# Analyzing Hidden Representations in Multimodal Language Models

Bachelor's Thesis of

Hyunji Lee

Artificial Intelligence for Language Technologies (AI4LT) Lab
Institute for Anthropomatics and Robotics (IAR)
KIT Department of Informatics

| | |
|---|---|
| Reviewer: | Prof. Dr. Jan Niehues |
| Second reviewer: | TT-Prof. Dr. Barbara Bruno |
| Advisor: | M.Sc. Danni Liu |
| Second advisor: | M.Sc. Supriti Sinhamahapatra |

13. May 2024 – 13. September 2024

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

**Karlsruhe, DATE**

Hyunji Lee

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

(Hyunji Lee)

# Abstract

Multimodal Language Models (MLMs) are designed to process input from different modalities. Unlike unimodal Large Language Models (LLMs), multimodality adds complexity to the models and introduces a unique set of tasks per modality that the MLM must learn in parallel. Therefore, it is crucial to understand their behavior and how they learn multimodal features. Since hidden representations capture the internal states of MLMs, this work focuses on analyzing how different or similar the representations of different modalities are. We analyze the cross-modal and cross-lingual similarities of representations based on inputs of the same semantic meaning, and additionally visualize these representations to examine the distribution of modalities. We performed our analysis on the MLMs SeamlessM4T, SONAR and SALMONN. Our results show that the similarities are influenced by a wide range of factors, from the architecture of the models to their training strategies and the resource levels of the languages analyzed. The tasks on which the MLMs were trained also strongly influence the similarity between speech and text representations. Translation models and embedders achieve high similarity between multimodal representations, while instruction-following models do not prioritize high cross-modal similarity. The cross-modal similarity of each model is quite high, which means that efforts are made to close the modality gap. The cross-lingual similarity within the text modality is generally higher than within the speech modality for each model, but it differs from model to model which gap - either modality or language - is more closed. Additionally, both cross-modal and cross-lingual similarity can be further increased only for high resource languages. Finally, the distributions of the multimodal representations indicate that the modality features are evident in all hidden representations of each model, which is consistent with our previous results.

# Zusammenfassung

Multimodale Sprachmodelle (Multimodal Language Models, MLMs) werden entwickelt, um Eingaben aus verschiedenen Modalitäten zu verarbeiten. Im Gegensatz zu unimodalen Large Language Models (LLMs) erhöht die Multimodalität die Komplexität der Modelle und führt neue Aufgaben für jede Modalität ein, die das MLM parallel erlernen muss. Daher ist es von großer Bedeutung, ihr Verhalten zu verstehen und wie sie multimodale Merkmale erlernen. Da verborgene Repräsentationen die internen Zustände von MLMs erfassen, konzentriert sich diese Arbeit auf die Analyse, wie unterschiedlich oder ähnlich die Repräsentationen der verschiedenen Modalitäten sind. Wir analysieren die intermodale und interlinguale Ähnlichkeiten von Repräsentationen, die auf derselben semantischen Bedeutung basieren. Zusätzlich werden Repräsentationen visualisiert, um die Verteilung der Modalitäten zu untersuchen. Wir haben unsere Analyse mit den MLMs SeamlessM4T, SONAR und SALMONN durchgeführt. Unsere Ergebnisse zeigen, dass die Ähnlichkeiten von einer Vielzahl von Faktoren beeinflusst werden: von der Architektur der Modelle über ihre Trainingsstrategien bis hin zu den Ressourcenniveaus der analysierten Sprachen. Die Aufgaben, für die die MLMs trainiert wurden, haben ebenfalls einen starken Einfluss auf die Ähnlichkeit zwischen Sprach- und Textrepräsentationen. Die Übersetzungs- und Embedding-Modelle erreichen eine hohe intermodale Ähnlichkeit, während Modelle, die Anweisungen befolgen, keine hohe Ähnlichkeit anstreben. Die multimodale Ähnlichkeit jedes Modells ist jedoch recht hoch, was bedeutet, dass die Modelle versuchen, die Modalitätslücke zu schließen. Die interlinguale Ähnlichkeit innerhalb der Textmodalität ist im Allgemeinen bei jedem Modell höher als innerhalb der Sprachmodalität, aber es ist von Modell zu Modell unterschiedlich, welche Lücke - entweder Modalität oder Sprache - stärker geschlossen wird. Darüber hinaus kann sowohl die intermodale als auch die interlinguale Ähnlichkeit nur für Sprachen mit hohem Ressourcenniveau weiter erhöht werden. Schließlich deuten die Verteilungen der multimodalen Repräsentationen darauf hin, dass die Modalität in allen verborgenen Repräsentationen jedes Modells erhalten bleibt, was mit unseren früheren Ergebnissen übereinstimmt.

# Contents

# List of Figures

# List of Tables

# 1. Introduction

## 1.1. Motivation

In recent years, Large Language Models (LLMs) have emerged as powerful general-purpose tools, significantly reshaping how we approach various tasks and perceive the world. The main reason for this phenomenon is the accessibility and convenience of these language models over traditional alternatives. LLMs can also specialize in a wide variety of tasks, from simple ones such as automatic speech recognition, translation, and question answering to more challenging tasks such as text generation and programming assistance.

Although LLMs are very present in our world today, it is important to mention that language models, like any other Artificial Intelligence (AI) model, are limited to the modality they have been developed for. Since LLMs have mostly been developed for text-based tasks, several successful unimodal text-based language models have emerged, such as GPT (Brown et al., 2020) and Llama (Touvron et al., 2023). Additionally, extensive research has focused on improving the quality and performance of these unimodal models (Kaddour et al., 2023; W. X. Zhao et al., 2023), leading to significant advances in their capabilities.

On the other hand, Multimodal Language Models (MLMs) aim to integrate several modalities, such as text, speech, and images, into one single model, thereby expanding the accessibility and flexibility of LLMs. By overcoming the limitations of unimodal language models in understanding and processing different modalities, MLMs open up a new field of research in Natural Language Processing (NLP) and new possibilities for handling and combining diverse modalities.

Despite the high potential of MLMs, a key challenge to understand how these models work internally. Since MLMs differ in terms of the modalities they support and the tasks for which they are designed, understanding their internal mechanisms becomes essential. Additionally, the ability of MLMs to process multimodal inputs can be viewed as performing multi-task learning, where each modality introduces a unique set of tasks that must be learned simultaneously. Thus, understanding how different tasks and modalities interact within the model is crucial to expanding our knowledge in MLMs.

MLMs encode the learned features and the complex relationship between different inputs and tasks in their hidden representations, which are then used to generate predictions and outputs. Therefore, understanding the internal states of MLMs and how they handle multimodality lies in analyzing their hidden representations. By analyzing the similarities between representations across modalities we gain further insights on the behaviour of

MLMs and how they manage the interaction between tasks and modalities, highlighting their strengths, limitations, and biases.

## 1.2. Research Questions

This work aims to answer the main question of how similar - or different - the hidden speech and text representations of MLMs truly are. Analysis in this field is crucial for our knowledge of MLMs and could tell us how MLMs learn and process their multimodal input. The main objective can be broken down into the following sub-questions:

- **Research Question 1: How does the similarity of representations change with the depth of the model's layers?**
  To fully understand how MLMs handle multimodality, the relationship between the inputs and the layers of the model's architecture is analyzed. As each layer processes modality-specific features differently, diving deeper into this relationship will give us further insights on how the MLMs capture different features in their hidden representations ranging from language, modality and semantic meaning.

- **Research Question 2: How does the similarity of representations change with varying language resource levels?**
  Since there are more languages spoken than all the countries in the world combined, some are less represented than others. This distribution is also evident in the available data used to train MLMs, with low resource languages especially lacking in high-quality speech data. With this research question, it is analyzed how MLMs perform under the limitations of language resource levels and how they affect the similarity between multimodal representations.

- **Research Question 3: How do the similarities of representations differ in a cross-modal and cross-lingual setting?**
  The complexity of multilingual MLMs is higher than that of unimodal LLMs, since they must be able to handle input data that differs across languages and modalities. By answering this research question, we can examine to what extent MLMs capture modality and language features and how they influence the similarities between representations.

- **Research Question 4: How does the architecture of the model affect the similarity of representations?**
  MLMs typically have special components focused on each modality in their architecture, as modalities are very different in their structures. Achieving a high similarity between multimodal representations with the same semantic value is crucial for the performance and robustness of MLMs. By analyzing how each component contributes to the similarity, and whether these components truly achieve what they aim for, we gain a better understanding of how MLMs handle multimodality.

# 2. Fundamentals and Related Work

## 2.1. Sequence-to-Sequence Models: Early Approaches

Sequence-to-Sequence (Seq2Seq) models are widely used in the field of Natural Language Processing (NLP) as they are designed to transform data sequences of a domain into another sequence of a different domain. First introduced by Sutskever, Vinyals, and Le (2014), Seq2Seq models are able to process inputs and outputs of varying lengths and are therefore often used for complex language problems such as machine translation, question answering and creating chatbots.

The architecture of early Seq2Seq models most commonly consists of two subsequent Recurrent Neural Networks (RNNs): an encoder and a decoder. In the context of NLP tasks, input data is first broken down into smaller single units of meaning, called tokens, which are sequentially processed by the encoder. With each token the encoder produces hidden states, which captures the relevant information from the input sequence seen up to that hidden layer. The encoder eventually creates a fixed-size context vector, which is used as the input for the decoder to generate the output sequence token by token, predicting the next sequence token based on the context vector and the previously generated tokens.

To improve the performance of the decoder in Seq2Seq models, the attention mechanism (Bahdanau, Cho, and Bengio, 2016) is applied. Attention acts as a dynamic weighting mechanism, allowing the decoder to focus on the relevant parts of the input sequence at each generation step, rather than relying only on a fixed-size context vector. This enables the decoder to gain more information from the encoder's hidden states, helping it to better capture dependencies across the input sequence and generate more accurate outputs.

However LLMs based on the Seq2Seq approach also have their downsides, as they struggle to handle long sequences due to the vanishing gradient problem, limiting the model to learn input data across a broader range of length. Additionally, due to its token-by-token procedure, Seq2Seq models are difficult to parallelize.

## 2.2. Transformers in Natural Language Processing

As Seq2Seq models have been extensively used in the field of NLP, improvements have also come along and a new variation of the Seq2Seq model has been introduced by Vaswani et al. (2017): Transformers. They also consist of a encoder and a decoder, however instead of using RNNs, transformers rely entirely on the self-attention mechanism to produce representations of the inputs.

Self-attention allows each token to attend to the other tokens in the same sequence, in both encoder and decoder of the transformer. This mechanism enables transformers on the one hand to attend to any token regardless of its distance from the current token and on the other to produce more context-aware representations, capturing dependencies and relationships across the entire length of the same sequence. Transformers are as a result more robust to long sequences and are parallizable, contrary to the early approaches of Seq2Seq models.

Additionally to the self-attention mechanism in each transformers layer, a feed-forward neural network (FFNN) and a layer normalization is followed afterwards. Both components increase the quality of the transformers outputs, since FFNNs add non-linearity to each token representation and the normalization stabilizes the training procedure by normalizing activations across each layer.

## 2.3. Large Language Models

Due to its many benefits, transformers have paved the way for the development of Large Language Models (LLMs) and have become the state-of-the-art NLP models. Due to the scalability of LLMs, they can be developed for a wide range of tasks with high capabilities by training on a large amount of data and an appropriate training strategy.

Transformer-based LLMs, such as BERT (Devlin et al., 2019), GPT (Brown et al., 2020) and Llama (Touvron et al., 2023), have been transformative in how NLP tasks are approached. The BERT (Bidirectional Encoder Representations from Transformers) model is used to represent text as a sequence of vectors and can be applied in a wide range of use cases. Thus, many LLMs and MLMs models use BERT as their foundation. Additionally, the decoder-only GPT (Generative Pretrained Transformer) models and the Llama models have proven to be excelling at text generation tasks.

## 2.4. Multimodal Language Models

The ability of Multimodal Language Models (MLMs) to accept and process inputs of different modalities is an additional advantage over unimodal LLMs and makes them more attractive in certain applications. MLMs are often sought after for their flexibility and recent works (X. Wang et al., 2023; Yin et al., 2024; Zhang et al., 2024) summarize advances in the research of MLMs while outlining the architecture, training strategies and performance evaluation methods of these models.

The most frequently supported input modalities of MLMs are text, audio (e.g. speech, music, and ambient noise) and images. The simplest MLMs have only two input modalities. For example, GPT-4 (OpenAI et al., 2024) accepts image and text inputs and can produce text outputs, while SpiRit-LM (Nguyen et al., 2024) has speech and text as both input and output modalities. Other MLMs such as Gemini (G. Team et al., 2024) can handle more than two input modalities: image, audio, video, and text. ImageBind (Girdhar et al., 2023) and OneLLM (Han et al., 2024) further extend the traditional definition of MLMs more by additionally accepting uncommon modalities such as depth and inertial measurement unit (IMU) data as input.

