) introduced subword splits to better handle unknown words and increase generalization. Another scale of representation of text is the character level (Boukkouri et al. 2020; J. H. Clark et al. 2022) which was shown to better handle user generated content that is much more noisy compared to clean text. On the other side of the spectrum, paragraph and document embeddings aim at representing larger units of text and have been mainly studied for retrieval tasks like Question Answering (Karpukhin et al. 2020).

In these different approaches, the modeling of text is biased towards some specific scale in order to solve some specific tasks. For example, as mentioned in Chapter 2, comparing two sentences can be easily done by concatenating them and feeding them to a language model, but it would be computationally intractable when comparing billions of sentences. By building sentence embeddings, we bias the representation of text in order to efficiently compare sentences in a vector space.

Choosing the right scale of the representations also helps to solve some technicalities, like the quadratic complexity of the self-attention mechanism of Transformers, which makes it difficult to take sequences of characters as input to represent long sentences or paragraphs. Interestingly, for a given architecture, the choice of the scale of input units may impact the language modeling performance too. For example, some work on spoken language modeling on HuBERT discrete units (Nguyen et al. 2023) found out that while acoustic consistency is really well preserved by those language models, the consistency in terms of meaning quickly drifts, leading to nonsensical continuations. This may be due to the fact that HuBERT units are phoneme-level representations and that long-term dependencies are difficult to handle on these low-scale units. Hierarchical representations for text modeling might be an answer to these technical obstacles as well as short-term and long-term consistency. For example, Megabyte (L. Yu et al. 2023) tries to address text with hierarchical representations trained in an end-to-end way to enable Large Language Models (LLM) to take smaller units as input.

While most of works in NLP are exploring methods which handle text at the sub-word level with heavily pre-trained LLMs, we advocate that other scales of representation should be explored and that pros and cons of each representation scale should be presented. The main focus of this thesis is the sentence scale. While the notion of sentence is not well defined for speech, representations of speech utterances in this thesis were learned to align with text sentence representations based on sentence-level training pairs. Text sentences have one big advantage compared to the word scale: it is a natural segmentation of

thoughts that human introduced to communicate. Moreover, these segments, representing one or more concepts, may be translated into different languages, spoken or written, or even illustrated. Therefore, it may be a good scale to bridge representations between languages and modalities, as presented in this thesis. Moreover, most of speech or text data is labeled at the sentence level in training datasets. In this thesis, we demonstrated again that leveraging MT labeled data can help to build strong semantic representations.

The question on how to successfully leverage all type of annotated data for several modalities and how to increase generalization in data modeling is still key nowadays. With an infinite amount of data and compute, end-to-end LLM modeling could potentially learn the desired underlying patterns in data, but in compute and data constrained setups, adding a bias in the way to model inputs has shown big improvements in several fields like Convolutional Neural Network (CNN) for images compared to general Multi-Layer Perceptron (MLP). In that context, exploring text modeling at the sentence level would be an interesting research perspective in the current LLM landscape.

When focusing on fixed-size sentence representations, one hyper-parameter, which was not explored in this thesis, is the dimensionality of the sentence embedding space. For both LASER and SONAR sentence embedding spaces, the dimensional is set to 1024, while LaBSE is using 768 dimensions. While this topic was explored for sentence embeddings before the transformer era (Conneau et al. 2018a), it may be interesting to study the impact of this hyper-parameter on the performance and generalization for new sentence embedding spaces like SONAR. Indeed, on one hand, a bigger dimension may boost linear probing tasks but, on the other hand, a constrained small dimensionality could act as regularization and therefore improve generalization or cross-lingual transfer.

6.2.2 From a modular framework to modality-agnostic and language-agnostic modeling

The SONAR framework offers compatible semantic representations for speech and text in different languages, as well as decoders to translate back to the text or speech domain. Moreover, non-semantic content for some modality specificity can be encoded in a disentangled representation, which can be taken into account when decoding in that specific modality. Such a framework, which provides competitive translation results, raises the question of how much full cross-attention is really needed. Indeed, the intermediate fixed-size SONAR representation, initially seen as a bottleneck, does not hurt much the translation performance. Interestingly, we noticed that the traditional Transformer

architecture (without fixed-size intermediate representation) biases the sequence-to-sequence modeling of the machine translation task towards good zero-shot auto-encoding performance. On the contrary, this is not the case when imposing a fixed-size intermediate representation, and an additional auto-encoding loss should be added to get similar performances. Once again, it highlights the fact that architectural choices and inductive biases have impact on generalization performance, but that constraining the model with alternative objective functions and enough training data may close the gap.

The multimodality aspect of the SONAR framework could theoretically be extended. One may interested in adding the visual modality to the SONAR embedding space, building an image decoder for instance. An image decoder should be rather straightforward as some auto-regressive or diffusion text-to-image models are already taking CLIP (Radford et al. 2021) text sentence embeddings as text conditioning to generate an image. With enough text-image training data, an image decoder could be trained to be conditioned on English text sentence embeddings. The zero-shot cross-lingual and cross-modal transfers could potentially enable zero-shot speech-to-image generation, from any spoken language handled by SONAR.

However, building a image encoder in the SONAR framework is not well defined, as images may be described by many semantically different captions. This one-to-many mapping is clearly different from speech-text synchronous ASR alignments. One could train an image encoder for retrieval purposes with a contrastive loss, as done in CLIP. But using existing SONAR decoders to decode such image embeddings is still undefined, as one may expect many possible captions as output. One way to overcome such issue would be to train a stochastic image encoder, that would randomly generate an embedding of one possible caption in the SONAR embedding space. However, such predicted embedding would be a partial representation of the input image only. One could also imagine to extend SONAR to other modalities. For example, some recent work showed that a model could be trained to translate non-invasive brain recordings into images (Benchetrit et al. 2023). In this context, the SONAR high-level semantic space could potentially be used as teacher for a brain recording encoder. Many other modalities of language may be included in such a framework like sign language.

Finally, we advocate that some NLP semantic transformation tasks may be directly modeled in the sentence latent space. This would leverage the efficient zero-shot cross-lingual and cross-modal transfers in the SONAR framework: one could learn a semantic transformation in the latent space for one language and one modality and apply this learned transformation in a zero-shot way to new languages and new modalities. This is described in Figure 6.1. In this context,

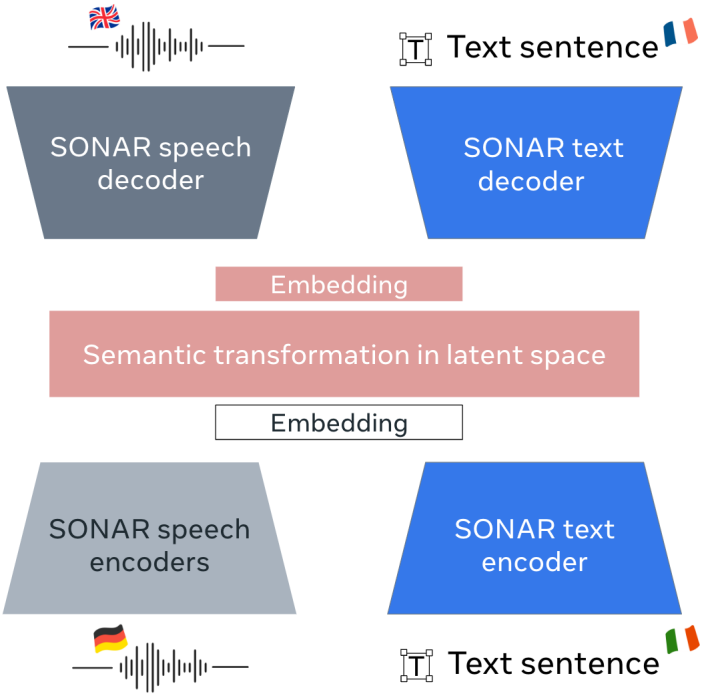

Figure 6.1. – **Semantic transformation in the SONAR sentence latent space.** After computing SONAR embeddings, semantic transformation learned on one language and one modality could be applied to the SONAR embeddings before decoding it into either speech or text.

semantic transformations as delta vectors in the SONAR embedding space were explored by a colleague at Meta, Marco Pennacchiotti, and proved effective for semantic manipulation like masculine to feminine or singular to plural operations (not yet released). An example, of these operations on an unseen sentence from FLEURS (Conneau et al. 2023) is shown in Figure 6.2. The zero-shot transfer to new languages and new modalities is a significant advantage of the SONAR framework. These semantic transformations could be coupled with a preservation of non-semantic properties of the transformed sentence decoded into speech, handled by the disentangled SpeechProp representation.

As a long-term perspective, more generic NLP tasks could be modeled directly in the sentence embedding space in the hope of benefiting from cross-lingual and cross-modal transfers. NLP tasks may be modeled as conditional probability distributions in this continuous space and techniques like diffusion models (Ho et al. 2020) could be leveraged to learn such tasks.

The prediction of sentence embeddings also raises the question of the organization of the embedding space as well as the robustness of decoders. Indeed, it may be the case that predicting precise sentence embeddings which can be correctly decoded is a difficult task. The question of the density of the space

```
emb = sonar_embed("He was subsequently relocated to Addenbrooke's Hospital in Cambridge.")
sonar_text_decode(emb + delta_gender, lang='eng')
```

```
She was subsequently transferred to Addenbrooke's Hospital in Cambridge.
```

```
sonar_text_decode(emb + delta_gender, lang='fra')
```

```
Elle a ensuite été transférée à l'hôpital d'Addenbrooke's à Cambridge.
```

```
sonar_text_decode(emb + delta_sing_plur, lang='eng')
```

```
They were subsequently relocated to Addenbrooke's Hospitals in Cambridge.
```

Figure 6.2. – **Results of an initial experiment on delta vectors in the SONAR space.** A sentence is transformed directly in the SONAR latent space and decoded into different languages.

for different sentence lengths remains an open question and could determine the level of difficulty of predicting sentence embeddings. Moreover, the robustness of the decoders to imperfect predicted embeddings would also be an area of research.

6.2.3 Multimodal communication tasks in the LLM landscape

Given the recent democratization and increasing popularity of instruction-finetuned Large Language Models (LLMs) (Ouyang et al. 2022; Touvron et al. 2023), it is important for any NLP modeling task to be discussed in the light of the current capabilities of these LLMs. Indeed, machine translation is one of the NLP tasks that LLMs are starting to master well for a few high-resource languages, compared to specialized models (Xu et al. 2023). These models are still limited in terms of language coverage compared to a state-of-the-art MT model like NLLB, but their multitask instruction finetuning may increase generalization and comes with several other advantages. Current state-of-the-art MT model are focusing on sentence level machine translation, whereas a LLM could more easily address document level translation, keeping long-term consistency in the translated document. Moreover, LLMs offer more controllability and allows for refinements of the translations with possible interactions with such general-purpose chatbots.

However, these Large Language Models come with some limitations. The first one is the limited language coverage for translation, as highlighted before. Then, the computational cost of inference with LLMs, which are often models with billions of parameters, is another important limitation, and prevents them from running on device. However, leveraging LLMs for data augmentation or back-

translation to further train specialized MT models may be a good way to benefit from the advantages of both LLMs and specialized models.

Regarding the speech modality, one way to leverage LLMs for speech inputs is simply to use an existing ASR model and transcribe the input speech to feed the generated text to the text-based LLM. This is a cascaded approach that works well even though suffering from some latency. However, such approach is not leveraging LLM training and methods to perform multimodal communication tasks. Other methods like Speech-LLaMA (Wu et al. 2023) learn to map speech to the continuous LLM input space and then perform low-rank adaptation of the LLM to perform speech-to-text tasks. Alternative approaches do not adapt speech to text-based LLMs, but rather try to natively perform language modeling on discrete speech units. This is the so-called textless NLP field, with audioLM (Borsos et al. 2023) being one of the state-of-the-art model of this research area. AudioLM extension for both speech and text inputs is audioPaLM (Rubenstein et al. 2023) which is fusing audio-based LLM with text-based LLM to enable a language model to take as input speech and text tokens but also output speech and text tokens and could therefore perform speech generation. Finetuned on multimodal communication tasks, such a model can perform MT, S2TT and S2ST, either directly producing the desired output or outputting intermediate steps in addition to the desired output (in a self-cascade manner, like performing ASR then MT to solve S2TT tasks) similarly to chain-of-thoughts techniques for text-based reasoning (Wei et al. 2022).

However, these LLMs applied to speech come with an important computational cost too. AudioPaLM is a 8B parameter model compared to seamlessM4T (Seamless Communication et al. 2023a) sequence-to-sequence model which outperforms audioPaLM and only has 2.3B parameters. These LLM-based solution for speech are too big to run on device. Moreover, such architectures are not suited for streaming applications of speech. For example, streaming speech translation introduced in the Seamless project (Seamless Communication et al. 2023b), could not be easily performed with an decoder-only architecture like audioPaLM.

6.3 Conclusion

To conclude, we introduced semantic representations for massively multilingual speech and text at the sentence level, enabling large-scale speech-to-text and speech-to-speech mining. Despite being fixed-size, these representations can encode a lot of information which can be efficiently decoded into different languages and modalities. Any modality-specific information can be modeled

in a disentangled additional embedding. These sentence-level representations enables efficient zero-shot cross-lingual and cross-modal transfer and pave the way to solving NLP tasks directly in sentence embedding spaces.

APPENDIX

A.1 SpeechMatrix evaluation details

We describe experimental details of speech-to-speech translation evaluations for SpeechMatrix.

A.1.1 HuBERT models

We train a multilingual HuBERT model for each language family (cf. Table 3.10). We collect unlabeled VoxPopuli speech for all languages of the same family as the training data. The HuBERT model consists of 7 convolutional layers and 12 Transformer encoder layers. Each encoder layer has 12 attention heads, the embedding dimension is 768 and the FFN dimension is 3072. Models are trained for 3 iterations. For each iteration, pseudo-labels are prepared as the training targets. In the first iteration, the target labels are based on Mel Frequency Cepstral Coefficients (MFCC) (cf. Chapter 2). In the second iteration, we extract speech features from the 6-th layer of the trained HuBERT model and apply k-means clustering to derive a set of 500 units. In the third iteration, speech features from the 9-th layer are clustered into 500 units. Lastly after these three iterations, we try feature extraction from different layers including layer 10, 11 and 12 of trained HuBERT. As for feature clustering, we also try different numbers of clusters, 800, 1000 and 1200, to derive multiple sets of target units.

To choose the optimal setup, we launch a resynthesis evaluation to select the HuBERT layer to extract speech features and the number of k-means clusters. We train a vocoder on each set of units. The synthesized speech is sent to off-the-shelf Automatic Speech Recognition (ASR) models, and Word Error Rate (WER) is reported to measure the speech quality. The resynthesis experiments are discussed in Section A.1.3. The optimal HuBERT layer and number of clusters is selected if their corresponding vocoder achieves lowest WER.

A.1.2 ASR models

We use ASR models publicly released on HuggingFace to transcribe the generated speech in order to calculate WER or BLEU scores in comparison with ground truth texts. ASR models used in our evaluation are listed in Table A.1.

Lang	cs	de
ASR	comodoro/wav2vec2-xls-r-300m-cs-250	jonatasgrosman/wav2vec2-xls-r-1b-german
Lang	**et**	**fi**
ASR	RASMUS/wav2vec2-xlsr-1b-et	jonatasgrosman/wav2vec2-large-xlsr-53-finnish
Lang	**hr**	**hu**
ASR	classla/wav2vec2-xls-r-parlaspeech-hr	jonatasgrosman/wav2vec2-large-xlsr-53-hungarian
Lang	**it**	**lt**
ASR	jonatasgrosman/wav2vec2-large-xlsr-53-italian	sammy786/wav2vec2-xlsr-lithuanian
Lang	**nl**	**pl**
ASR	jonatasgrosman/wav2vec2-xls-r-1b-dutch	jonatasgrosman/wav2vec2-xls-r-1b-polish
Lang	**pt**	**ro**
ASR	jonatasgrosman/wav2vec2-xls-r-1b-portuguese	gigant/romanian-wav2vec2
Lang	**sk**	**sl**
ASR	anuragshas/wav2vec2-xls-r-300m-sk-cv8-with-lm	anuragshas/wav2vec2-xls-r-300m-sl-cv8-with-lm

Table A.1. – HuggingFace ASR models for each language.

A.1.3 Vocoders

Data preprocessing. We applied a denoiser [1] (Defossez et al. 2020) to the speech of VoxPopuli and Common Voice as the speech preprocessing to increase Signal-to-Noise Ratio (SNR) given that they are noisier than CSS10 audios. Then we generate speech units with HuBERT models using the k-means clustering results. Single-speaker vocoders are trained in CSS10, and languages from VoxPopuli and Common Voice have multi-speaker vocoders where speaker embeddings are learned. During inference, we select the speaker with the longest speech duration to synthesize speech from predicted unit sequences.

Vocoder training and evaluation. Vocoders are trained to synthesize speech from a given sequence of units. The train sets are speech data from CSS10, VoxPopuli and Common Voice. As mentioned before, units are derived from HuBERT models. Table A.2 summarizes WER of ASR models, which reflects the transcription quality in each language. Besides, we report the training dataset, vocoder WER of synthesized speech. We include only the vocoder results obtained from the optimal HuBERT layer and k-means cluster size. Layer 11 is the best

1. https://github.com/facebookresearch/denoiser

HuBERT layer for feature extraction in all languages, and most languages have the best k-means size of 1000 except Italian (it) whose best label size is 800.

Lang	Data	ASR WER	HuBERT	Vocoder WER	Lang	Data	ASR WER	HuBERT	Vocoder WER
deu	CSS10	0.10	Germanic HuBERT layer 11, km 1000	0.16	nld	CSS10	0.19	Germanic HuBERT layer 11, km 1000	0.27
fin	CSS10	0.02	Uralic HuBERT layer 11, km 1000	0.15	hun	CSS10	0.21	Uralic HuBERT layer 11, km 1000	0.21
est	Common Voice	0.14	Uralic HuBERT layer 11, km 1000	0.44	ita	VoxPopuli	0.23	Uralic HuBERT layer 11, km 800	0.27
por	Common Voice	0.06	Uralic HuBERT layer 11, km 1000	0.31	ron	VoxPopuli	0.42	Uralic HuBERT layer 11, km 1000	0.50
ces	VoxPopuli	0.15	Slavic HuBERT layer 11, km 1000	0.23	pol	VoxPopuli	0.14	Slavic HuBERT layer 11, km 1000	0.23
hrv	VoxPopuli	0.21	Slavic HuBERT layer 11, km 1000	0.29	lit	VoxPopuli	0.38	Slavic HuBERT layer 11, km 1000	0.57
slk	VoxPopuli	0.28	Slavic HuBERT layer 11, km 1000	0.41	slv	VoxPopuli	0.37	Slavic HuBERT layer 11, km 1000	0.46

Table A.2. – Benchmark results of ASR models and vocoder resynthesis.