To integrate these non-text input modalities drastically differing in structure and information, MLMs often have separate components, aside from isolated encoders for each input modality, dedicated to transform inputs varying in modality to the shared space of an MLM. For example, Querying Transformers (Q-Formers) are used to align audio features to a text-based LLM (Tang et al., 2024).

## 2.5. Analysis of Hidden Representations

Similar work in analyzing the hidden representations of language models has been done in G. Wang et al. (2023), where the text and speech representations of models with separate speech and text encoders followed by an additional shared encoder were analyzed. The t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and G. Hinton, 2008) and a probing method were used to compare the speech and text representations before and after the joint space. It was shown that the joint representations after the shared encoder were more unified in modality and domain than before the shared encoder.

In Sun et al. (2023), the hidden multilingual representations of end-to-end speech translation models, trained on three separate translation directions (Eng → X, X → X and X → Eng) were compared using the Singular Vector Canonical Correlation Analysis (SVCCA) (Raghu et al., 2017) and the Linear Discriminant Analysis (LDA) (Tharwat et al., 2017). It is highlighted that the SVCCA similarities between representations of similar languages increase with the depth of the encoder. Unique languages, such as the Indo-European language Persian, create their own subspace in LDA, resulting in low SVCCA scores compared to other languages in the same family.

Conversely, Seyssel et al. (2022) analyzes the phonetic class, gender, and language information encoded in the Contrastive Predictive Coding (CPC) (Oord, Li, and Vinyals, 2019) representations of self-supervised speech models - two monolingual models (English, French) and one multilingual model (English and French). These representations were visualized with t-SNE, and a probing method with a logistic regression classifier was trained to evaluate the error scores on phonetic class, gender and language across the three models. This work concludes that information about phonetic class and gender are similarly represented in all three models. However, the distinction between English and French was only visible in the multilingual model, while in the monolingual models, the language information is diffused across multiple dimensions.

The V-Measure was used in Sicherman and Adi (2023) to determine the phoneme, gender, and speaker ID information between discrete self-supervised speech representations of a CPC model, HuBERT (Hsu et al., 2021) and a Mel-Frequency Cepstrum Coefficients (MFCC) model, while also considering the total number of discrete speech units. The findings indicate that the representations show a strong relationship with phonemes, as well as with gender and speaker ID. To visualize the phoneme information, t-SNE was also used, demonstrating that units of the same phoneme and phoneme family are more closer to one another in all three models.

# 3. Methodology

Given a pre-trained MLM, a pre-selected set of model architecture layers and a pre-selected set of model-supported languages, the hidden speech and text representations are extracted for further analysis. In this work, only the hidden representations produced directly from the speech and text inputs of the same semantic meaning are extracted, before any output generation takes place. For each layer, the extracted hidden representation of a speech or text input is of size *(input length, feature size)*, where both values vary depending on the input and the model. This hidden representation matrix is then averaged over the input length dimension, resulting in a feature size vector for each layer.

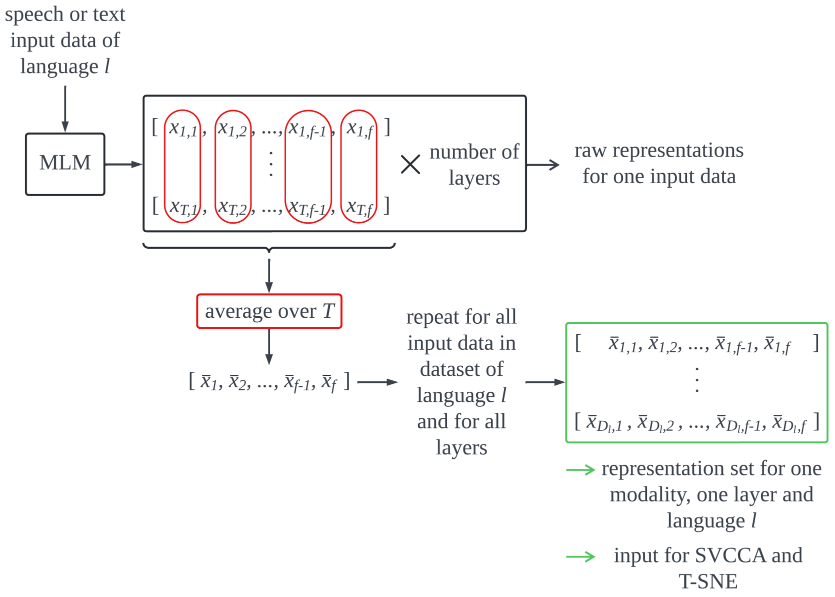

speech or text input data of language $l$
MLM
$[x_{1,1}, x_{1,2}, ..., x_{1,f-1}, x_{1,f}]$
$[x_{T,1}, x_{T,2}, ..., x_{T,f-1}, x_{T,f}]$
× number of layers
raw representations for one input data
average over $T$
$[\bar{x}_1, \bar{x}_2, ..., \bar{x}_{f-1}, \bar{x}_f]$
repeat for all input data in dataset of language $l$ and for all layers
$[\bar{x}_{1,1}, \bar{x}_{1,2}, ..., \bar{x}_{1,f-1}, \bar{x}_{1,f}]$
$[\bar{x}_{D_l,1}, \bar{x}_{D_l,2}, ..., \bar{x}_{D_l,f-1}, \bar{x}_{D_l,f}]$
representation set for one modality, one layer and language $l$
input for SVCCA and T-SNE

**Figure 3.1.: Extraction of Representation Sets.** With $f$ = feature size, $T$ = input data length and $D_l$ = number of input data of language $l$.

As shown in Figure 3.1, this averaging process is repeated for all speech and text representations of each pre-selected language and it returns two sets of representations of size *(number of input data, feature size)* per language and layer, one for the speech inputs and one for the text inputs. In each row of both representation sets, the averaged hidden representation from the input sentence with the same semantic meaning is found. Formally, there are $D_l$ many speech and text representations $x_i = [x_{i,1}, x_{i,2}, ..., x_{i,f-1}, x_{i,f}]$ with $D_l$ = number of input data of language $l$, $f$ = feature size, $x_i$ = averaged representation of input $i$ and $x_{i,j}$ = averaged representation of input $i$ and feature $j$. These representation sets are then used for further analysis, as listed in the sections below.

## 3.1. Cross-Modal Similarity Analysis

The Singular Vector Canonical Correlation Analysis (SVCCA) (Raghu et al., 2017) is used to evaluate the similarly of the extracted speech and text representations. Two sets of representations $X \in (F_x, M)$ and $Y \in (F_y, N)$, with $F_x$ and $F_y$ being the feature sizes and $M$ and $N$ being the number of data points, can be given to calculate the SVCCA similarity. The representation sets may differ in feature sizes ($F_x \neq F_y$), however the number of data points have to be the same ($M = N$). SVCCA first performs a singular value decomposition on both $X$ and $Y$, resulting in two sets of singular vectors and singular values. After that, Canonical Correlation Analysis (CCA) is applied on only the top $m \leq M$ and top $n \leq N$ singular vectors that explain 90% variance of $X$ and $Y$ with the top $m$ and $n$ singular values. CCA will then find linear transformations that maximize the correlation between two vector sets, returning CCA correlation coefficients. The averaged value of all coefficients is the SVCCA similarity value $\in [0, 1]$, depending on how similar (= 1) or different (= 0) the two sets of representations are. The goal of this analysis is to have one SVCCA cross-modal similarity score for each layer of one model to see how the similarity changes with the depth of the model architecture. To achieve this, the following steps are carried out.

1. Firstly, all speech and text representation sets of each layer are reduced to match the size of the smallest representation set, which is from the language with the smallest total number of input data. For example, if English has a total of $M$ input data and Dutch has a the least with a total of $N$, with $M > N$, the English speech and text representation sets for each layer would be reduced from ($M$, *feature size)* to ($N$, *feature size)* only leaving the first $N$ input data behind. This step ensures consistency by performing all SVCCA comparisons on sets with the same number of representations.

2. Since the feature sizes of MLMs are much greater than the total number of representations of one modality, the feature dimension must first be reduced before any SVCCA calculations are performed. To achieve this, the representation sets are reduced once more to the dimension explaining 90% of the total variance. The resulting smaller dimension is different for each modality, language and layer. For instance, the reduced English speech and text representation sets for each layer from step 1 would be now of size $(N, d_s)$ and $(N, d_t)$ respectively, if the dimensions $d_s$ and $d_t$ explain 90% of the total speech and text variance.

3. With all representation sets reduced twice to the desired size, pairs consisting of the speech and text representation set of the same language and layer are given to the SVCCA algorithm to compute the modality similarity value between 0 and 1.

4. At this point, there are a total of *(number of languages)* × *(number of layers)* SVCCA scores to analyze. For each layer, the similarity scores of all pre-selected languages are averaged together, now resulting in *(number of layers)* SVCCA scores.

## 3.2. Cross-Lingual Similarity Analysis

A similar analysis as in Section 3.1 was also implemented for the cross-lingual similarity analysis, in order to get the SVCCA similarity scores for each possible language pair in the pre-selected language set. The same procedure was followed as described in Section 3.1 with a few changes:

- Before reducing each representation set of each layer to the size of the smallest set, as in the first step of Section 3.1, the intersection of the input data of each language pair is first determined. Each intersection is then reduced to the size of the smallest intersection, which is subsequently used to assemble new speech and text representation sets for the cross-lingual SVCCA calculations by accumulating all the representations that are produced from the input data in the intersection. For example, if the number of intersecting input data of English and German is $M$ and there are only $N$ intersecting inputs for English and Dutch, with $M > N$, the English-German intersection is reduced to the first $N$ intersecting input data, so that all speech and text representation intersections have the size $(N,$ *feature size)*. Thus, all intersecting representation sets are of the same size and the sets of each language pair comparison are based on the same input sentences.

- For cross-lingual similarity comparisons, each representation set is also reduced once more to a smaller feature dimension. The reduced sets explain at least 90% of the total variance.

- For each layer, there are 4 different modality comparisons to consider: (1) speech-speech, (2) text-text, (3) speech-text and (4) text-speech. (4) is left out in this work, because the comparison results are the transposed of those of (3).

- Due to the symmetry of the cross-lingual comparisons, this procedure results in $($*(number of languages)*$^2 \div 2)$ SVCCA similarity scores for each layer of the same modality comparisons (text-text and speech-speech). For speech-text similarity comparisons, there are in total of *(number of languages)*$^2$ similarity scores for each layer.

## 3.3. Visualisation of Hidden Representations

Apart from explicit similarity values, this work also analyzes the hidden representations with t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and G. Hinton, 2008). For each layer, the speech and text representations are concatenated as one and given to the t-SNE algorithm with representation labels according to modality, language and input data to visualize the distribution of the hidden representations on a two-dimensional map.

T-SNE is a non-linear dimensionality reduction method to visualize high-dimensional data into a more interpretable lower dimensionality space. It is a modified version of Stochastic Neighbour Embedding (SNE) developed by G. E. Hinton and Roweis (2002), using a different cost function with simpler gradients, which is easier to optimize. Additionally, t-SNE solves the problem of crowding points in the center of the visualization map which was evident in SNE, creating visualization results of higher quality.

The t-SNE algorithm starts just as the SNE algorithm, by calculating the probability $p_{ij}$ for each possible pair $(x_i, x_j)$ in the dataset $X = x_1, ..., x_n(i \neq j and i, j \in [1, n])$, with higher probabilities indicating a higher similarity of a pair. Then for $T$ iterations, the positions of the data points $Y = y_1, ..., y_n$ in the low-dimensional space is computed using its Student t-distributed similarity probability $q_{ij}$ of two points $y_i$ and $y_j$ and the probabilities $p_{ij}$ of the higher dimension with minimizing the Kullback-Leibler divergence of the two distributions. After the last iteration, the elements in $Y$ show the distribution of the high dimensional data in a low-dimensional space.

# 4. Experimental Setup

## 4.1. Models

This work applies the aforementioned methods in Chapter 3 on three separate MLMs: SeamlessM4T, SONAR and SALMONN. In the following sections, the architecture of the analyzed MLM are illustrated and it is explained of which part of the architecture layers the speech and text hidden representations are extracted.

### 4.1.1. Encoder-Decoder Model: SeamlessM4T

Seamless **M**assively **M**ultilingual & **M**ultimodal **M**achine **T**ranslation (SeamlessM4T) is a MLM for tasks such as Automatic Speech Recognition (ASR) and translation in all for possible directions (speech-to-speech, speech-to-text, text-to-speech and text-to-text) (Communication, Barrault, Chung, Mariano Coria Meglioli, et al., 2023). SeamlessM4T supports over 100 languages varying in resource levels and is a new advancement in the field of multimodal machine translation. Its goal is to bride the modality gap between speech and text of recent direct and cascaded models by combining a multilingual text-to-text translation model with a speech representation model.