As shown in Table A.2, ASR models are of good quality for high-resource languages such as deu, fin and por, while suffering from high error rates in languages such as ron, lit and slv. It is expected to have higher vocoder WER than ASR WER since the former is obtained from synthesized speech. By measuring the gap between the two error rates, we can tell how good a vocoder is and also infer the quality of HuBERT units. For est, por and lit, the gaps are obviously larger than other languages. It not surprising since we do not have much good-quality vocoder data for these languages. For example, there is only around 10-hour of noisy speech from Common Voice for est and por for vocoder training.

A.1.4 Mined data selection

We performed an analysis of translation performance varying with thresholds from 1.06 to 1.09 on three language pairs: spa-eng, ron-eng and hrv-eng. Figure A.1 shows the evaluated thresholds, their corresponding speech mined data size and the resulting BLEU score.

For low-resource directions such as hrv-eng, it is best to include all the mined data. For high- and medium-resource directions, spa-eng and ron-eng, the optimal amount of mined data is around 1k hours and it does not bring further gains to go beyond that data size. Given these observations, we choose the highest threshold that keeps the source speech duration of mined data more than 1k hour for each direction. For example, we use a threshold of 1.09 for es-en and of 1.06 for hr-en.

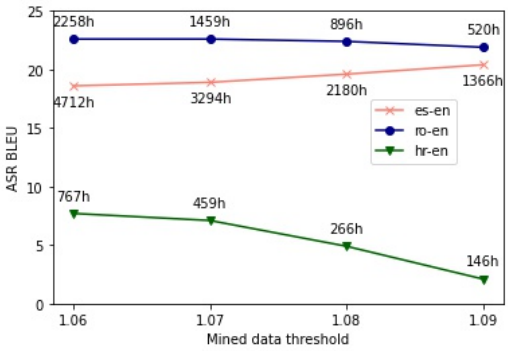

Figure A.1. – Bilingual Speech-to-Speech Translation (S2ST) BLEU by mined data at different thresholds.

A.1.5 Bilingual evaluation on FLEURS

We report here the speech-to-speech translation performance of bilingual models on the Wikipedia domain, i.e., FLEURS test data. We notice that these results are much lower than the ones on EuroparlST (EPST) and VoxPopuli data, likely because of the domain mismatch between train and test data.

	ces	deu	eng	spa	est	fin	fra	hrv	hun	ita	lit	nld	pol	por	ron	slk	slv
cs	-	2.0	4.2	4.6	0.1	0.2	7.5	2.1	0.2	2.5	0.1	1.0	2.3	2.8	1.4	3.5	1.7
de	2.3	-	8.3	3.8	0.1	0.2	6.5	2.2	0.2	1.8	0.0	1.2	0.9	3.1	2.1	0.8	1.0
en	2.7	2.7	-	6.0	0.7	0.6	10.4	2.4	0.3	3.6	0.1	3.8	1.3	5.1	2.0	1.2	1.2
es	1.9	1.8	7.5	-	0.1	0.2	9.2	1.0	0.2	4.2	0.1	1.5	1.4	5.9	2.3	0.9	0.8
et	2.1	0.7	8.2	3.0	-	0.7	6.3	1.0	0.7	2.3	0.1	1.5	1.2	1.7	1.4	0.4	0.8
fi	1.5	0.9	5.5	3.8	0.5	-	6.2	0.5	0.0	1.2	0.0	0.8	1.2	2.0	1.1	0.7	0.7
fr	1.5	2.1	9.8	7.6	0.1	0.2	-	1.7	0.2	3.1	0.1	1.3	1.5	5.8	2.4	0.6	0.6
hr	2.5	0.9	7.7	3.1	0.2	0.1	5.8	-	0.2	1.1	0.0	0.9	1.1	2.0	0.6	0.9	0.8
hu	1.3	1.0	4.6	3.0	0.1	0.2	5.7	0.7	-	1.2	0.0	0.1	0.4	2.3	0.9	0.2	0.3
it	1.3	1.0	6.3	8.3	0.1	0.1	11.3	1.3	0.2	-	0.0	0.9	1.1	5.6	1.9	0.4	0.6
lt	0.1	0.0	0.9	0.2	0.0	0.0	0.2	0.0	0.0	0.4	-	0.1	0.0	0.0	0.0	0.0	0.0
nl	1.4	3.1	5.7	4.9	0.2	0.2	7.5	1.8	0.2	1.7	0.0	-	0.9	3.3	1.4	0.4	1.0
pl	1.6	1.6	4.9	4.4	0.1	0.2	5.4	1.2	0.1	1.5	0.0	0.3	-	2.5	1.2	1.1	0.7
pt	1.2	1.0	6.1	8.7	0.1	0.3	11.1	1.1	0.1	1.1	0.1	0.6	0.8	-	1.5	0.6	0.6
ro	1.9	2.2	7.8	7.0	0.4	0.3	11.3	0.9	0.2	3.8	0.1	0.9	1.1	6.0	-	0.7	0.2
sk	9.1	2.1	5.5	5.1	0.3	0.2	7.8	3.0	0.4	2.1	0.0	0.7	1.9	2.3	1.9	-	1.5
sl	2.2	2.0	7.3	3.4	0.2	0.3	4.5	1.1	0.1	1.2	0.0	1.0	1.2	1.5	0.1	0.3	-

Table A.3. – **Mined data evaluation on FLEURS.** BLEU scores of bilingual S2ST models on FLEURS test sets.

A.2 SONAR ablation on pooling methods for speech

We compare different pooling methods to extract a fixed-size representation from speech utterances, namely mean-pooling, max-pooling and attention pooling. We focus on zero-shot Speech-to-Text Translation (S2TT) evaluation of the different speech encoders, combining, at inference time, the speech encoders with the SONAR text decoder. Best results are obtained with attention pooling (cf. Table A.4)

BLEU	fra-eng	spa-eng
SONAR mean-pooling	25.2	20.6
SONAR max-pooling	31.6	24.5
SONAR attention-pooling	33.3	25.5

Table A.4. – **Pooling methods experiments.** spBLEU X-eng zero-shot S2TT on FLEURS test set for different pooling methods.

BIBLIOGRAPHY

Akuzawa, Kei, Yusuke Iwasawa, and Yutaka Matsuo (2018). "Expressive speech synthesis via modeling expressions with variational autoencoder". In: *arXiv preprint arXiv:1804.02135* (cit. on p. 30).

Alinejad, Ashkan and Anoop Sarkar (2020). "Effectively pretraining a speech translation decoder with machine translation data". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8014–8020 (cit. on pp. 29, 65).

Ammar, Waleed, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith (2016). "Massively multilingual word embeddings". In: *arXiv preprint arXiv:1602.01925* (cit. on p. 19).

Anil, Rohan, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. (2023). "Palm 2 technical report". In: *arXiv preprint arXiv:2305.10403* (cit. on p. 28).

Antoun, Wissam, Fady Baly, and Hazem Hajj (May 2020). "AraBERT: Transformer-based Model for Arabic Language Understanding". English. In: *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. Ed. by Hend Al-Khalifa, Walid Magdy, Kareem Darwish, Tamer Elsayed, and Hamdy Mubarak. Marseille, France: European Language Resource Association, pp. 9–15. URL: https://aclanthology.org/2020.osact-1.2 (cit. on p. 13).

Ao, Junyi, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. (2021). "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing". In: *arXiv preprint arXiv:2110.07205* (cit. on p. 15).

Ardila, Rosana, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber (2020). "Common Voice: A Massively-Multilingual Speech Corpus". In: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis. European Language Resources Association, pp. 4218–4222 (cit. on pp. 38, 50, 55, 99).

Arivazhagan, Naveen, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey (2019a). "The missing ingredient in zero-shot neural machine translation". In: *arXiv preprint arXiv:1903.07091* (cit. on p. 26).

Arivazhagan, Naveen, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. (2019b). "Massively multilingual neural machine translation in the wild: Findings and challenges". In: *arXiv preprint arXiv:1907.05019* (cit. on p. 25).

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma (2017). "A simple but tough-to-beat baseline for sentence embeddings". In: *International conference on learning representations* (cit. on p. 2).

Artetxe, Mikel, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho (2017). "Unsupervised neural machine translation". In: *arxiv 1710.11041* (cit. on p. 35).

Artetxe, Mikel and Holger Schwenk (2019a). "Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL Short Papers)* (cit. on p. 38).

Artetxe, Mikel and Holger Schwenk (2019b). "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond". In: *TACL*, pp. 597–610 (cit. on pp. 2, 19, 21, 22, 88, 89, 92, 102).

Audhkhasi, Kartik, Andrew Rosenberg, Abhinav Sethy, Bhuvana Ramabhadran, and Brian Kingsbury (2017). "End-to-end ASR-free keyword search from speech". In: *IEEE Journal of Selected Topics in Signal Processing* 11.8, pp. 1351–1359 (cit. on p. 23).

Babu, Arun, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. (2021). "Xls-r: Self-supervised cross-lingual speech representation learning at scale". In: *arXiv preprint arXiv:2111.09296* (cit. on pp. xi, 14, 15, 27, 28, 49, 58, 72, 76, 82).

Baevski, Alexei, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli (2022). "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language". In: `https://arxiv.org/abs/2202.03555` (cit. on p. 15).

Baevski, Alexei, Steffen Schneider, and Michael Auli (2019). "vq-wav2vec: Self-supervised learning of discrete speech representations". In: *arXiv preprint arXiv:1910.05453* (cit. on p. 13).

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in Neural Information Processing Systems* 33, pp. 12449–12460 (cit. on pp. 11, 13, 44).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (cit. on pp. 2, 10, 24).

Bahdanau, Dzmitry, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio (2016). "End-to-end attention-based large vocabulary speech recognition". In: *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 4945–4949 (cit. on p. 27).