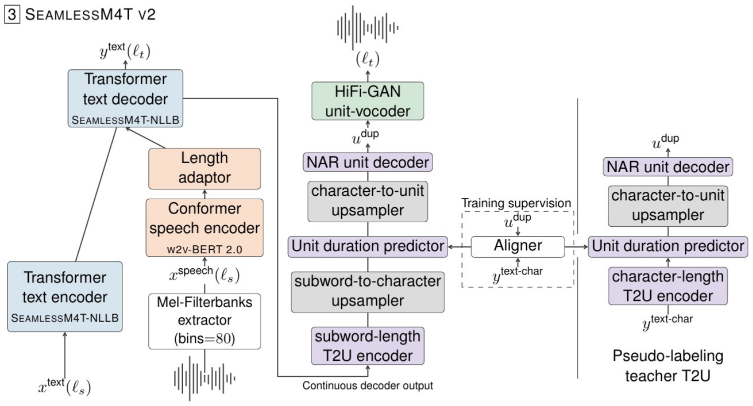


**Figure 4.1.: Model Architecture of SeamlessM4T.** (Communication, Barrault, Chung, Mariano Coria Meglioli, et al., 2023)

The SeamlessM4T architecture, as shown in Figure 4.1, can be split into two parts, the text and the speech generation. The text generation part consists of the components before and including the transformer text decoder, while the components after the text decoder are used to generate speech out of the text decoder output. This work focuses on the hidden representations before the transformer text decoder, which are those of the conformer speech and transformer text encoder, additionally analysing the input embeddings and the representations after the length adaptor for speech inputs.

Text inputs go through the SeamlessM4T's transformer text encoder and decoder, which are initialized with SeamlessM4T-NLLB (Communication, Barrault, Chung, Mariano Coria Meglioli, et al., 2023) - a multilingual text-to-text translation model supporting 200 languages. Meanwhile, speech inputs first pass through the mel filterbank feature extraction, where the outputs are given to the conformer speech encoder, initialized with the speech representation learning model W2v-BERT 2.0 (Chung et al., 2021) and is post-fixed with a length adaptor. The length adaptor of SeamlessM4T is a modified version of the M-Adaptor of J. Zhao et al. (2022), used to adapt speech representations to text by downsizing the speech sequence and building features for speech-to-text translation.

For the analysis, the pre-trained transformer SeamlessM4T model `facebook/seamless-m4t-v2-large` from Hugging Face[1] is used. Both speech and text encoders have 24 layers with a feature size of 1024. Aside from the speech and text representations of the layers $\{1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 23, 24\}$, both speech and text input embeddings and the speech representations after the length adaptor were also analyzed. All representation sets of SeamlessM4T are of size $(D_l, 1024)$ for each language and layer, since the input embeddings and the speech representations after the length adaptor also have the feature size of 1024. The components including and after the shared text decoder was not analyzed in this work, as we focus on the multimodality of the representations. SeamlessM4T generates speech outputs with the text outputs of the shared text decoder through a cascaded system, making the analysis of the decoder side of SeamlessM4T difficult and beyond the scope of this work.

### 4.1.2. Sentence Embedder: SONAR

**S**entence-level multim**O**dal and la**N**guage-**A**gnostic **R**epresentations (SONAR) is a multimodal and multilingual sentence embedding space. Apart from its functionality to embed sentences of 200 languages, SONAR can also be used for translating speech and text inputs to text outputs (Duquenne, Schwenk, and Sagot, 2023).

As shown in Figure 4.2, the SONAR architecture consists of one multilingual text encoder initialized with NLLB (N. Team et al., 2022) and multiple monolingual speech encoders initialized with W2v-BERT 2.0 (Chung et al., 2021) followed by a multilingual text decoder also initialized with NLLB. For comparing the multimodal representations of SONAR, this work focuses on the hidden representations of the encoders. Speech or text inputs given to SONAR surpass all layers of the corresponding encoder, which then the last encoder representations are used to produce language-agnostic sentence embeddings by pooling along the sequence dimension. While mean pooling is used for text encoder outputs, learning (attention) pooling is used for the speech outputs. Additionally, the mean squared error (MSE) loss is used in the SONAR embedding space, which encourages

[1] https://huggingface.co/facebook/seamless-m4t-v2-large

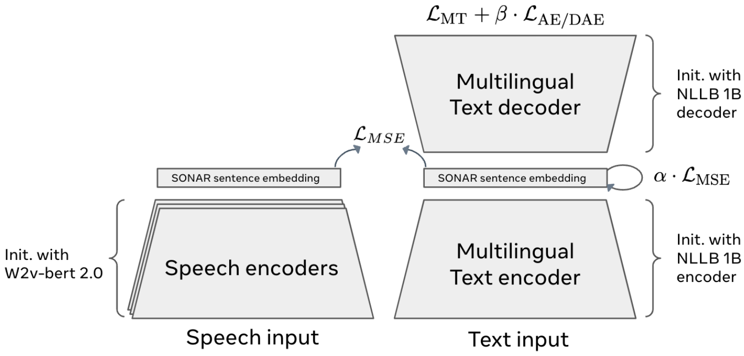

$\mathcal{L}_{\mathrm{MT}} + \beta \cdot \mathcal{L}_{\mathrm{AE/DAE}}$
Multilingual Text decoder
Init. with NLLB 1B decoder
$\mathcal{L}_{MSE}$
SONAR sentence embedding
SONAR sentence embedding
$\alpha \cdot \mathcal{L}_{\mathrm{MSE}}$
Init. with W2v-bert 2.0
Speech encoders
Multilingual Text encoder
Init. with NLLB 1B encoder
Speech input
Text input

**Figure 4.2.: Model Architecture of SONAR.** (Duquenne, Schwenk, and Sagot, 2023)

the SONAR to correctly align sentences in the shared embedding space by reducing the differences between embeddings of the same semantic meaning but of different languages and modality.

The pre-trained SONAR model from `fairseq2`[2] was chosen for this work. Similar to SeamlessM4T in Section 4.1.1, all encoders have in total of 24 layers with the same feature size of 1024. Also for SONAR as well, speech and text representations of the same layers $\{1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 23, 24\}$ were extracted, with the addition of the input embeddings, the final speech and text embeddings after the pooling, which are all of the same feature size of 1024. Thus, the representation sets of both modalities for each layer and language are the same size as those extracted from SeamlessM4T, specifically $(D_l, 1024)$.

### 4.1.3. Decoder-Only Model: SALMONN

**S**peech **A**udio **L**anguage **M**usic **O**pen **N**eural **N**etwork (SALMONN) is a MLM developed to process music, speech and also ambient noise in combination with a text instruction prompt (Tang et al., 2024).

SALMONN is based on the pre-trained Vicuna[3] model (Zheng et al., 2023), which is a text-based LLM fine-tuned from the Llama2 model (Touvron et al., 2023) to follow text instructions, and is equipped with low-rank adaptation (LoRA) (Hu et al., 2021) to align the two cross-modal input and output space of Vicuna. The audio and the text instruction referring to the audio are simultaneously given to SALMONN. While the text inputs are embedded for the Vicuna model by the Llama tokenizer and embedder in a fairly simple way, audio inputs have to surpass several components. As seen in Figure 4.3, the audio inputs are first fed into the encoder of the ASR model Whisper[4] (Radford et al., 2023) and the BEATs[5] (Chen et al., 2022) encoder, which can process a wide range of audio data

[2] https://github.com/facebookresearch/SONAR
[3] https://huggingface.co/lmsys/vicuna-7b-v1.5
[4] https://huggingface.co/openai/whisper-large-v2
[5] https://github.com/microsoft/unilm/tree/master/beats

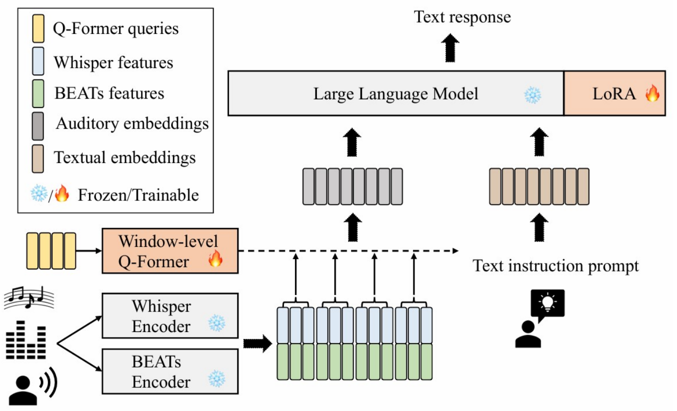


**Figure 4.3.: Model Architecture of SALMONN.** (Tang et al., 2024)

beside speech. The resulting two outputs are then given to a Window-level Q-Former that unifies the two encoder outputs into auditory embeddings of the input space of Vicuna by transforming the encoder output sequence varying in length to audio tokens of fixed length.

Since SALMONN only accepts audio and text inputs simultaneously and the auditory and textual embeddings are given to Vicuna as one concatenated input, the extracted raw representations equal the concatenated speech and text representations. To analyze hidden speech and text representations separately, the raw representations are split into speech and text representations with the input length dimension.

For this work, the 7B version[6] of SALMONN was used and the decoder layers of the Vicuna LLM were analyzed. The decoder has 32 layers and the representation sets of layers $\{1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 31, 32\}$ with addition to the speech encoder outputs before the Q-Former, the textual and the auditory embeddings were extracted. The speech encoder outputs before the Q-former has the feature size of 2048, while the text embeddings, speech encoder outputs after the Q-former and all decoder layers have a feature size of 4096. The different feature sizes cause no problem for SVCCA, as it can handle representations of different sizes. However for t-SNE, the input size matters and the speech encoder outputs before the Q-former were padded at the end with zeros from 2048 to 4096.

| MLM | Speech Representations & Size of Sets | Text Representations & Size of Sets |
|---|---|---|
| SeamlessM4T | • input embeddings $\Rightarrow (D_I, 1024)$<br>• encoder hidden representations of layers $\{1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 23, 24\}$ $\Rightarrow (14, D_I, 1024)$ | • input embeddings $\Rightarrow (D_I, 1024)$<br>• encoder hidden representations of layers $\{1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 23, 24\}$ $\Rightarrow (14, D_I, 1024)$ |

[6] https://huggingface.co/tsinghua-ee/SALMONN-7B

| MLM | Speech Representations & Size of Sets | Text Representations & Size of Sets |
|---|---|---|
| SONAR | • input embeddings<br>⇒ $(D_l, 1024)$<br>• encoder hidden representations of layers {1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 23, 24}<br>⇒ $(14, D_l, 1024)$<br>• final SONAR embeddings<br>⇒ $(D_l, 1024)$ | • input embeddings<br>⇒ $(D_l, 1024)$<br>• encoder hidden representations of layers {1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 23, 24}<br>⇒ $(14, D_l, 1024)$<br>• final SONAR embeddings<br>⇒ $(D_l, 1024)$ |
| SALMONN | • encoder outputs before Q-Former<br>⇒ $(D_l, 2048)$<br>• auditory embeddings after Q-Former<br>⇒ $(D_l, 4096)$<br>• decoder hidden representations of layers {1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 31, 32}<br>⇒ $(18, D_l, 4096)$ | • textual embeddings<br>⇒ $(D_l, 4096)$<br>• decoder hidden representations of layers {1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 31, 32}<br>⇒ $(18, D_l, 4096)$ |

**Table 4.1.: List of Analyzed Model Representations.** With *$D_l$ = number of input data of language l.*

## 4.2. Data and Languages

For creating the speech and text representation sets, the FLEURS[7] dataset (Conneau et al., 2023) was used. FLEURS supports 102 languages with an even distribution in language resource levels and is based on the FLORES[8] dataset, which sources its sentences from multiple Wikimedia sources from varying domains (Goyal et al., 2022). The main reason for choosing FLEURS for this work's data source is due to its n-way parallel sentences, which are crucial to the similarity analysis explained in Chapter 3.

The categorization of languages into its resource-levels is based on the available hours of speech-to-text translation data into English and pseudo labeled ASR data (Communication, Barrault, Chung, Mariano Cora Meglioli, et al., 2023). The languages in each category is used as the language set in Chapter 3 to analyze the influence of resource levels on the hidden representations. A language is a high resource language, if there are at least 1000 hours of data, and a low resource language, if the hours of data are less than or equal to 500. Every language with the volume of available data in between 1000 and 500 are medium resource languages.

The 102 languages supported by FLEURS was reduced to a set of 30 languages for each MLM, depending on which languages each model supports. While deciding on the languages, care was taken into maintaining an even distribution of different language characteristics such as script, family and resource-level (high, medium and low). As shown

[7] https://huggingface.co/datasets/google/fleurs

[8] https://huggingface.co/datasets/facebook/flores

in Table 4.2, SeamlessM4T and SALMONN share the same language set. However for SONAR, some languages had to be swapped out for other languages of the same resource-level because SONAR did not support languages such as Amharic, Greek and Khmer.

The FLEURS dataset is divided into three splits: train, validation and test. For this work, the text split was used to extract hidden representations and the normalized transcriptions was used for the text representations instead of raw transcriptions. As FLEURS has multiple dataset entries for the same sentence but with different speakers, these duplicates were removed randomly before extracting the representation for both modalities, so that each representation set would not have two averaged representations that have the same semantic meaning. Even though the speech representations based on the same sentence may differ due to different speakers and audio recording environments, duplicate representations in the text set would influence the similarity computations in our work, since the normalized transcriptions do not change across duplicates.

For the cross-modal analysis of Section 3.1 each representation sets are reduced to the first 251 representations, as this is the smallest number of input data without duplicates (see Dutch in Table 4.2). For the cross-lingual analysis of Section 3.2, each intersects were reduced to the first 194 intersecting representations for SeamlessM4T and SALMONN, and 192 for SONAR.

| Code | Name | Script | Family | Resource-Level | SeamlessM4T/ SALMONN | SONAR | Text Split Size | Without Duplicates |
|---|---|---|---|---|---|---|---|---|
| amh | Amharic | Ethiopic | Afro-Asiatic | low | x |  | 516 | 296 |
| arb | Arabic | Arabic | Afro-Asiatic | high | x | x | 428 | 283 |
| asm | Assamese | Bengali | Indo-European | low |  | x | 984 | 349 |
| bul | Bulgarian | Cyrillic | Indo-European | low | x | x | 658 | 344 |
| cat | Catalan | Latin | Indo-European | high | x | x | 940 | 350 |
| cmn | Chinese Mandarin | Hant | Sino-Tibetan | high | x | x | 945 | 349 |
| deu | German | Latin | Indo-European | high | x | x | 862 | 347 |
| ell | Greek | Greek | Indo-European | medium | x |  | 650 | 333 |
| eng | English | Latin | Indo-European | high | x | x | 647 | 350 |
| est | Estonian | Latin | Uralic | medium | x | x | 893 | 345 |
| fin | Finnish | Latin | Uralic | high | x | x | 918 | 348 |
| fra | French | Latin | Indo-European | high | x | x | 676 | 332 |
| heb | Hebrew | Hebrew | Afro-Asiatic | low |  | x | 792 | 347 |
| hin | Hindi | Devanagari | Indo-European | medium | x | x | 418 | 265 |
| hye | Armenian | Armenic | Indo-European | low | x |  | 932 | 350 |
| ind | Indonesian | Latin | Austronesian | medium | x | x | 687 | 328 |
| ita | Italian | Latin | Indo-European | high | x | x | 865 | 346 |
| jpn | Japanese | Japanese | Japonic | high | x | x | 650 | 321 |
| kat | Georgian | Georgian | Kartvelian | low | x |  | 979 | 350 |
| khm | Khmer | Khmer | Austroasiatic | low | x |  | 949 | 335 |
| kor | Korean | Korean | Koreanic | medium | x | x | 382 | 270 |
| lao | Lao | Lao | Tai-Kadai | low |  | x | 405 | 260 |
| lit | Lithuanian | Latin | Indo-European | low | x | x | 986 | 349 |
| mal | Malayalam | Malayalam | Dravidian | low |  | x | 985 | 344 |
| mar | Marathi | Devanagari | Indo-European | low | x | x | 1020 | 349 |
| nld | Dutch | Latin | Indo-European | high | x | x | 364 | 251 |
| pes | Persian | Arabic | Indo-European | low | x | x | 871 | 324 |
| rus | Russian | Cryrillic | Indo-European | medium | x | x | 775 | 344 |
| sna | Shona | Latin | Atlantic-Congo | low | x |  | 925 | 348 |
| snd | Sindhi | Arabic | Indo-European | low | x | x | 980 | 350 |
| swh | Swahili | Latin | Atlantic-Congo | low |  | x | 487 | 312 |

| Code | Name | Script | Family | Resource-Level | SeamlessM4T/ SALMONN | SONAR | Text Split Size | Without Duplicates |
|---|---|---|---|---|---|---|---|---|
| tam | Tamil | Tamil | Dravidian | medium | x | x | 591 | 336 |
| tel | Telugu | Telugu | Dravidian | medium | | x | 472 | 302 |
| tha | Thai | Thai | Tai-Kadai | medium | x | x | 1020 | 349 |
| tur | Turkish | Latin | Turkic | medium | x | x | 743 | 329 |
| yue | Cantonese | Hant | Sino-Tibetan | low | x | x | 819 | 339 |

**Table 4.2.: List of Analyzed Languages.** For each language, its language code, name, script, family, resource-level and on which models the language has been analyzed is given. The text split size is the number of sentence entries of a language in the FLEURS dataset. The number of unique sentences in the test split for each language is given under the 'Without Duplicates' column.

## 4.3. Configurations and Parameters

The code for the SVCCA computations[9] was provided by Raghu et al. (2017). To stabilize the similarity computations an `epsilon` of `1e-10` was applied. All speech and text representation sets were reduced to a target dimension according to the similarity analysis. The cross-modal target dimensions are listed in Appendix A.1 and in Appendix A.2 the target dimensions for the cross-lingual analysis can be found.

For the visualization analysis, the t-SNE library `sklearn.manifold.TSNE`[10] from scikit-learn was used, initialized with only the default values (`n_components=2, perplexity=30, early_exaggeration=12.0, learning_rate='auto', max_iter=1000, etc.`).

[9] https://github.com/google/svcca

[10] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

# 5. Results and Discussion

In the following chapter we present the results of the analysis explained in Chapter 3 and attempt to answer our research questions stated in Section 1.2. Firstly, we explore the cross-modal similarity results within one language (see Section 5.1), providing an in-depth analysis of on how the MLMs attempt to close the modality gap with the depth of its architecture, additionally analyzing the impact of language resource levels on the similarity between hidden speech and text representations. Following this, we delve into the cross-lingual similarity analysis results (see Section 5.2), presenting how the models perform with each language across or within modalities. Lastly, in Section 5.3, we visualize and explore the distribution of hidden speech and text representations, additionally drawing connections to the previous sections.

## 5.1. Cross-Modal Similarity Results

### 5.1.1. General Observations

The results of the cross-modal similarity analysis described in Section 3.1 applied on the pre-selected 30 languages and on all three models are shown in Figure 5.1. For each model, the cross-modal SVCCA similarities are consistently better than the baseline, which equal the SVCCA similarity of randomly initialized representation sets of the same size as other sets of the model. The higher cross-modal similarity than the baseline in all layers shows that the encoder/decoder of all three models are generally capable of capturing the shared features of speech and text inputs. Another consistent finding across all MLMs is the increase in the cross-modal similarity with the depth of the encoder/decoder. For SeamlessM4T and SONAR (see (1) and (2) in Figure 5.1), the similarity increase between in the first and last layer is approximately +0.04, while for SALMONN (see (3) in Figure 5.1) the increase is about +0.02, even though the similarity decreases in the last few layers of the decoder. Thus, the encoder/decoder of all MLMs are also capable of aligning shared information of speech and text inputs, not letting the difference in modality influence the hidden representations.

However, a drop in similarity in the early layers is visible in all three models, before recovering to the gradual increase. The lowest point is found in the fourth layer of the encoders of SeamlessM4T and SONAR, and likewise in the fourth layer of the SALMONN decoder. In both SeamlessM4T and SONAR, the decrease from the embedding similarity equals about $-0.039$ and $-0.028$, respectively, reaching the lowest similarity of all encoder layers. For SALMONN, the similarity drops from the second decoder layer with a decrease of $-0.039$. We assume that the drop in all models is caused by the different

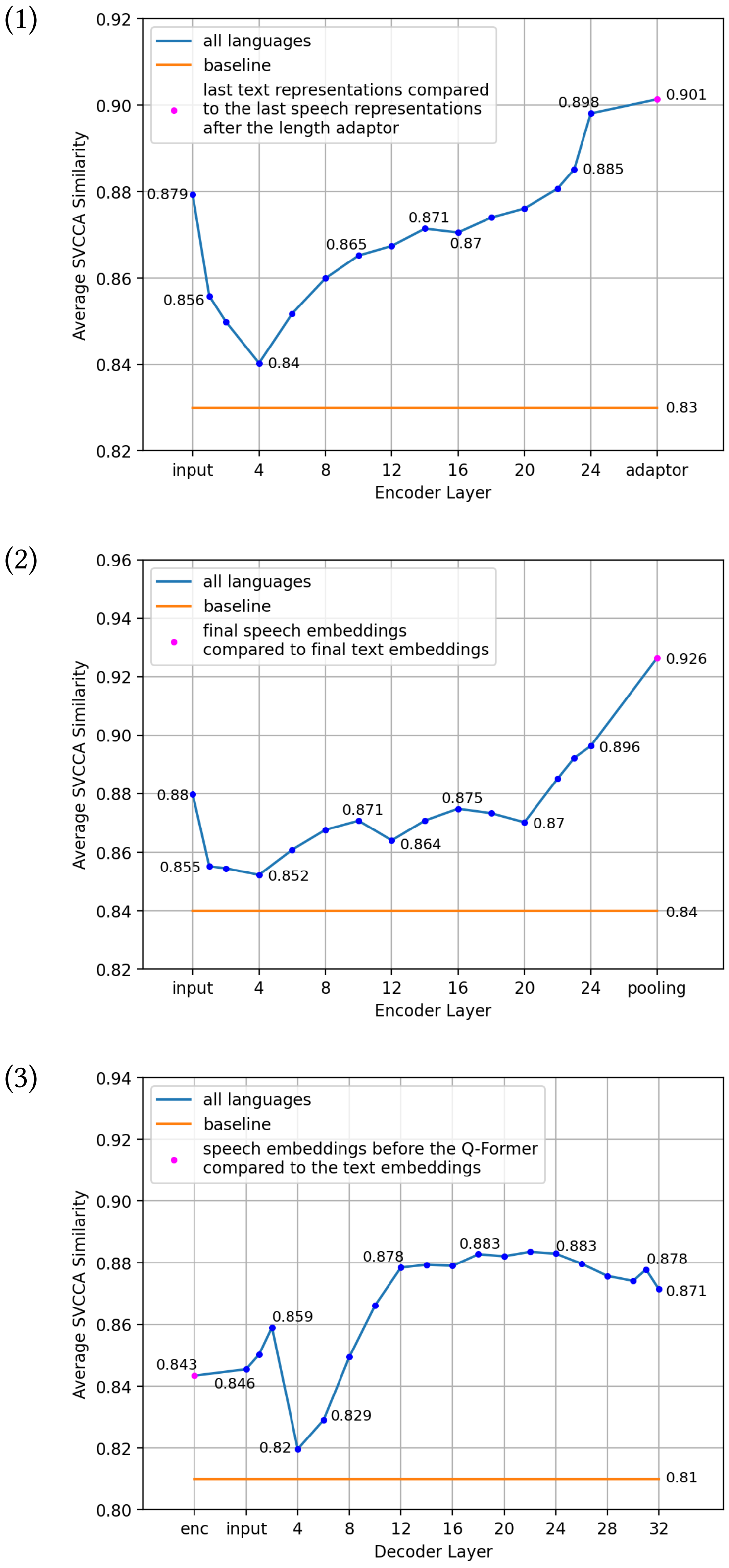


**Figure 5.1.: Cross-Modal Similarity Analysis Results For All Languages.** With (1) SeamlessM4T Encoder, (2) SONAR Encoder and (3) SALMONN Decoder.

role of the first few layers in comparison to the deeper layers. Since the earlier layers are highly influenced by the input and may focus on capturing low-level information and modality specific features, such as phonemes for speech or sentence structure for text inputs, early hidden representations hold a lower cross-modal similarity. After that, the encoder/decoder restructures the hidden representations converting them into more abstract, modality-independent representations, leading to the recovery.

Additional explanation on the course of each graph in Figure 5.1 and further insights into the factors that could lead to the reason behind these similarity scores can be found in Section 5.3, where the t-SNE results are presented and associated with the cross-modal similarity analysis.

### 5.1.2. Impact on Modality Gap

**SeamlessM4T** As mentioned in Section 4.1.1, the SeamlessM4T conformer speech encoder is post-fixed with a length adaptor, which should adapt the speech representations to the text representations for the input space of the text decoder. In the resulting graph of the cross-modality comparison of SeamlessM4T (see (1) in Figure 5.1), we can observe that the similarity between the text representations of the last encoder layer and the speech representations after the length adaptor is marginally higher than the cross-modal similarity before the length adaptor by +0.003, just barely pushing the similarity between the hidden speech and text representations to over 0.9. Even though +0.003 is a minor increase, the length adaptor is still a crucial component for the SeamlessM4T architecture. Since SeamlessM4T consists of separate speech and text encoders followed by a shared decoder, it is important for the length adaptor to be trained to preserve the high similarity across modalities by shortening the length-variable speech representations and aligning them to text representations.