Bahl, Lalit R, Peter F Brown, Peter V de Souza, and Robert L Mercer (1987). "Speech recognition with continuous-parameter hidden Markov models". In: *Computer Speech & Language* 2.3-4, pp. 219–234 (cit. on p. 8).

Baker, James (1975). "The DRAGON system–An overview". In: *IEEE Transactions on Acoustics, speech, and signal Processing* 23.1, pp. 24–29 (cit. on p. 27).

Banerjee, Satanjeev and Alon Lavie (June 2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ed. by Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72. URL: https://aclanthology.org/W05-0909 (cit. on p. 26).

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza (July 2020). "ParaCrawl: Web-Scale Acquisition of Parallel Corpora". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 4555–4567. URL: https://aclanthology.org/2020.acl-main.417 (cit. on p. 20).

Bapna, Ankur, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau (2022). "mSLAM: Massively multilingual joint pre-training for speech and text". In: *arXiv preprint arXiv:2202.01374* (cit. on pp. xi, 15, 27, 76, 82).

Bapna, Ankur, Yu-an Chung, Nan Wu, Anmol Gulati, Ye Jia, Jonathan H Clark, Melvin Johnson, Jason Riesa, Alexis Conneau, and Yu Zhang (2021). "SLAM: A unified encoder for speech and language modeling via speech-text joint pre-training". In: *arXiv preprint arXiv:2110.10329* (cit. on p. 15).

Barrault, Loïc, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri (Aug. 2019). "Findings of the 2019 Conference on Machine

Translation (WMT19)". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Ed. by Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor. Florence, Italy: Association for Computational Linguistics, pp. 1–61. URL: https://aclanthology.org/W19-5301 (cit. on p. 2).

Baum, Leonard E and Ted Petrie (1966). "Statistical inference for probabilistic functions of finite state Markov chains". In: *The annals of mathematical statistics* 37.6, pp. 1554–1563 (cit. on p. 27).

Benchetrit, Yohann, Hubert Banville, and Jean-Remi King (2023). "Brain decoding: toward real-time reconstruction of visual perception". In: *arXiv preprint* (cit. on p. 123).

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult". In: *IEEE transactions on neural networks* 5.2, pp. 157–166 (cit. on p. 10).

Bérard, Alexandre, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin (2018). "End-to-end automatic speech translation of audiobooks". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6224–6228 (cit. on p. 28).

Bérard, Alexandre, Olivier Pietquin, Christophe Servan, and Laurent Besacier (2016). "Listen and translate: A proof of concept for end-to-end speech-to-text translation". In: *arXiv preprint arXiv:1612.01744* (cit. on p. 28).

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching word vectors with subword information". In: *Transactions of the association for computational linguistics* 5, pp. 135–146 (cit. on p. 9).

Bojar, Ondřej, Dominik Macháček, Sangeet Sagar, Otakar Smrž, Jonáš Kratochvíl, Ebrahim Ansari, Dario Franceschini, Chiara Canton, Ivan Simonini, Thai-Son Nguyen, Felix Schneider, Sebastian Stücker, Alex Waibel, Barry Haddow, Rico Sennrich, and Philip Williams (Nov. 2020). "ELITR: European Live Translator". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Ed. by André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada. Lisboa, Portugal: European Association for Machine Translation, pp. 463–464. URL: https://aclanthology.org/2020.eamt-1.53 (cit. on p. 23).

Booth, Andrew Donald and R. H. Richens (1952). "Some methods of mechanized translation". In: *EARLYMT*. URL: https://api.semanticscholar.org/CorpusID:34140950 (cit. on p. 1).

Borsos, Zalán, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. (2023). "Audiolm: a language modeling approach to audio generation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (cit. on pp. 28, 31, 126).

Boukkouri, Hicham El, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Junichi Tsujii (2020). "CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters". In: *arXiv preprint arXiv:2010.10392* (cit. on p. 121).

Bourlard, H. and N. Morgan (1993). "Continuous speech recognition by connectionist statistical methods". In: *IEEE Transactions on Neural Networks* 4.6, pp. 893–909 (cit. on p. 27).

Bowman, Samuel R, Gabor Angeli, Christopher Potts, and Christopher D Manning (2015). "A large annotated corpus for learning natural language inference". In: *arXiv preprint arXiv:1508.05326* (cit. on p. 16).

Brown, Peter F, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin (1990). "A statistical approach to machine translation". In: *Computational linguistics* 16.2, pp. 79–85 (cit. on pp. 1, 8).

Buck, Christian and Philipp Koehn (Aug. 2016). "Findings of the WMT 2016 Bilingual Document Alignment Shared Task". In: *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, pp. 554–563. URL: http://www.aclweb.org/anthology/W/W16/W16-2347 (cit. on p. 20).

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn (Apr. 2006). "Re-evaluating the Role of Bleu in Machine Translation Research". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Diana McCarthy and Shuly Wintner. Trento, Italy: Association for Computational Linguistics, pp. 249–256. URL: https://aclanthology.org/E06-1032 (cit. on p. 26).

Casanova, Edresson, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Golge, 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*. PMLR, pp. 2709–2720 (cit. on p. 30).

Cauchy, Augustin et al. (1847). "Méthode générale pour la résolution des systemes d'équations simultanées". In: *Comp. Rend. Sci. Paris* 25.1847, pp. 536–538 (cit. on p. 7).

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al.

(2018). "Universal sentence encoder". In: *arXiv preprint arXiv:1803.11175* (cit. on p. 16).

Chan, William, Navdeep Jaitly, Quoc Le, and Oriol Vinyals (2016). "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition". In: *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 4960–4964 (cit. on p. 27).

Chen, Leiyu, Shaobo Li, Qiang Bai, Jing Yang, Sanlong Jiang, and Yanming Miao (2021). "Review of image classification algorithms based on convolutional neural networks". In: *Remote Sensing* 13.22, p. 4712 (cit. on p. 8).

Chen, Mingda, Kevin Heffernan, Onur Çelebi, Alex Mourachko, and Holger Schwenk (2023). "xSIM++: An Improved Proxy to Bitext Mining Performance for Low-Resource Languages". In: *arXiv preprint arXiv:2306.12907* (cit. on pp. 22, 91, 99).

Chen, Sanyuan, 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". In: *IEEE J. Sel. Top. Signal Process.* 16.6, pp. 1505–1518 (cit. on p. 110).

Chidambaram, Muthuraman, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil (2018). "Learning cross-lingual sentence representations via a multi-task dual-encoder model". In: *arXiv preprint arXiv:1810.12836* (cit. on p. 19).

Chiu, Chung-Cheng, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu (2022). "Self-supervised learning with random-projection quantizer for speech recognition". In: *International Conference on Machine Learning*. PMLR, pp. 3915–3924 (cit. on p. 14).

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078* (cit. on pp. 2, 24).

Chuang, Yung-Sung, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass (2022). "DiffCSE: Difference-based contrastive learning for sentence embeddings". In: *arXiv preprint arXiv:2204.10298* (cit. on p. 17).

Chung, Yu-An, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu (2021). "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training". In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 244–250 (cit. on pp. 14, 98, 99).

Chung, Yu-An, Chenguang Zhu, and Michael Zeng (2020). "Splat: Speech-language joint pre-training for spoken language understanding". In: *arXiv preprint arXiv:2010.02295* (cit. on p. 15).

Chung, Yu-An, Chao-Chung Wu, Chia-Hao Shen, Hung-yi Lee, and Lin-Shan Lee (2016). "Audio Word2Vec: Unsupervised Learning of Audio Segment Representations using Sequence-to-sequence Autoencoder". In: *CoRR* abs/1603.00982. arXiv: 1603.00982. URL: http://arxiv.org/abs/1603.00982 (cit. on p. 22).

Church, Kenneth and Patrick Hanks (1990). "Word association norms, mutual information, and lexicography". In: *Computational linguistics* 16.1, pp. 22–29 (cit. on p. 9).

Clark, Jonathan H, Dan Garrette, Iulia Turc, and John Wieting (2022). "Canine: Pre-training an efficient tokenization-free encoder for language representation". In: *Transactions of the Association for Computational Linguistics* 10, pp. 73–91 (cit. on p. 121).

Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning (2020). "Electra: Pre-training text encoders as discriminators rather than generators". In: *arXiv preprint arXiv:2003.10555* (cit. on p. 13).

Conneau, Alexis, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli (2020a). "Unsupervised cross-lingual representation learning for speech recognition". In: *arXiv preprint arXiv:2006.13979* (cit. on p. 14).

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020b). "Unsupervised Cross-lingual Representation Learning at Scale". In: *ACL* (cit. on pp. 2, 13).

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes (2017). "Supervised learning of universal sentence representations from natural language inference data". In: *arXiv preprint arXiv:1705.02364* (cit. on p. 16).

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni (2018a). "What you can cram into a single vector: Probing sentence embeddings for linguistic properties". In: *arXiv preprint arXiv:1805.01070* (cit. on p. 122).

Conneau, Alexis and Guillaume Lample (2019). "Cross-lingual language model pretraining". In: *Advances in neural information processing systems* 32 (cit. on pp. 13, 15, 19, 39).

Conneau, Alexis, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov (2018b). "XNLI: Evaluating cross-lingual sentence representations". In: *arXiv preprint arXiv:1809.05053* (cit. on p. 13).

Conneau, Alexis, 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)*. IEEE, pp. 798–805 (cit. on pp. 54, 98, 109, 124).

Dahl, George E, Dong Yu, Li Deng, and Alex Acero (2011). "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition". In: *IEEE Transactions on audio, speech, and language processing* 20.1, pp. 30–42 (cit. on p. 27).