**SONAR** Apart from the first drop mentioned in Section 5.1.1, the similarity of the SONAR speech and text representations experiences more drops in the 12th layer and the 20th layer of the encoder (see (2) in Figure 5.1). These drops are however not as striking as the first, as they do not fall under the similarity of the previous drop. We assume that this observation is occurring due to the same reason behind the drop in the fourth layer, as stated in Section 5.1.1. Since every layer in the encoder has different weights and thus different roles in processing the representations of the previous layer, some layers may focus on the input modality rather than producing abstract representations, resulting in a decline in similarity. These drops are however always followed by an increase in similarity above the previous peak, proving that both SONAR encoders are able to recover and produce more abstract and modality-independent speech and text representations.

Both SeamlessM4T and SONAR reach a fairly high similarity score in the last encoder layer, just falling behind of 0.9. However, with its pooling mechanism, SONAR performs better in closing the modality gap than the other two MLMs, reaching the final similarity of 0.926, which is about +0.025 and +0.055 higher than the final similarities of SeamlessM4T and SONAR, respectively. The pooling method along the sequence dimension shortens the representations to a fixed size, therefore increasing the cross-modal similarities by a very sharp increase of +0.03 from the last encoder layer. The MSE loss used in the SONAR embedding space additionally helps to increase the cross-modal similarity, since it better aligns sentences to the language-agnostic embedding space, by reducing the representation differences in both language and modality.

**SALMONN** In contrast to SeamlessM4T and SONAR, the cross-modal similarities in the decoder input space are lower than the similarity in the first decoder layer (see (3) in Figure 5.1). To be more precise, the similarity between the BEATs & Whisper speech encoder outputs and the Vicuna text embeddings are about −0.037 lower than the input embedding similarities of SeamlessM4T and SONAR. We assume the reason behind the low similarity is lies within the BEATs encoder. Even if the Whisper encoder is trained to distinguish speech from audio noise, the BEATs encoder subsequently magnifies noise, such as music and ambient noise, limiting the similarity between the speech and text representations of the same semantic meaning to increase. Additionally, the window-level Q-Former of SALMONN just marginally increases the cross-modal similarity (compare first two data points in (3) of Figure 5.1), as its only role is to downsize the audio encoder outputs to the Vicuna input space tokens, by maintaining the diverse audio features including noise.

Another difference to the previous MLMs is that the similarity scores do not gradually increase after the recovery from the drop in the forth decoder layer. Instead, the cross-modal similarity remains approximately at 0.88 from the 10th to the 26th decoder layer (see (3) in Figure 5.1). This is then followed by a small decrease, reaching the final similarity score of 0.871. The reason behind this decrease lies in the nature of decoders taking abstract representations from the encoder to generate outputs in a specific modality or language, resulting in more modality- and language-specific representations. Since SALMONN, unlike the other two MLMs, also processes music and ambient noise in addition to speech with the BEATs encoder, these various audio features are still evident in the speech representations, limiting the cross-modal similarity to increase. Additionally, SALMONN uses the text-based LLM Vicuna that follows text instructions based on audio inputs, instead of translating speech and text inputs like SeamlessM4T or producing language-agnostic embeddings like SONAR. Its priority is therefore not to increase the cross-modal similarity, rather engaging in the modality specific features of the inputs.

### 5.1.3. Impact of Language Resource Levels

**SeamlessM4T** The course of the SeamlessM4T graph of each three language resource levels (see Figure 5.2) are similar to that of all 30 languages combined: A major drop in similarity from the input space to the fourth encoder layer, followed by a gradual increase until the highest similarity score is reached. However, the influence of the language resource level is visible through the variation of the similarity scores based on the resource level. While the similarities of the high and medium resource languages are overlapping in almost all encoder layers and are just slightly higher than the averaged similarities of all 30 languages (see (1) of Figure 5.1), the similarities of the low resource languages are noticeably lower than the other resource levels, with the lowest similarity in the fourth layer falling below 0.84 and the highest similarity score in the last encoder layer not reaching 0.9. This proves that the cross-modal similarity depends on how extensive the encoder has been trained on languages of different resource levels. As high resource languages have a high volume of available training data than low resource languages, the encoder can accurately extract the shared features of the speech and text representations, resulting in higher similarity scores.

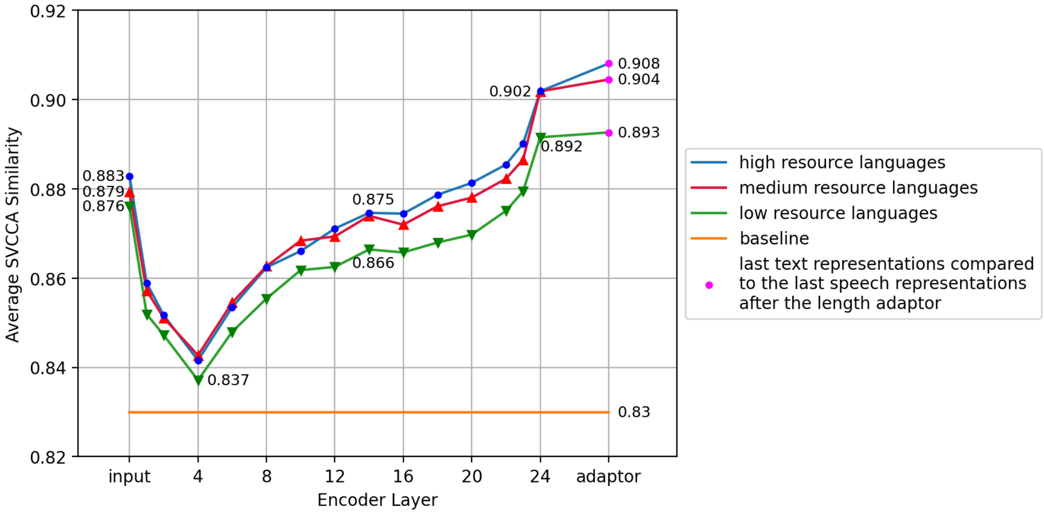


**Figure 5.2.: SeamlessM4T Cross-Modal Similarity Analysis Results For Each Resource Level.**

However, the highest similarity, which is the score of the last encoder layer, is not proportional to the resource levels, as both high and medium resource cross-modal comparisons reach 0.902. This may be a result of regularization techniques used by SeamlessM4T in the training phase (e.g. dropouts), to prevent the speech and text encoders from overfitting to languages with large amount of training data. Regularization ensures generalization across languages with different resource levels, and encourages the SeamlessM4T encoders to prioritize accurate modality-independent hidden representations over perfect similarity scores for high and medium resource languages, therefore limiting them to 0.902. Nevertheless, it is also important to mention that cross-modal similarities should necessarily completely close the modality gap by reaching similarities of 1.0, as this would mean that models are not capable of capturing language specific feature in their representations.

In contrary, the increase in similarity after the length adaptor is proportional to the language resource level. With only high resource languages, the SeamlessM4T length adaptor increases the similarity by +0.006 from the last encoder layer. This is three times more of that what the length adaptor can reach for medium resource languages, and for low resource languages six times of its amount. Even if these increases differ only marginally, these results show the how high quantity and quality of training data allows the length adaptor to accurately align speech representations to those of text inputs, as it has a better understanding of how to shorten speech sequences to text, resulting in higher cross-modal representation similarity scores for higher resource languages.

**SONAR** The course of the SONAR graph is the same across all four analyses with different language sets. Same as SeamlessM4T, the similarities of each layer are slightly shifted according to the language resource level (see Figure 5.3). While the results on all 30 languages with varying resource levels (see Figure 5.1) resembles the medium resource language graph, the influence of language resource levels are observed in the last few SONAR encoder layers. The similarities of high resource languages are at maximum by +0.009 higher than the same values of the analysis on all languages, while the similarities of low resource languages is at maximum lower by −0.006. Since higher quantity and quality of training data of higher resource languages ensure the ability of the SONAR encoders to accurately extract the shared semantic features of speech and text inputs, the SONAR cross-modal similarity for high resource languages experiences a stronger increase in similarity from the 12th to the 16th layer, following with a very minor decrease in the 20th encoder layer, ultimately reaching the highest similarity score of 0.905 in the last encoder layer across all SONAR analysis. The decrease in the 20th layer is consecutively stronger with low resource languages, almost reaching the same level as the previous drop in the 12th layer, making the recovery in the 16th layer insignificant. With low resource languages, SONAR only reaches the similarity score of 0.89, −0.015 lower than the score of the same layer with high resource languages.

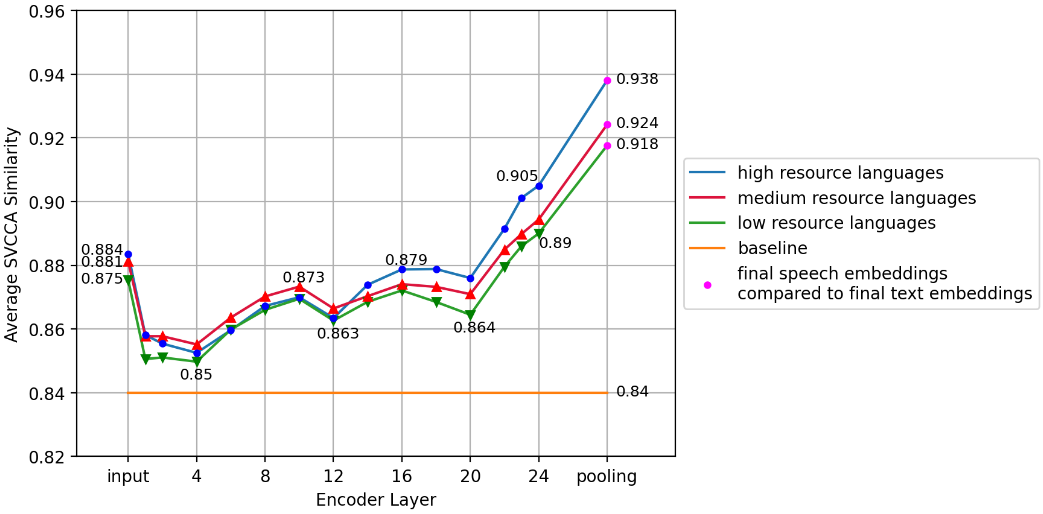


**Figure 5.3.: SONAR Cross-Modal Similarity Analysis Results For Each Resource Level**

Same as the SONAR analysis for all 30 languages (see (2) in Figure 5.1), the pooling methods also further closes the modality gap between the hidden speech and text representations for all three language resource levels. However, the increase from the last encoder layer to the shared embedding space is similar across all language resource levels, being approximately 0.03. This can be explained with the pooling methods effectively smoothing out all features that does not add value to the semantic meaning of the representation. The embeddings therefore become insensitive to different language resource levels to create consistent language-agnostic embeddings, resulting in similar increases. Even though the increase in similarity in the embedding layer is not influenced by the language resource levels, varying final similarity scores can be still achieved by the effort of the SONAR encoders to accurately align speech and text representations, making SONAR to achieve the highest final cross-modal similarity across all three analyzed models with it being just slightly below 0.94 for high resource languages (see Figure 5.3).

**SALMONN** Same as SeamlessM4T and SONAR, the course of the similarity graph does not change with varying language resource levels (see Figure 5.4). The overall graph is only shifted upwards or downwards depending on the resource level, also with the same reasoning of SeamlessM4T and SONAR. The SALMONN analysis on medium resource languages is the most similar to the similarity scores on all 30 languages, with the similarities only varying on average by +0.003. The influence of the resource levels are more visible in the layers before and in the last few layers of the decoder.

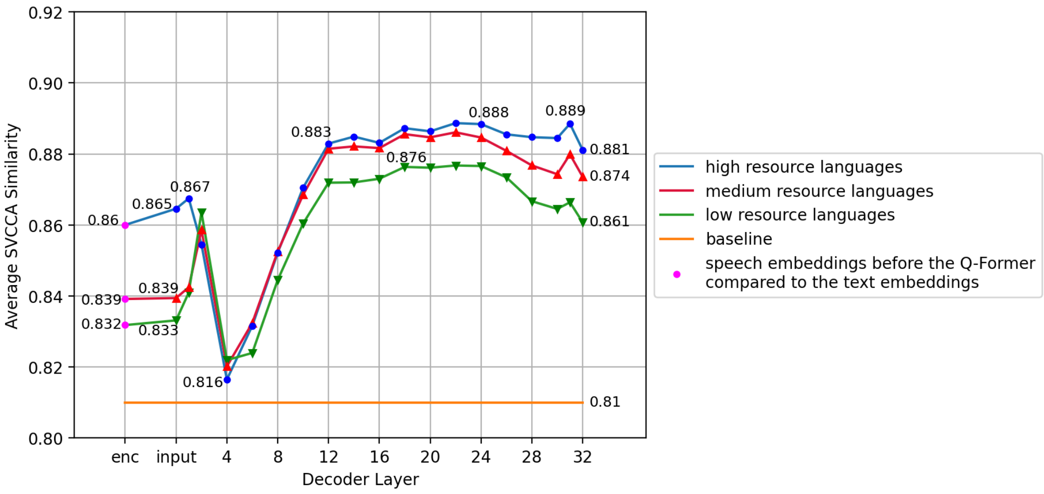


**Figure 5.4.: SALMONN Cross-Modal Similarity Analysis Results For Each Resource Level.**

The cross-modal similarities before the Q-Former are highly dependent on the language resource level. SALMONN achieves a high cross-modal similarity only with high resource languages, almost being at the same level as the similarity of the first decoder layer (see Figure 5.4). Compared to the medium and low resource languages, this similarity is respectively +0.021 and +0.028 higher. This observation is heavily influenced by the Llama2 model (Touvron et al., 2023), on which is the Vicuna model of SALMONN is based. Llama2 has been predominantly trained on high resource languages, more precisely, 90% of the whole pre-training data are English (Touvron et al., 2023, Chapter 5.2). As a result, the

Vicuna embedder performs best for English and other high resource languages in capturing linguistic and semantic features compared to other resource levels.

We also assume that the Whisper encoder contributes partly in increasing the embedding similarity, as Whisper is also mostly trained on English and other high resource language speech recognition data (Radford et al., 2023, Appendix E) with English making about 65% out of all training data. Whisper also states that it achieves <50% word error rate (WER) in multilingual transcription for all medium and for some low resource languages of this work (Radford et al., 2023, Appendix D.2). However, the amount of hours of medium resource training data differs drastically compared to the amount of English data. For example, while for English there are approximately 440,000 hours of speech recognition data, there are only 41 and 12 hours for Estonian and Hindi, respectively. For low resource languages, the amount of training data lies below 10 hours or Whisper was entirely not trained on them (Radford et al., 2023, Appendix E). Consequently, due to both Whisper encoder and Vicuna failing to extract accurate features from languages that are not English or from any other high resource language, the cross-modal similarity before the Q-Former is lower than those of medium and low resource languages.

Same as the SALMONN analysis with all 30 languages, a similarity increase with the Q-Former on medium and low resource languages does not exist or is barely noticeable, compared to the increase of +0.005 with high resource languages. Since the speech encoders of SALMONN are most likely not accurate for medium and low resource languages, the Q-Former is not able to add to the similarity just by downsizing the speech sequence of varying lengths to a match text sequences with preserving the important language and semantic features, as there is no accurate representation of the speech inputs to begin with.

For the same reason behind the varying similarities before the Q-Former, the previously mentioned decrease starting from the 26th layer in Section 5.1.2 is more prominent in the SALMONN analysis on low resource languages and barely noticeable with high resource languages. Thus, the similarity for high resource languages stays above 0.88, which is the highest final similarity across the SALMONN analysis on varying resource levels. Due to Vicuna and Whisper being mostly an high resource language model and more capable in processing English, we assume that these languages prevent the cross-modal similarity to decrease after the 26th decoder layer.

<table>
<tr><th></th><th>SeamlessM4T</th><th>SONAR</th><th>SALMONN</th></tr>
<tr><td>Common Observations Across Models</td><td colspan="3">• cross-modal similarity of the encoder/decoder representations is higher than a random baseline<br>• cross-modal similarity increases with the depth of the encoder/decoder<br>• course of cross-modal similarity does not change with varying language resource levels<br>• overall cross-modal similarity is proportional to the language resource level<br>→ Each model is capable of capturing shared semantic information independent of the input modality.<br>→ The quantity and the quality of the training data proportionally impacts the cross-modal similarity.</td></tr>
</table>

| | SeamlessM4T | SONAR | SALMONN |
|---|---|---|---|
| Final Cross-Modal Similarity (all/high/medium/low) | (0.901/0.908/0.904/0.893) | (0.926/0.938/0.924/0.918) | (0.878/0.881/0.874/0.861) |
| Impact of ... | Length Adaptor<br>• marginal similarity increase<br>• increase proportional to language resource level | Pooling<br>• steep similarity increase due to the MSE loss in the language-agnostic embedding space<br>• increase same in all resource levels | Window-Level Q-Former<br>• marginal similarity increase<br>• increase only visible with high resource languages |
| | → The cross-modal similarity is also affected by the architecture and the training setup of the model. | | |

**Table 5.1.: Summary of the Cross-Modal Analysis Results.**

## 5.2. Cross-Lingual Similarity Analysis Results

### 5.2.1. General Observations

To examine how the cross-lingual similarity within one modality changes with the depth of the model's architecture, we averaged the similarity scores of every possible language pair for each layer and modality. We did not add the similarities of the same-language pairs to the calculations, as they all equal 100.0 and do not add value to the final results. The averaged cross-lingual similarities within each modality are shown in Figure 5.5.

One common observation across all three models is that the cross-lingual intra-text similarities for each encoder/decoder layer are higher than the intra-speech similarities, meaning that each model is more capable of closing the language gap within the text modality. For SeamlessM4T and SONAR, the difference is visible through all encoder layers, while for SALMONN it is more visible in the decoder input space and the last half of the decoder layers. The intra-speech similarity start at a relatively low score compared to those of intra-text comparisons, even though all cross-lingual comparisons are based on inputs of the same intersecting semantic meaning. For SeamlessM4T and SONAR, the scores start at around 0.63, while for SALMONN it starts at around 0.79. We assume that this higher similarity of SALMONN is caused by the two speech encoders having more resources than the SeamlessM4T and SONAR embedders to capture the important features of the speech inputs.

The reason behind the higher initial cross-lingual similarity within the text modality lies in the speech inputs being more versatile than the text inputs. Compared to the static and normalized transcriptions, where the language and the meaning of words can be directly identified, the speech inputs vary with different speakers and audio noise, making it difficult to capture the language and semantic meaning. In addition, since languages produce different audio features and the embeddings in the encoder/decoder input space are still highly tied to low-level features, the cross-lingual intra-speech similarity is limited. However, the encoder/decoder of each model is capable of handling with different speech data, increasing the final intra-speech similarity to reach approximately the same level as

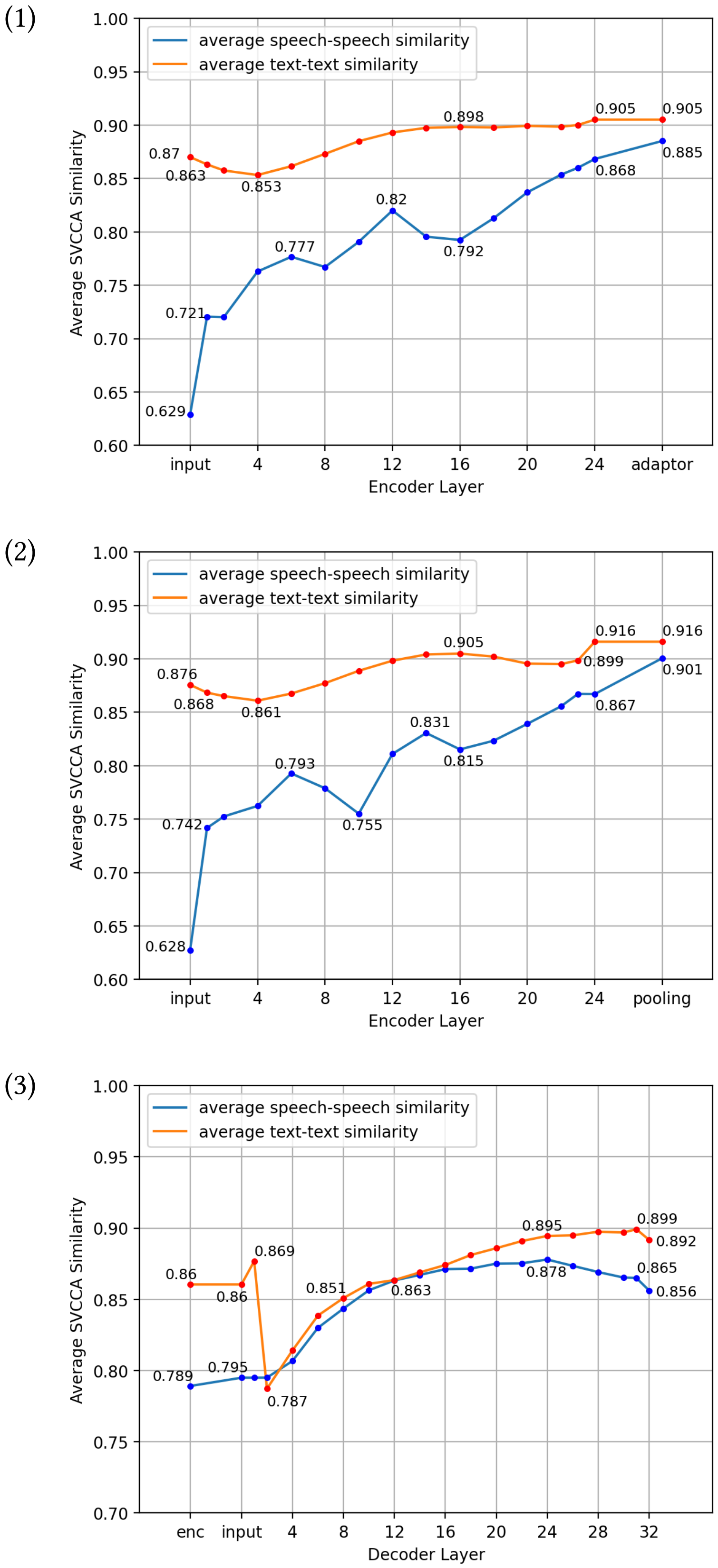


**Figure 5.5.: Averaged Cross-Lingual Similarities for Same Modality Comparisons.** With (1) SeamlessM4T, (2) SONAR and (3) SALMONN.

the final intra-text similarity. In contrast, all three models only needed a small increase to reach the final intra-text similarity, since the stating similarities were high to begin with.

Same as the results of the cross-modal similarity analysis in Section 5.1, all cross-lingual similarities within one modality increase with the depth of the encoder/decoder. Due to both speech and text encoders of SeamlessM4T and SONAR being initialized with the same models (see Section 4.1.1), they return similar graphs (see (1) and (2) in Figure 5.5). Nevertheless, it is important to mention that both final cross-lingual similarities of SONAR are at least +0.011 higher than those of SeamlessM4T.

For text inputs, SONAR reaches a high level of language-independence in the last encoder layer. The mean pooling does not increase the cross-lingual intra-text similarity, as it only downsizes the last encoder representation to match the shared embedding space of SONAR. The learning pooling with the embedding space MSE loss on the other hand increases the intra-speech similarity by downsizing the last encoder speech representation and emphasizing semantic meaning to match the language-agnostic embedding space.

The length adaptor does also further close the language gap, however not as effective as the pooling method of SONAR, resulting in a final cross-lingual intra-speech similarity score of under 0.9. In contrast to the other speech sequence shortening methods, the Q-Former of SALMONN only marginally increases the intra-speech similarity (see (3) in Figure 5.5), similar to the cross-modal results of SALMONN in Section 5.1.2.

However, evident in all three models, the language gap is more closed in the cross-lingual comparisons within the text modality than speech. Even if the cross-lingual similarities of intra-speech comparisons increase with the depth of the encoder/decoder, we assume that the noise and various audio features from the speech inputs still remain in the hidden representations, limiting their similarity to increase like the intra-text inputs.

Comparing these results to the cross-modal results in Figure 5.1, we can observe that it is different from MLM to MLM in which conditions the model can close the similarity gap the most. SeamlessM4T is more capable of closing the language gap for cross-lingual text inputs and the modality gap for same-language inputs, as both final similarity scores reach above 0.9 (see (1) in Figures 5.1 and 5.5). We assume the reason behind this observation is that SeamlessM4T can process the normalized text inputs more accurately with capturing the shared features in the hidden representations, since text is more stable and static compared to speech. Additionally, while the cross-modal similarity only increases marginally after the length adaptor, the cross-lingual intra-speech similarity increases by about 0.02 after the length adaptor. This is due to the role of the length adaptor to shorten and align speech representations of varying length to the text representations, making speech representations more comparable across languages, but it does not bridge the gap between different modalities.

Contrary to SeamlessM4T, SONAR performes better in closing the modality gap of same language inputs, as both final cross-lingual similarities within one modality are lower than the final cross-modal similarity for same-language inputs (see (2) of Figure 5.1). This result shows us that even though SONAR aims to produce language-agnostic embeddings, language features are still evident in the final embeddings, making the final cross-modal similarity at maximum by +0.025 higher than the cross-lingual similarities within one modality. Another difference between SONAR and SeamlessM4T lies within the similarity increases after the last encoder layer. While the increases of the length adaptor differs in the cross-modal and cross-lingual analysis, the similarity increase after the pooling method of SONAR remain approximately the same in both analysis.