Defossez, Alexandre, Gabriel Synnaeve, and Yossi Adi (2020). "Real Time Speech Enhancement in the Waveform Domain". In: *Interspeech* (cit. on p. 130).

Défossez, Alexandre, Jade Copet, Gabriel Synnaeve, and Yossi Adi (2022). "High fidelity neural audio compression". In: *arXiv preprint arXiv:2210.13438* (cit. on pp. 30, 88, 104).

Devlin, Jacob, 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)*, pp. 4171–4186 (cit. on pp. 2, 12, 14, 38).

Di Gangi, Mattia A, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi (2019). "Must-c: a multilingual speech translation corpus". In: *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 2012–2017 (cit. on pp. 28, 46, 50, 99).

Diderot, Denis and Jean le Rond d'Alembert (1751). *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers* (cit. on p. 1).

Dinh, Tu Anh (2021). "Zero-shot Speech Translation". In: *arXiv preprint arXiv:2107.06010* (cit. on p. 29).

Dinh, Tu Anh, Danni Liu, and Jan Niehues (2022). "Tackling data scarcity in speech translation using zero-shot multilingual machine translation techniques". In: *arXiv preprint arXiv:2201.11172* (cit. on p. 29).

Do, Quoc Truong, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura (2015). "Improving translation of emphasis with pause prediction in speech-to-speech translation systems". In: *Proceedings of the 12th International Workshop on Spoken Language Translation: Papers, IWSLT 2015, Da Nang, Vietnam, December 3-4, 2015*. Ed. by Marcello Federico, Sebastian Stüker, and Jan Niehues (cit. on p. 29).

Dong, Linhao, Shuang Xu, and Bo Xu (2018). "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition". In: *2018 IEEE*

international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp. 5884–5888 (cit. on p. 27).

Dong, Qianqian, 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". In: *arXiv preprint arXiv:2306.02982* (cit. on p. 31).

Dong, Qianqian, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li (2021). "Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 14, pp. 12749–12759 (cit. on pp. 29, 65).

Dong, Qianqian, Fengpeng Yue, Tom Ko, Mingxuan Wang, Qibing Bai, and Yu Zhang (2022). "Leveraging pseudo-labeled data to improve direct speech-to-speech translation". In: *arXiv preprint arXiv:2205.08993* (cit. on p. 116).

Dufter, Philipp, Mengjie Zhao, Martin Schmitt, Alexander Fraser, and Hinrich Schütze (July 2018). "Embedding Learning Through Multilingual Concept Induction". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 1520–1530. URL: https://aclanthology.org/P18-1141 (cit. on p. 18).

Duquenne, Paul-Ambroise, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswami, Changhan Wang, Juan Pino, Benoit Sagot, and Holger Schwenk (July 2023a). "SpeechMatrix: A Large-Scale Mined Corpus of Multilingual Speech-to-Speech Translations". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 16251–16269. URL: https://aclanthology.org/2023.acl-long.899 (cit. on pp. 4, 33).

Duquenne, Paul-Ambroise, Hongyu Gong, Benoit Sagot, and Holger Schwenk (Dec. 2022). "T-Modules: Translation Modules for Zero-Shot Cross-Modal Machine Translation". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5794–5806. URL: https://aclanthology.org/2022.emnlp-main.391 (cit. on pp. 5, 63).

Duquenne, Paul-Ambroise, Hongyu Gong, and Holger Schwenk (2021). "Multimodal and multilingual embeddings for large-scale speech mining". In: *Advances in Neural Information Processing Systems* 34 (cit. on pp. 4, 23, 28, 33).

Duquenne, Paul-Ambroise, Kevin Heffernan, Alexandre Mourachko, Benoît Sagot, and Holger Schwenk (2023b). *SONAR EXPRESSIVE: Zero-shot Expressive Speech-to-Speech Translation* (cit. on pp. 6, 86).

Duquenne, Paul-Ambroise, Holger Schwenk, and Benoit Sagot (2023c). "Modular Speech-to-Text Translation for Zero-Shot Cross-Modal Transfer". In: *Proc. INTERSPEECH 2023*, pp. 32–36 (cit. on pp. 5, 64).

Duquenne, Paul-Ambroise, Holger Schwenk, and Benoit Sagot (2023d). *SONAR: Sentence-Level Multimodal and Language-Agnostic Representations*. URL: https://arxiv.org/abs/2308.11466 (cit. on pp. 6, 86).

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier (2018). "Understanding back-translation at scale". In: *arXiv preprint arXiv:1808.09381* (cit. on pp. 25, 115).

Escolano, Carlos, Marta R Costa-Jussà, and José AR Fonollosa (2021a). "From bilingual to multilingual neural-based machine translation by incremental training". In: *Journal of the Association for Information Science and Technology* 72.2, pp. 190–203 (cit. on pp. 26, 29).

Escolano, Carlos, Marta R Costa-jussà, José AR Fonollosa, and Mikel Artetxe (2020a). "Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders". In: *arXiv preprint arXiv:2004.06575* (cit. on pp. 26, 29).

Escolano, Carlos, Marta R Costa-jussà, José AR Fonollosa, and Mikel Artetxe (2020b). "Training multilingual machine translation by alternately freezing language-specific encoders-decoders". In: *arXiv preprint arXiv:2006.01594* (cit. on pp. 26, 29).

Escolano, Carlos, Marta R Costa-jussà, José AR Fonollosa, and Carlos Segura (2021b). "Enabling zero-shot multilingual spoken language translation with language-specific encoders and decoders". In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 694–701 (cit. on pp. xi, 29, 77).

España-Bonet, Cristina, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith (2017). "An Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification". In: *IEEE Journal of Selected Topics in Signal Processing*, pp. 1340–1348 (cit. on pp. 19, 21).

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. (2021). "Beyond english-centric multilingual machine translation". In: *Journal of Machine Learning Research* 22.107, pp. 1–48 (cit. on pp. xi, 21, 25, 26, 35, 74, 75, 80).

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang (2020). "Language-agnostic bert sentence embedding". In: *arXiv preprint arXiv:2007.01852* (cit. on pp. 2, 19, 21, 22, 36, 37, 90).

Firth, John Rupert (1957). "A synopsis of linguistic theory, 1930-1955". In: *Studies in Linguistic Analysis* (cit. on p. 9).

Fung, Pascale and Percy Cheung (July 2004). "Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and E". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Ed. by Dekang Lin and Dekai Wu. Barcelona, Spain: Association for Computational Linguistics, pp. 57–63. URL: https://aclanthology.org/W04-3208 (cit. on p. 20).

Gafni, Oran, 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*. Springer, pp. 89–106 (cit. on pp. 108, 111).

Gao, Tianyu, Xingcheng Yao, and Danqi Chen (2021). "Simcse: Simple contrastive learning of sentence embeddings". In: *arXiv preprint arXiv:2104.08821* (cit. on p. 17).

Giorgi, John, Osvald Nitski, Bo Wang, and Gary Bader (2020). "Declutr: Deep contrastive learning for unsupervised textual representations". In: *arXiv preprint arXiv:2006.03659* (cit. on p. 17).

Gouws, Stephan, Yoshua Bengio, and Greg Corrado (2015). "Bilbowa: Fast bilingual distributed representations without word alignments". In: *International Conference on Machine Learning*. PMLR, pp. 748–756 (cit. on p. 18).

Goyal, Naman, 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". In: *Transactions of the Association for Computational Linguistics* 10. Ed. by Brian Roark and Ani Nenkova, pp. 522–538. URL: https://aclanthology.org/2022.tacl-1.30 (cit. on pp. 22, 54, 66, 80).

Graves, Alex and Navdeep Jaitly (2014). "Towards end-to-end speech recognition with recurrent neural networks". In: *International conference on machine learning*. PMLR, pp. 1764–1772 (cit. on p. 27).

Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. (2020). "Conformer: Convolution-augmented transformer for speech recognition". In: *arXiv preprint arXiv:2005.08100* (cit. on p. 11).

Guo, Mandy, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil (2018). "Effective Parallel Corpus Mining using Bilingual Sentence Embeddings". In: *arXiv:1807.11906* (cit. on pp. 19, 21, 43).

Harris, Zellig S (1954). "Distributional structure". In: *Word* 10.2-3, pp. 146–162 (cit. on p. 9).

Harwath, David, Wei-Ning Hsu, and James Glass (2019). "Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech". In: *International Conference on Learning Representations* (cit. on p. 23).

Harwath, David, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass (2018). "Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input". In: *European Conference on Computer Vision*. Springer, pp. 659–677 (cit. on p. 23).

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou (2018). "Achieving Human Parity on Automatic Chinese to English News Translation". In: *arXiv:1803.05567* (cit. on p. 21).

Heffernan, Kevin, Onur Çelebi, and Holger Schwenk (2022). "Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (cit. on pp. 4, 19, 50, 71, 90).

Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. (2012). "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". In: *IEEE Signal processing magazine* 29.6, pp. 82–97 (cit. on p. 27).

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33, pp. 6840–6851 (cit. on p. 124).

Hochreiter, Sepp (Apr. 1991). "Untersuchungen zu dynamischen neuronalen Netzen". In: (cit. on p. 10).

Hochreiter, Sepp and Jurgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780 (cit. on p. 10).

Holzenberger, Nils, Mingxing Du, Julien Karadayi, Rachid Riad, and Emmanuel Dupoux (2018). "Learning Word Embeddings: Unsupervised Methods for Fixed-size Representations of Variable-length Speech Segments". In: *INTERSPEECH* (cit. on p. 22).

Hsu, Wei-Ning, 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". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, pp. 3451–3460 (cit. on p. 14).

Hsu, Wei-Ning, 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". In: *arXiv preprint arXiv:1810.07217* (cit. on p. 30).