For SALMONN, the cross-lingual intra-text similarities are the highest by about +0.032 higher than the cross-lingual intra-speech similarity scores and +0.021 higher than the final cross-modal similarity (see (3) of Figure 5.1). We assume this is due to the Vicuna model, which SALMONN is based on, originally being a text based model. Even though Vicuna was fine-tuned to process speech inputs in the form of BEATs and Whisper encoder outputs, the final cross-lingual intra-speech similarity is far less of what the other MLMs can achieve. Additionally, the cross-lingual similarities within one modality face the same consistency in similarity in the last few layers of the decoder, especially more noticeable in the intra-speech analysis. The reason behind this is the same as the one mentioned in Section 5.1.2. Since SALMONN main purpose is not to generate translations or produce modality- and language-independent embeddings, it does not aim at capturing the shared semantic meaning and to achieve a high cross-lingual similarity, but rather at engaging in the unique features of the inputs to generate answers for the text instructions.

Further insights on the course of the cross-lingual similarity can be found with our t-SNE results in Section 5.3, where we explain the potential causes of the observations in this section, such as the impacts of the pooling methods of SONAR and the major drop in similarity after the first decoder layer of SALMONN.

### 5.2.2. Impact of Languages Across Modalities

For all three MLMs, the cross-lingual similarities of the input embeddings, the first and the last layer of the encoder/decoder and lastly, the model specific components handling length-variant speech sequences were taken into analysis. The similarities for each language pair of all 30 languages are presented in heatmaps listed in Appendix A.3 with all similarity scores multiplied by 100 for better visualization.

**SeamlessM4T** The cross-lingual similarity analysis results of SeamlessM4T in Figures A.1 and A.2 resemble the cross-modal results of Section 5.1, as the similarity ranges of the analyzed layers stay around the cross-modal similarity scores of the corresponding layers, as shown in graph (1) in Figure 5.1. The cross-lingual scores in this section are higher or lower than the score evaluated on all 30 languages, depending on the language pair.

In the input embedding space of SeamlessM4T, there is no defined structure in the similarity distribution, as the embeddings are still highly influenced by the input data varying in modality and language. For example, non-related languages like Estonian and

German have a higher similarity score of 90.1 than more linguistically similar languages, such as German and Dutch with 88.6 (see (1) in Figure A.1). After the input embeddings enter the encoder, the similarity decreases for all language pairs, as noticed in Figure 5.1. There is still no structure visible in the first encoder layer (see (2) in Figure A.1), however the diagonal of the heatmap, which holds the similarities from the same-language pairs, is more noticeable than before, as it shows higher similarities relative to the other comparisons. This observation shows us that the SeamlessM4T encoders are able to recognize and capture the language the speech and text inputs are based on already in the first hidden representations, resulting in higher similarities on the diagonal.

In the last encoder layer of SeamlessM4T, an increase in similarity is visible for all language pairs (see (1) in Figure A.2). The most noticeable increases are on the diagonal of the heatmap (see (1) in Figure A.3), since both representation sets are based on the same language. Differences in the input language besides the different input modality therefore adds more complexity to the comparison, resulting in lower cross-lingual similarities across modalities. It can also be observed that SeamlessM4T performs the better in closing the language gap with English, as all cross-lingual comparisons with English, regardless of which modality, have higher increases from the first encoder layer. The reason behind this observation can be explained with the amount of English training data and the strategy used to fine-tune the SeamlessM4T encoders. Since English is a high resource language, SeamlessM4T was able to use a large amount of ASR data to train the model (Communication, Barrault, Chung, Mariano Coria Meglioli, et al., 2023, Appendix I.2). In addition, the X-to-text model of SeamlessM4T, which includes both speech and text encoders, was fine-tuned on the X–eng and eng-X translation directions (Communication, Barrault, Chung, Mariano Cora Meglioli, et al., 2023), resulting in higher cross-lingual similarity scores with English.

Additionally, same as what we have examined in Section 5.1.3, the increase in similarity is proportional to the language resource level, as cross-lingual comparisons including high resource languages have higher increases from the first to the last encoder layer than other resource levels (see (1) in Figure A.3), resulting in cross-lingual comparisons with only high resource languages to hold higher similarities in the last encoder layer (see (1) in Figure A.2). However, differences in similarity except for the diagonal in the last encoder layer are very small, meaning SeamlessM4T has also been extensively trained on the selected medium and low resource languages to bridge the language barrier, capturing the cross-lingual shared semantic meanings for all language pairs.

The lack of training data is noticeable in the minor increases from the first encoder layer to the last in all comparisons including the speech representations of Shona (sna) and Sindhi (snd) (see (1) in Figure A.3). This is caused by SeamlessM4T not being sufficiently trained on these languages, since the speech-to-text and speech-to-speech translation tasks with the two languages being the source language was trained on zero-shot. The semantic meaning of the Shona and Sindhi speech inputs are therefore not accurately captured in the hidden speech representations, resulting in lower similarity scores.

As seen in Section 5.1.2, the length adaptor only minimally changes the cross-modal similarity scores. This observation is also evident in the cross-lingual comparisons, as most differences range between −0.8 and +0.8. This is due to the goal of the length adaptor to align the length of speech representations to text, not primarily further extracting the cross-lingual shared semantic meaning. However, for the two zero-shot languages, the length adaptor only to decreases the similarity at maximum by −1.4 and achieves the opposite of what it aims for (see (2) in Figure A.3). We assume this is the result of the length adaptor's main purpose and the fact that it was not trained sufficiently on the zero-shot languages. While adapting the speech representations, it might falsely discard the few shared features that was left from the speech encoder while downsizing the speech features, resulting in the decrease in similarity.

**SONAR** SONAR and SeamlessM4T return very similar cross-lingual results until the last encoder layer, since the speech and text encoders of both MLMs are based on the same models, similar of what we have observed previously in the previous analysis in Sections 5.1.2 and 5.2.1. More precisely, the cross-lingual similarity results across modalities of the input embeddings and the first encoder layer of SONAR also have no define structure (see (1) and (2) in Figure A.11) and the cross-lingual similarities of the same language pairs in the last encoder layer are also higher compared to the other comparisons (see diagonal in heatmap (1) of Figure A.12).

One difference is that the increases from the first encoder layer to the last layer highly depend on the language resource levels of the speech representations, as the columns of the high and medium resource languages (e.g. English, Finnish, Japanese) have generally higher increases compared to the low resource level columns in (1) of Figure A.13, resulting in the columns of the same high and medium resource languages to have higher similarity scores in the last encoder layer (see (1) in Figure A.12). This shows is that the cross-lingual similarities across modalities are dependent on the quality of the speech representations of the encoder and how much shared semantic meaning has been captured in them. Since speech is more variant than normalized text transcriptions, the SONAR speech encoder has to be capable of capturing the true semantic meaning behind the language and the modality. With high and medium resource languages, the encoder can be sufficiently trained on capturing these features, resulting in higher cross-lingual similarity scores.

Even though SONAR is a model developed to produce language-agnostic sentence embeddings independent from the input modality and language, the unique language features are still evident in the representations after the pooling, resulting in higher increases and similarity scores of the language pairs along the diagonal (see (2) of Figures A.12 and A.13), where the speech and text representations of the same language are compared. Same as SeamlessM4T, representations varying in language and modality brings more complexity to the comparison, resulting in the similarities on the diagonal to reach at maximum 95.8 while the cross-lingual similarities reach at maximum 93.4. Nevertheless, SONAR reaches the highest overall final cross-lingual similarity scores across modalities compared to SeamlessM4T and SALMONN (see Figures 5.1 and 5.5), achieving language-agnostic embeddings to a fairly high extent.

**SALMONN** Same as SeamlessM4T and SONAR, not linguistically similar languages hold higher similarities in the decoder embedding space before the Q-Former, as the text embeddings focus on character-level features. Nevertheless, a pattern is visible in the distribution of the cross-lingual similarities between the speech encoder outputs and text embeddings (see (1) of Figure A.21). The similarity scores depend on the Vicuna text embeddings, however they are not proportional to the language resource level as we could have expected, since the similarities of cross-lingual comparisons including text embeddings of medium and low resource languages are generally higher than those including high resource languages. We assume that this is due to Vicuna being mainly trained on English and some other high resource languages (Touvron et al., 2023, Chapter 5.2), which results in its text embeddings of medium and low resource languages to be highly generalized to match the high resource languages Vicuna was trained on, increasing the similarity scores across all languages.

The same aforementioned observations of the encoder outputs are also visible after the window-level Q-Former (see (2) in Figure A.21), with the similarity between the speech and text representations of the same language additionally being higher compared to the other comparisons, due to the language features being evident in the representations. The overall cross-lingual similarity barely changes after the Q-Former, as seen in the marginal differences in the heatmap of Figure A.23, matching our observations in Sections 3.1 and 3.2.

After the first decoder layer, the complete rows of medium and low resource languages holding higher similarities are not as visible as before (see (1) in Figure A.22), even though they have increased in similarity by an average of 2.0. Instead, the cross-modal same-language comparisons of the languages with generalized text embeddings hold the highest similarities (e.g. Amharic, Greek and Tamil), increasing by a maximum of +6.4 from the decoder input embeddings to the first decoder layer (see (2) in Figure A.23). We assume that the first layer of the SALMONN decoder is able to capture some language features out of the input embeddings limiting the similarities to rise for cross-lingual comparisons including the generalized text representations of medium and low resource languages. However, the addition to the language features on top of the generalized text representations increases the similarity score of medium to low resource languages when compared to the speech representations of the same language.

Fortunately, with the depth of the decoder, SALMONN seems to recognize overly generalized text representations, as there is a noticeable decrease in similarity for those languages from the first decoder layer to the last, while for the similarities with the text representations of high resource languages decrease marginally (see Figure A.24). As a result, the cross-lingual similarity across modalities is proportional to the language resource levels in the final decoder layer (see (2) of A.22), as comparisons with high resource languages, such as English and French, hold higher similarities than with medium and low resource languages. This proportionality also matches with our cross-modal similarity results previously seen in Section 5.1.3. Same as SeamlessM4T and SONAR, the cross-modal similarities across the diagonal of the heatmap of the last decoder layer are significantly higher than the cross-lingual similarities, with the similarity of English being the highest. The language features of the input are therefore still evident in the decoder output of SALMONN and since Vicuna was primarily trained on English, the decoder is able to further close the modality gap for English.

### 5.2.3. Impact of Languages Within One Modality

After analyzing the cross-lingual similarities of all 30 languages across modalities for each model, the same analysis can also be conducted on the cross-lingual comparisons within one modality for the same layers. The results of this analysis only contain the similarity scores of the lower triangle, due to the symmetry of the similarity matrix (see Appendix A.3). Since the diagonals in the results of this section are similarities between the same representation set and language, they are normalized to 100.0 and are not included in the heatmap scale. The averaged similarity scores without the diagonal equal the scores in the graphs of Figure 5.5.

**SeamlessM4T** For SeamlessM4T, the cross-lingual intra-speech analysis presents similar results as the cross-lingual analysis across modalities throughout each observed SeamlessM4T layers, meaning that the cross-lingual and inter-modal analysis (see Figures A.1 and A.2) is highly dependent on the information the speech representations carry. Especially, the randomness of the similarity distribution of the input embeddings and in first encoder layer (see Figure A.4) and the proportional similarity to the resource levels in the last two SeamlessM4T layers (see Figure A.5) are visible in both cross-lingual analysis results.

The only difference lies in the inter-layer similarity increases, as shown in Figures A.6 and A.7. As mentioned in Section 5.2.1, the cross-lingual intra-speech similarities start at a relatively low score, which increases with the depth of the encoder. Contrary to the cross-lingual comparisons across modalities in Figure A.3, high increases can be seen between the input embeddings and the first encoder layer and also between the first and last encoder layer. However, the former does not add value to the results, since the increases also have no define structure, and not linguistically similar language pairs such as Italian with Arabic have the highest increases (see (1) in Figure A.6). Instead, the increases between the first and last encoder layer is proportional to the resource levels of the language pairs, since comparisons with the zero-shot languages Shona (sna) and Sindhi (snd) have the lowest increases and those with English have the highest, similar to the cross-lingual increases across modalities in heatmap (1) of Figure A.3.

In contrast to the cross-lingual and inter-modal similarity decreases for low resource languages after the length adaptor in Figure A.3, the length adaptor only increases the similarities in the cross-lingual intra-speech analysis by at least 1.0, as well as the similarities of the language pairs with zero-shot languages (see Figure A.7). This consistent with our results of the averaged cross-modal similarity in Figure 5.5 in Section 5.2.1, where the reason behind this observation can be found.

The cross-lingual intra-text comparisons do not have many peculiarities either, with most observations already stated in Section 5.2.1: High initial similarity in the input embedding space, followed by small gradual increases though the depth of the encoder. Same as the cross-lingual intra-speech analysis, the similarity scores are proportional to the language resource level throughout all analyzed layers (see Figures A.8 and A.9). For example, while the cross-lingual comparisons with English have the highest scores overall, low resource languages such as Amharic, Khmer and Shona have the lowest.