Huang, Qingqing, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis (2022). "Mulan: A joint embedding of music audio and natural language". In: *arXiv preprint arXiv:2208.12415* (cit. on p. 23).

Huang, Wen-Chin, 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)*, pp. 1–5 (cit. on p. 110).

Ilharco, Gabriel, Yuan Zhang, and Jason Baldridge (2019). "Large-scale representation learning from visually grounded untranscribed speech". In: *arXiv preprint arXiv:1909.08782* (cit. on p. 23).

Iranzo-Sánchez, Javier, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan (2020). "Europarl-ST: A multilingual corpus for speech translation of parliamentary debates". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 8229–8233 (cit. on pp. 28, 45, 54).

Jegou, Herve, Matthijs Douze, and Cordelia Schmid (2010). "Product quantization for nearest neighbor search". In: *IEEE transactions on pattern analysis and machine intelligence* 33.1, pp. 117–128 (cit. on p. 21).

Jia, Ye, 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)*. IEEE, pp. 7180–7184 (cit. on p. 116).

Jia, Ye, 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*. PMLR, pp. 10120–10134 (cit. on pp. 29, 31).

Jia, Ye, 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". In: *arXiv preprint arXiv:1904.06037* (cit. on pp. 29, 31, 47).

Jia, Ye, 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 multispeaker text-to-speech synthesis". In: *Advances in neural information processing systems* 31 (cit. on p. 30).

Jiang, Dongwei, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li (2020). "Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning". In: *arXiv preprint arXiv:2010.13991* (cit. on p. 23).

Johnson, Jeff, Matthijs Douze, and Hervé Jégou (2019). "Billion-scale similarity search with GPUs". In: *IEEE Transactions on Big Data* 7.3, pp. 535–547 (cit. on pp. 21, 43).

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy (2020). "Spanbert: Improving pre-training by representing and predicting spans". In: *Transactions of the association for computational linguistics* 8, pp. 64–77 (cit. on p. 13).

Kahn, Jacob, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux (2019). "Libri-Light: A Benchmark for ASR with Limited or No Supervision". In: *CoRR* abs/1912.07875. arXiv: 1912.07875. URL: http://arxiv.org/abs/1912.07875 (cit. on p. 42).

Kalchbrenner, Nal and Phil Blunsom (Oct. 2013). "Recurrent Continuous Translation Models". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Ed. by David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1700–1709. URL: https://aclanthology.org/D13-1176 (cit. on p. 24).

Karpukhin, Vladimir, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih (2020). "Dense passage retrieval for open-domain question answering". In: *arXiv preprint arXiv:2004.04906* (cit. on p. 121).

Kharitonov, Eugene, 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". In: *arXiv preprint arXiv:2302.03540* (cit. on p. 30).

Khurana, Sameer, Antoine Laurent, and James Glass (2020). "Cstnet: Contrastive speech translation network for self-supervised speech representation learning". In: *arXiv preprint arXiv:2006.02814* (cit. on p. 23).

Khurana, Sameer, Antoine Laurent, and James Glass (2022). "SAMU-XLSR: Semantically-Aligned Multimodal Utterance-level Cross-Lingual Speech Representation". In: *arXiv preprint arXiv:2205.08180* (cit. on p. 23).

Kim, Jaehyeon, 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. PMLR, pp. 5530–5540 (cit. on p. 30).

Kiros, Ryan, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Skip-thought vectors". In: *Advances in neural information processing systems* 28 (cit. on pp. 2, 16).

Klementiev, Alexandre, Ivan Titov, and Binod Bhattarai (Dec. 2012). "Inducing Crosslingual Distributed Representations of Words". In: *Proceedings of COLING 2012*. Ed. by Martin Kay and Christian Boitet. Mumbai, India: The COLING 2012 Organizing Committee, pp. 1459–1474. URL: https://aclanthology.org/C12-1089 (cit. on p. 18).

Koehn, Philipp (2005). "Europarl: A Parallel Corpus for Statistical Machine Translation". In: *MT summit* (cit. on p. 25).

Koehn, Philipp, Franz J. Och, and Daniel Marcu (2003). "Statistical Phrase-Based Translation". In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127–133. URL: https://aclanthology.org/N03-1017 (cit. on p. 24).

Kreuk, Felix, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi (2022). "Audiogen: Textually guided audio generation". In: *arXiv preprint arXiv:2209.15352* (cit. on pp. 108, 111).

Kudo, Taku and John Richardson (2018). "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 66–71. URL: https://www.aclweb.org/anthology/D18-2012 (cit. on p. 121).

Kvapilíková, Ivana, Mikel Artetxe, Gorka Labaka amd Eneko Agirre, and Ondřej Bojar (2020). "Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL Short Papers)* (cit. on p. 19).

Lakhotia, Kushal, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. (2021). "On generative spoken language modeling from raw audio". In: *Transactions of the Association for Computational Linguistics* 9, pp. 1336–1354 (cit. on p. 14).

Lample, Guillaume, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato (2018). "Phrase-Based & Neural Unsupervised Machine Translation". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5039–5049 (cit. on p. 35).

Lang, Kevin J., Alex H. Waibel, and Geoffrey E. Hinton (1990). "A time-delay neural network architecture for isolated word recognition". In: *Neural Networks*

3.1, pp. 23–43. URL: https://www.sciencedirect.com/science/article/pii/089360809090044L (cit. on p. 11).

Le, Matthew, 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". In: *CoRR* abs/2306.15687. URL: https://doi.org/10.48550/arXiv.2306.15687 (cit. on pp. 30, 110).

LeCun, Yann et al. (1989). "Generalization and network design strategies". In: *Connectionism in perspective* 19.143-155, p. 18 (cit. on p. 11).

Lee, Ann, 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 (2022a). "Direct speech-to-speech translation with discrete units". In: *ACL*, pp. 3327–3339. URL: https://aclanthology.org/2022.acl-long.235 (cit. on pp. 29, 56).

Lee, Ann, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu (2022b). "Textless Speech-to-Speech Translation on Real Data". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 860–872. URL: https://aclanthology.org/2022.naacl-main.63 (cit. on pp. 48, 55, 73, 78).

Lepikhin, Dmitry, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen (2021). "GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021* (cit. on p. 58).

Li, Bohan, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li (2020). "On the sentence embeddings from pre-trained language models". In: *arXiv preprint arXiv:2011.05864* (cit. on p. 17).

Li, Naihan, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu (2019). "Neural speech synthesis with transformer network". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 6706–6713 (cit. on p. 29).

Li, Xian, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli (2020). "Multilingual speech translation with efficient finetuning of pretrained models". In: *arXiv preprint arXiv:2010.12829* (cit. on pp. 28, 29, 44).

Liao, Junwei, Yu Shi, Ming Gong, Linjun Shou, Hong Qu, and Michael Zeng (2021). "Improving Zero-shot Neural Machine Translation on Language-specific Encoders-Decoders". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8 (cit. on p. 26).

Lison, P. and J. Tiedemann (2016). "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles". In: *LREC* (cit. on p. 25).

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer (2020a). "Multilingual denoising pre-training for neural machine translation". In: *Transactions of the Association for Computational Linguistics* 8, pp. 726–742 (cit. on pp. 25, 44, 90).

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (cit. on p. 12).

Liu, Yuchen, Junnan Zhu, Jiajun Zhang, and Chengqing Zong (2020b). "Bridging the modality gap for speech-to-text translation". In: *arXiv preprint arXiv:2010.14920* (cit. on pp. 3, 29, 65).

Lu, Yichao, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun (2018). "A neural interlingua for multilingual machine translation". In: *arXiv preprint arXiv:1804.08198* (cit. on p. 26).

Luong, Thang, Hieu Pham, and Christopher D. Manning (June 2015). "Bilingual Word Representations with Monolingual Quality in Mind". In: *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Ed. by Phil Blunsom, Shay Cohen, Paramveer Dhillon, and Percy Liang. Denver, Colorado: Association for Computational Linguistics, pp. 151–159. URL: https://aclanthology.org/W15-1521 (cit. on p. 18).

Maas, Andrew, Ziang Xie, Dan Jurafsky, and Andrew Ng (May 2015). "Lexicon-Free Conversational Speech Recognition with Neural Networks". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Rada Mihalcea, Joyce Chai, and Anoop Sarkar. Denver, Colorado: Association for Computational Linguistics, pp. 345–354. URL: https://aclanthology.org/N15-1038 (cit. on p. 27).

Martin, Louis, Angela Fan, Eric De La Clergerie, Antoine Bordes, and Benoit Sagot (2020). "MUSS: multilingual unsupervised sentence simplification by mining paraphrases". In: *arXiv preprint arXiv:2005.00352* (cit. on p. 3).

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoit Sagot (2019). "CamemBERT: a tasty French language model". In: *arXiv preprint arXiv:1911.03894* (cit. on pp. 4, 13).

Merkx, Danny, Stefan L Frank, and Mirjam Ernestus (2019). "Language learning using speech to image retrieval". In: *arXiv preprint arXiv:1909.03795* (cit. on p. 23).

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (cit. on pp. 2, 9).

Mohamed, Abdel-rahman, George Dahl, Geoffrey Hinton, et al. (2009). "Deep belief networks for phone recognition". In: *Nips workshop on deep learning for speech recognition and related applications*. Vol. 1. 9. Vancouver, Canada, p. 39 (cit. on p. 27).

Monfort, Mathew, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva (2021). "Spoken moments: Learning joint audio-visual representations from video descriptions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14871–14881 (cit. on p. 23).

Munteanu, Dragos Stefan and Daniel Marcu (2005). "Improving Machine Translation Performance by Exploiting Non-Parallel Corpora". In: *Computational Linguistics* 31.4, pp. 477–504. URL: http://www.aclweb.org/anthology/J05-4003 (cit. on p. 20).

Nakamura, Satoshi, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, Jinsong Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto (2006). "The ATR multilingual speech-to-speech translation system". In: *IEEE Trans. Speech Audio Process.* 14.2, pp. 365–376 (cit. on p. 29).

Neelakantan, Arvind, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. (2022). "Text and code embeddings by contrastive pre-training". In: *arXiv preprint arXiv:2201.10005* (cit. on p. 17).

Nguyen, Toan Q and David Chiang (2017). "Transfer learning across low-resource, related languages for neural machine translation". In: *arXiv preprint arXiv:1708.09803* (cit. on p. 25).

Nguyen, Tu Anh, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. (2023). "Generative spoken dialogue language modeling". In: *Transactions of the Association for Computational Linguistics* 11, pp. 250–266 (cit. on p. 121).

Ni, Jianmo, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang (2021). "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models". In: *arXiv preprint arXiv:2108.08877* (cit. on p. 17).

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*. URL: https://arxiv.org/abs/2207.04672 (cit. on pp. 21, 22, 58, 92, 102, 116).

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. (2022). "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems* 35, pp. 27730–27744 (cit. on p. 125).

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur (2015). "Librispeech: an asr corpus based on public domain audio books". In: *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 5206–5210 (cit. on p. 99).

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (July 2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. URL: https://aclanthology.org/P02-1040 (cit. on p. 26).

Park, Daniel S, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le (2019). "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition". In: (cit. on p. 37).

Park, Kyubyong and Thomas Mulc (2019). "CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages". In: *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*. Ed. by Gernot Kubin and Zdravko Kacic. ISCA, pp. 1566–1570 (cit. on p. 55).

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543 (cit. on pp. 9, 17).

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). *Deep contextualized word representations*. arXiv: 1802.05365 [cs.CL] (cit. on p. 12).

Pham, Ngoc-Quan, Jan Niehues, Thanh-Le Ha, and Alex Waibel (2019). "Improving zero-shot translation with language-independent constraints". In: *arXiv preprint arXiv:1906.08584* (cit. on p. 26).

Pino, Juan, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang (2020). "Self-training for end-to-end speech translation". In: *arXiv preprint arXiv:2006.02490* (cit. on pp. 28, 116).

Polyak, Adam, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux (2021). "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations". In: *Proc. Interspeech 2021* (cit. on pp. 29, 55).

Popović, Maja (Sept. 2015). "chrF: character n-gram F-score for automatic MT evaluation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395. URL: https://aclanthology.org/W15-3049 (cit. on p. 26).

Popuri, Sravya, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee (2022). "Enhanced Direct Speech-to-Speech Translation Using Self-supervised Pre-training and Data Augmentation". In: pp. 5195–5199 (cit. on p. 58).

Post, Matt, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur (2013). "Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus". In: *Proc. IWSLT* (cit. on pp. 30, 47).

Pratap, Vineel, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. (2023). "Scaling speech technology to 1,000+ languages". In: *arXiv preprint arXiv:2305.13516* (cit. on pp. 14, 29).

Pratap, Vineel, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert (Oct. 2020). "MLS: A Large-Scale Multilingual Dataset for Speech Research". In: *Interspeech 2020*. URL: http://dx.doi.org/10.21437/Interspeech.2020-2826 (cit. on p. 42).

Qu, Leyuan, Taihao Li, Cornelius Weber, Theresa Pekarek-Rosin, Fuji Ren, and Stefan Wermter (2023). "Disentangling prosody representations with unsupervised speech reconstruction". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (cit. on p. 30).

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. (2021). "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR, pp. 8748–8763 (cit. on p. 123).

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever (2023). "Robust speech recognition via large-scale weak supervision". In: *International Conference on Machine Learning*. PMLR, pp. 28492–28518 (cit. on pp. xi, 27, 28, 82, 102).

Ramesh, Gowtham, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee,

Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra (2022). "Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages". In: *Transactions of the Association for Computational Linguistics* 10. Ed. by Brian Roark and Ani Nenkova, pp. 145–162. URL: https://aclanthology.org/2022.tacl-1.9 (cit. on pp. 2, 21).

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie (2020). "COMET: A neural framework for MT evaluation". In: *arXiv preprint arXiv:2009.09025* (cit. on pp. 26, 27).

Reimers, Nils and Iryna Gurevych (2019). "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (cit. on pp. 16, 17, 19, 92).

Reimers, Nils and Iryna Gurevych (2020). "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–4525 (cit. on pp. 19, 69, 90).

Ren, Yi, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu (2020). "Fastspeech 2: Fast and high-quality end-to-end text to speech". In: *arXiv preprint arXiv:2006.04558* (cit. on pp. 29, 30).

Resnik, Philip (1999). "Mining the Web for Bilingual Text". In: *ACL*. URL: http://www.aclweb.org/anthology/P99-1068 (cit. on p. 20).

Resnik, Philip and Noah A. Smith (2003). "The Web as a Parallel Corpus". In: *Computational Linguistics* 29.3, pp. 349–380. URL: http://www.aclweb.org/anthology/J03-3002 (cit. on p. 20).

Riad, Rachid, Corentin Dancette, Julien Karadayi, Neil Zeghidour, Thomas Schatz, and Emmanuel Dupoux (2018). "Sampling strategies in siamese networks for unsupervised speech representation learning". In: *arXiv preprint arXiv:1804.11297* (cit. on p. 23).

Rubenstein, Paul K, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. (2023). "AudioPaLM: A Large Language Model That Can Speak and Listen". In: *arXiv preprint arXiv:2306.12925* (cit. on pp. 28, 31, 126).

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard (2019). "A survey of cross-lingual word embedding models". In: *Journal of Artificial Intelligence Research* 65, pp. 569–631 (cit. on p. 18).

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536 (cit. on pp. 7, 10).

Salesky, Elizabeth, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post (2021). "The Multilingual TEDx Corpus for Speech Recognition and Translation". In: *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*. Ed. by Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek. ISCA, pp. 3655–3659 (cit. on p. 50).

Schneider, Steffen, Alexei Baevski, Ronan Collobert, and Michael Auli (2019). "wav2vec: Unsupervised pre-training for speech recognition". In: *arXiv preprint arXiv:1904.05862* (cit. on p. 13).

Schwenk, Holger (2009). "Investigations on large-scale lightly-supervised training for statistical machine translation". In: *IWSLT*, pp. 182–189 (cit. on p. 115).

Schwenk, Holger (2018). "Filtering and Mining Parallel Data in a Joint Multilingual Space". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL Short Papers)*, pp. 228–234 (cit. on p. 18).

Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán (2019). "WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia". In: *http://arxiv.org/abs/1907.05791* (cit. on pp. 3, 21).

Schwenk, Holger and Matthijs Douze (2017). "Learning Joint Multilingual Sentence Representations with Neural Machine Translation". In: *ACL workshop on Representation Learning for NLP* (cit. on pp. 18, 19).

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan (2021). "CCMatrix: Mining billions of high-quality parallel sentences on the web". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL Short Papers)*, pp. 6490–6500 (cit. on pp. 2, 3, 21, 42, 43, 71).

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*. URL: https://arxiv.org/abs/2308.11596 (cit. on pp. xiii, 4, 27, 28, 58, 102, 103, 126).

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* (cit. on pp. 4, 23, 108–111, 116, 126).

Sellam, Thibault, Dipanjan Das, and Ankur P Parikh (2020). "BLEURT: Learning robust metrics for text generation". In: *arXiv preprint arXiv:2004.04696* (cit. on p. 26).

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2015). "Improving neural machine translation models with monolingual data". In: *arXiv preprint arXiv:1511.06709* (cit. on p. 115).

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725 (cit. on p. 121).

Sennrich, Rico and Martin Volk (2010). "MT-based sentence alignment for OCR-generated parallel texts". In: (cit. on p. 20).

Settle, Shane and Karen Livescu (2016). "Discriminative acoustic word embeddings: Recurrent neural network-based approaches". In: *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 503–510 (cit. on p. 23).

Shen, Kai, 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". In: *arXiv preprint arXiv:2304.09116* (cit. on p. 30).

Shih, Yi-Jen, Hsuan-Fu Wang, Heng-Jui Chang, Layne Berry, Hung-yi Lee, and David Harwath (2023). "Speechclip: Integrating speech with pre-trained vision

and language model". In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 715–722 (cit. on p. 23).

Silero-Team (2021). *Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier*. https://github.com/snakers4/silero-vad (cit. on pp. 52, 111).

Stentiford, Fred and M.G. Steer (Apr. 1988). "MACHINE TRANSLATION OF SPEECH." In: 6, pp. 116–123 (cit. on p. 27).

Stevens, Stanley S and John Volkmann (1940). "The relation of pitch to frequency: A revised scale". In: *The American Journal of Psychology* 53.3, pp. 329–353 (cit. on p. 11).

Su, Jianlin, Jiarun Cao, Weijie Liu, and Yangyiwen Ou (2021). "Whitening sentence representations for better semantics and faster retrieval". In: *arXiv preprint arXiv:2103.15316* (cit. on p. 17).

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems* 27 (cit. on pp. 18, 24).

Tan, Weiting, Kevin Heffernan, Holger Schwenk, and Philipp Koehn (2022). "Multilingual Representation Distillation with Contrastive Learning". In: *arXiv preprint arXiv:2210.05033* (cit. on p. 20).

Tang, Yun, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, et al. (2022). "Unified speech-text pre-training for speech translation and recognition". In: *arXiv preprint arXiv:2204.05409* (cit. on p. 15).

Thakur, Nandan, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych (2021). *Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks*. arXiv: 2010.08240 [cs.CL] (cit. on p. 17).

Thiolliere, Roland, Ewan Dunbar, Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux (2015). "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling". In: *Sixteenth Annual Conference of the International Speech Communication Association* (cit. on p. 23).

Thompson, Brian and Philipp Koehn (2019). "Vecalign: Improved sentence alignment in linear time and space". In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 1342–1348 (cit. on p. 21).

Tiedemann, J. (2012). "Parallel Data, Tools and Interfaces in OPUS". In: *LREC* (cit. on pp. 25, 44).

Tiyajamorn, Nattapong, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka (2021). "Language-Agnostic Representation from Multilingual Sentence Encoders for Cross-Lingual Similarity Estimation". In: *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing, pp. 7764–7774 (cit. on p. 20).

Tjandra, Andros, Sakriani Sakti, and Satoshi Nakamura (2019). "Speech-to-speech translation between untranscribed unknown languages". In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 593–600 (cit. on p. 29).

Touvron, Hugo, 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". In: *arXiv preprint arXiv:2307.09288* (cit. on pp. 116, 125).

Tsai, Yao-Hung Hubert, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov (2019). "Multimodal transformer for unaligned multimodal language sequences". In: *Proceedings of the conference. Association for Computational Linguistics. Meeting*. Vol. 2019. NIH Public Access, p. 6558 (cit. on p. 23).

Tsiamas, Ioannis, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà (2022). "Shas: Approaching optimal segmentation for end-to-end speech translation". In: *arXiv preprint arXiv:2202.04774* (cit. on p. 108).

Utiyama, Masao and Hitoshi Isahara (2003). "Reliable Measures for Aligning Japanese-English News Articles and Sentences". In: *ACL*. URL: http://www.aclweb.org/anthology/P03-1010 (cit. on p. 20).

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón (2007). "Parallel corpora for medium density languages". In: *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4* 292, p. 247 (cit. on p. 20).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *NIPS*, pp. 6000–6010 (cit. on pp. 2, 3, 10).

Vázquez, Raúl, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz (2018). "Multilingual NMT with a language-independent attention bridge". In: *arXiv preprint arXiv:1811.00498* (cit. on p. 26).

Waibel, Alexander, Toshiyuki Hanazawa, G. Hinton, Kiyohiro Shikano, and K.J. Lang (Apr. 1989). "Phoneme recognition using time-delay neural networks". In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 37, pp. 328–339 (cit. on p. 11).

Wang, Changhan, Juan Pino, Anne Wu, and Jiatao Gu (2020a). "CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus". In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4197–4203 (cit. on p. 50).

Wang, Changhan, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux (Aug. 2021a). "VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 993–1003. URL: https://aclanthology.org/2021.acl-long.80 (cit. on pp. 48, 50, 52, 54, 55, 78, 99).

Wang, Changhan, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino (2020b). "Fairseq S2T: Fast Speech-to-Text Modeling with Fairseq". In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 33–39 (cit. on p. 46).

Wang, Changhan, Anne Wu, Jiatao Gu, and Juan Pino (2021b). "CoVoST 2 and Massively Multilingual Speech Translation." In: *Interspeech*, pp. 2247–2251 (cit. on pp. 28, 38, 44, 50).

Wang, Chengyi, 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". In: *arXiv preprint arXiv:2301.02111* (cit. on pp. 30, 88, 105, 111).

Wang, Hongyu, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei (2022). "DeepNet: Scaling Transformers to 1,000 Layers". In: *arXiv preprint arXiv:2203.00555* (cit. on pp. xi, 21, 74, 75, 80).

Wang, Kexin, Nils Reimers, and Iryna Gurevych (2021). "Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning". In: *arXiv preprint arXiv:2104.06979* (cit. on p. 17).

Wang, Liang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei (2022). "Text embeddings by weakly-

supervised contrastive pre-training". In: *arXiv preprint arXiv:2212.03533* (cit. on p. 17).

Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei (2023). "Improving Text Embeddings with Large Language Models". In: *arXiv preprint arXiv:2401.00368* (cit. on p. 17).

Wang, Yongqi, Jionghao Bai, Rongjie Huang, Ruiqi Li, Zhiqing Hong, and Zhou Zhao (2023b). "Speech-to-Speech Translation with Discrete-Unit-Based Style Transfer". In: *arXiv preprint arXiv:2309.07566* (cit. on p. 31).

Wang, Yuxuan, 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*. PMLR, pp. 5180–5189 (cit. on p. 30).

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. (2022). "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in Neural Information Processing Systems* 35, pp. 24824–24837 (cit. on p. 126).

Wenzek, Guillaume, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave (2019). "Ccnet: Extracting high quality monolingual datasets from web crawl data". In: *arXiv preprint arXiv:1911.00359* (cit. on pp. 42, 67, 72).

Wieting, John, Graham Neubig, and Taylor Berg-Kirkpatrick (2019). "A bilingual generative transformer for semantic sentence embedding". In: *arXiv preprint arXiv:1911.03895* (cit. on p. 20).

Williams, Adina, Nikita Nangia, and Samuel R Bowman (2017). "A broad-coverage challenge corpus for sentence understanding through inference". In: *arXiv preprint arXiv:1704.05426* (cit. on p. 16).

Wu, Jian, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. (2023). "On decoder-only architecture for speech-to-text and large language model integration". In: *arXiv preprint arXiv:2307.03917* (cit. on p. 126).

Xu, Chen, Bojie Hu, Yanyang Li, Yuhao Zhang, Qi Ju, Tong Xiao, Jingbo Zhu, et al. (2021). "Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders". In: *arXiv preprint arXiv:2105.05752* (cit. on p. 29).

Xu, Haoran, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla (2023). "A paradigm shift in machine translation: Boosting translation performance of large language models". In: *arXiv preprint arXiv:2309.11674* (cit. on p. 125).

Yang, Kaicheng, Hua Xu, and Kai Gao (2020). "CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20. Seattle, WA, USA: Association for

Computing Machinery, pp. 521–528. URL: https://doi.org/10.1145/3394171.3413690 (cit. on p. 23).

Yang, Yinfei, Gustavo Hernandez Abrego, Steve Yuan, Qinlan Shen Mandy Guo, Daniel Cer, Brian Strope Yun-hsuan Sun and, and Ray Kurzweil (2019a). "Improving Multilingual Sentence Embedding using Bi-directional Dual Encoder with Additive Margin Softmax". In: *IJCAI*, pp. 5370–5378 (cit. on pp. 19, 21).

Yang, Yinfei, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil (2019b). "Multilingual Universal Sentence Encoder for Semantic Retrieval". In: *https://arxiv.org/abs/1907.04307* (cit. on p. 19).

Yang, Ziyi, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve (2020). "Universal sentence representation learning with conditional masked language model". In: *arXiv preprint arXiv:2012.14388* (cit. on p. 19).

Ye, Rong, Mingxuan Wang, and Lei Li (2022). "Cross-modal contrastive learning for speech translation". In: *arXiv preprint arXiv:2205.02444* (cit. on p. 65).

Yu, Katherine, Haoran Li, and Barlas Oguz (July 2018). "Multilingual Seq2seq Training with Similarity Loss for Cross-Lingual Document Classification". In: *Proceedings of the Third Workshop on Representation Learning for NLP*. Ed. by Isabelle Augenstein, Kris Cao, He He, Felix Hill, Spandana Gella, Jamie Kiros, Hongyuan Mei, and Dipendra Misra. Melbourne, Australia: Association for Computational Linguistics, pp. 175–179. URL: https://aclanthology.org/W18-3023 (cit. on p. 18).

Yu, Lili, Dániel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis (2023). "Megabyte: Predicting million-byte sequences with multiscale transformers". In: *arXiv preprint arXiv:2305.07185* (cit. on p. 121).

Zeghidour, Neil, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi (2021). "Soundstream: An end-to-end neural audio codec". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30, pp. 495–507 (cit. on pp. 30, 31).

Zhang, Hao, Xukui Yang, Dan Qu, and Zhen Li (2022). "Bridging the Cross-Modal Gap Using Adversarial Training for Speech-to-Text Translation". In: *Digital Signal Processing*, p. 103764 (cit. on p. 65).

Zhang, Junlei, Zhenzhong Lan, and Junxian He (2023). "Contrastive Learning of Sentence Embeddings from Scratch". In: *arXiv preprint arXiv:2305.15077* (cit. on p. 17).

Zhang, Meng, Liangyou Li, and Qun Liu (2022). "Triangular Transfer: Freezing the Pivot for Triangular Machine Translation". In: *arXiv preprint arXiv:2203.09027* (cit. on p. 26).

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi (2019). "Bertscore: Evaluating text generation with bert". In: *arXiv preprint arXiv:1904.09675* (cit. on p. 26).

Zhang, Yu, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. (2023). "Google usm: Scaling automatic speech recognition beyond 100 languages". In: *arXiv preprint arXiv:2303.01037* (cit. on pp. 15, 28).

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books". In: *Proceedings of the IEEE international conference on computer vision*, pp. 19–27 (cit. on p. 12).

Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen (2016). "The united nations parallel corpus v1. 0". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3530–3534 (cit. on p. 25).

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight (2016). "Transfer learning for low-resource neural machine translation". In: *arXiv preprint arXiv:1604.02201* (cit. on p. 25).

Zweigenbaum, Pierre, Serge Sharoff, and Reinhard Rapp (2017). "Overview of the Second BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora". In: *BUCC*, pp. 60–67. URL: http://aclweb.org/anthology/W17-2512 (cit. on p. 22).