One striking difference is that the SeamlessM4T text encoder performs differently for the Mandarin Chinese and Cantonese. In the input embedding space and the first encoder layer, the intra-text similarity between Mandarin and Cantonese is the highest, just falling behind 90.0 (see Figure A.8). This observation is most likely due to their linguistic similarity in their written from. As Mandarin and Cantonese share a significant amount of vocabulary, they also share tokens in the embedding space. This causes their similarity in the encoder input space to be higher than with other languages. Even after the first encoder layer, the similarity stays high, since early hidden representations are more tied to the character level features rather than the semantic meaning.

This is nevertheless reversed with the depth of the text encoder, as SeamlessM4T recognizes the differences between the two languages, resulting in a minor decrease in similarity (see (2) in Figure A.10), while all other comparisons rise in similarity. However, all cross-lingual similarities involving Mandarin and Cantonese have a smaller increase from the first encoder layer to the last. For Cantonese, we can argue that it is a low resource language, but this is not the case for the high resource language Mandarin. We assume that the SeamlessM4T text encoder has some difficulties in understanding languages of the Hant family due to their complex syntactic and morphological structures compared to other languages like English or French, which have the highest similarities in the last encoder layer.

**SONAR** As SONAR and SeamlessM4T are initialized with the same speech and text models, the cross-lingual intra-modality similarity analysis results in the encoder input space and after the first encoder of SONAR are very similar to those of SeamlessM4T (see Figures A.14 and A.18). Differences between the two models are only visible after the last encoder layer, since both models are trained on different tasks.

Same as the last speech encoder layer of SeamlessM4T, the SONAR intra-speech similarity scores in the last encoder layer are proportional to the resource level of the language pair, with some exceptions (see (1) in Figure A.15). Foremost, all comparisons with the low resource languages Assamese and Sindhi have lower increases from the first encoder layer (see (2) in Figure A.16) and as a result, have lower similarities than other comparisons. In contrast, we can observe exceptionally high similarity increases for certain language pairs, such as Dutch-German, Catalan-French/Italian, Bulgarian-Russian and Estonian-Finnish (see (2) in Figure A.16). Since the aforementioned language pairs each share the same script and are linguistically very similar, SONAR is able to produce similar speech representations based on their shared semantic meaning, minimizing the language gap only for linguistically similar languages (see (1) in Figure A.15).

As mentioned before, the cross-lingual intra-text analysis results of SeamlessM4T and SONAR are mostly the same, as the same observations in the SeamlessM4T part of Section 5.2.3 are also visible in the SONAR results, with the same reasoning behind it. The only difference is that apart from the low increases and similarity scores of language pairs including Mandarin and Cantonese in the last encoder layer, all comparisons with the low resource languages Assamese and Sindhi also have smaller increases from the first encoder layer to the last (see (2) in Figure A.20) and therefore lower similarity scores than the rest (see Figure A.19).

The pooling method of SONAR does not change the cross-lingual intra-text similarities, as seen in Section 5.2.1, but averages the similarities in the cross-lingual intra-speech comparisons, giving these high-similarity language pairs smaller increases than those with low similarity (see (2) in Figure A.17). For both intra-modal analysis, SONAR reaches final cross-lingual similarities of about 88.0 − 94.0 (see (2) in Figure A.15 and Figure A.19), achieving its goal of producing language-agnostic embeddings within one modality to a high extent. However, influences of language resource levels are unavoidable even with SONAR, as for both final cross-lingual and intra-modal similarities, the comparisons with English have the highest scores, while it is the lowest for low resource languages such as Assamese and Sindhi.

**SALMONN** Same as SeamlessM4T and SONAR, the distribution of the cross-lingual and intra-speech similarities in the decoder input space before the Q-Former have no distinct structure, with non-related language pairs such as Bulgarian-Finnish having the highest similarity score (see (1) in Figure A.25). However, the reason behind this observation is not the same as the other two models (see Section 5.2.2), as the varying input features are first processed through the Whisper and BEATs encoder. We assume that the lack of structure is caused by the BEATs encoder outputs, which is trained to capture background audio noise, magnifying uncertainty to the Whisper encoder outputs.

The overall intra-speech similarity does not increase drastically, as seen previously in Figure 5.5, but the minimal increases from the encoder outputs to the outputs after the Q-Former depend on the language resource level (see (1) in Figure A.27). As stated in Tang et al. (2024), a large amount of speech and audio data was used to close the gap between the pre-trained components of SALMONN and the Q-Former. We assume that the Q-Former was trained on similar languages as the Whisper encoder and Vicuna, meaning the training data consists mostly of high and medium resource languages. The Q-Former can therefore accurately align speech representations for these languages with leaving the shared semantic features behind, resulting in higher increases compared to low resource languages such as Amharic and Georgian. This pattern in the cross-lingual intra-speech similarity scores does not change with the depth of the decoder as the similarity increases (see Figures A.25 and A.26), meaning that the fine-tuned Vicuna model is not able to accurately capture shared semantic features if the language comparison pair includes a low resource language Vicuna was not sufficiently trained on.

Which languages Vicuna was mostly trained on is more evident in the cross-lingual intra-text analysis, because these languages have a very high similarity score in the Vicuna input space. As seen in heatmap (1) in Figure A.29, comparisons of only high resource European languages with Latin and Cyrillic script, such as Catalan, English, German and Russian, have higher similarity scores than others. The generalization of low resource language text inputs, mentioned in Section 5.2.2, is therefore limited in the cross-lingual intra-text similarity, since embeddings of high resource languages that capture the shared semantic features more accurately.

For text embeddings, the fined-tuned Vicuna model of SALMONN flattens the cross-lingual similarities, which can already be observed in the disproportionate increases to the language resource level in the first decoder layer (see (1) of Figure A.31), and consequently reaching cross-lingual intra-text comparisons with less drastic differences in similarity than before (see Figure A.30). However, similar to the intra-speech analysis, a bias for high resource and linguistically similar languages is still visible, meaning SALMONN can only further close the language gap if these conditions are met.

<table>
<tr><th></th><th>SeamlessM4T</th><th>SONAR</th><th>SALMONN</th></tr>
<tr><td>Common Observations Across Modalities</td><td colspan="3">• comparisons with same-language pairs (diagonal of heatmaps) hold the highest similarities<br>• final cross-lingual similarity is proportional to the resource level of language pairs (exception: SALMONN, similarities influenced by Whisper not recognizing unseen languages)<br>→ Comparisons differing in only one attribute (modality or language) hold higher similarities, as they have more common features aside from same semantic meaning.<br>→ The quantity and the quality of the training data proportionally impacts the cross-modal similarity.</td></tr>
<tr><td>Common Observations Within Modalities</td><td colspan="3">• high initial intra-text similarity, low initial intra-speech similarity<br>• cross-lingual similarity increases with the depth of the encoder/decoder<br>• intra-speech similarity < intra-text similarity in almost all layers (exception: SALMONN)<br>• final cross-lingual similarity is proportional to the resource level of language pairs<br>→ As speech is more varying in language features and length than text, the cross-lingual similarities within the speech modality are lower than those of text.</td></tr>
<tr><td rowspan="2">Final Cross-Lingual Similarity (intra-speech/intra-text)</td><td>(0.885/0.905)</td><td>(0.901/0.916)</td><td>(0.856/0.892)</td></tr>
<tr><td colspan="3">→ The language gap is more closed in a cross-lingual intra-text setting, due to the stability of normalized text inputs.</td></tr>
</table>

| | SeamlessM4T | SONAR | SALMONN |
|---|---|---|---|
| Impact of ... | Length Adaptor:<br>• increases cross-lingual intra-speech similarity more than cross-modal analysis<br>→ The downsizing of speech representations makes them more comparable within the same modality. | Pooling:<br>• mean pooling does not increase cross-lingual intra-text similarity<br>• learning pooling increases cross-lingual intra-speech similarity with a similar increase as in the cross-modal analysis<br>→ The cross-modal similarity is highly influenced by the speech learning pooling.<br>→ Text representations reach language-agnostic embeddings with the encoder (see SONAR t-SNE results)<br>→ The speech encoder representations need learning pooling and MSE loss to produce language-agnostic embeddings. | Window-Level Q-Former<br>• marginal similarity increase, same as cross-modal analysis<br>→ The Q-Former is used as a connection module between the encoder speech representations and the text-based LLM. |

**Table 5.2.: Summary of the Cross-Lingual Analysis Results.**

## 5.3. Representation Visualization Results

In the following section, the results of the t-SNE analysis are presented and discussed. For capacity and redundancy reasons, we only present the t-SNE results that have a correlation to our other analysis results or show a significant change in the distribution of the representations.

As mentioned in Section 3.3, we use t-SNE to visualize the distributions of multimodal hidden representations. The results are shown in Appendix A.4. The common ground that is shared across all three MLMs is that throughout all model architecture layers (except for the last few layers of the SONAR decoder), a clear separation between the speech and text representations is visible, meaning all models are not capable of completely closing the modality gap, since modality features are evident regardless of the modality alignment strategies that have been used. This observation matches the cross-modal results seen in Figure 5.1, as the modality gap is never fully closed for all models. Additionally, another view of the cross-modal and cross-lingual similarity results of Sections 5.1 and 5.2 is gained though the visualization of the representation distribution, giving us a further explanation on the course of the similarity.

**SeamlessM4T** In the encoder input space of SeamlessM4T (see (1) Figure A.32), we can observe that the text embeddings are already clustered into languages, with a minimal number of text embeddings that have not been correctly aligned with their language cluster. These clusters are mostly distinct, except for clusters of linguistically similar languages such as Bulgarian-Russian, Mandarin-Cantonese and Estonian-Finnish. In contrast, the speech embeddings in the encoder input space are not fully separated into languages. Instead, they are mostly aggregated into one spot, since it is not easy to define the language of the speech input in comparison to text inputs. While the clusters of text representations remain the mostly same after the first encoder layer (see (2) Figure A.32), the aggregation of speech embeddings is however separated into several smaller speech representation clusters for each language, meaning the speech encoder is able to capture language features as early as in the first encoder layer.

This separation of speech representations continue in the following encoder layers, first forming multiple bigger clusters of each language near the modality separation line in the fourth encoder layer (see (1) in Figure A.33) to becoming distinct speech representation clusters for each language in the 14th encoder layer (see (1) Figure A.34), just like the text representations were clustered in the encoder input space.

After the speech and text representations are clustered into languages, the distribution of both modalities follow the same path: linguistically similar clusters, such as French-Italian and Dutch-German, first begin to join each other until the representations are evenly spread out into smaller bundles across all languages, with a few independent clusters. For text representations, this final distribution is already reached in the 14th encoder layer (see (1) in Figure A.34), since the text representations start as languages clusters in the encoder input space and the merging starts as early as in the fourth layer (see (1) Figure A.33). These smaller bundles of text representations across different languages are the representations of the same semantic meaning (see (1) in Figure A.37), meaning that SeamlessM4T is capable of extracting the shared semantic features of text representations with the depth of the encoder. The merging of representations also correlates with the increase in the cross-modal and cross-lingual similarities from the fourth to the 14th encoder layer, as seen in Figures 5.1 and 5.5.

For speech representations, the merging starts in the 18th encoder layer (see (2) in Figure A.34) after the text representations have been fully separated into semantic meanings, and is also separated into small semantic bundles only after the last encoder layer (see (2) in Figure A.35 and (1) in Figure A.37). Same as text, the increase in similarity within the speech modality between the 18th and last encoder layer is also visible in the SeamlessM4T graphs in Figure 5.1 and 5.5. While the text representations are more evenly spread throughout the t-SNE map with relatively constant distances between the small semantic bundles, some speech representations have not been assigned to its bundle, resulting in a large aggregation with no structure across several languages in the last encoder layer, meaning that the Seamless encoder can not effectively capture the shared semantic meaning for certain speech inputs. This observation is also reflected in the averaged cross-lingual graph in Figure 5.5, as the intra-speech similarity in the last encoder layer is smaller than the intra-text similarity.

The speech clusters of Japanese and the two zero-shot languages Sindhi and Shona with the Cantonese, Mandarin and Sindhi text clusters are very noticeable in the last encoder layer, as the representations have not been aligned with their corresponding semantic bundles, meaning that the encoder is not fully capable of extracting the semantic meaning for these languages. As a result, complete rows and columns of lower cross-lingual similarities in the analysis results of Section 5.2 can be observed if the language comparison pair includes one of the aforementioned languages (see (1) in Figures A.2, A.5 and Figure A.9).

The length adaptor does not change much in the distribution of the speech representations, only slightly separating the large aggregation of speech representations into more distinct clusters of low resource languages or languages with a unique script, such as Mandarin, Korean and Persian, and also creating a new Amharic speech cluster (see (1) in Figure A.36). Thus, for these highlighted languages in the t-SNE map (2) of Figure A.36, the length adaptor of SeamlessM4T emphasizes language-specific features rather than the semantic features, resulting in lower similarity scores for these languages, as shown in heatmap (2) of Figure A.3.

**SONAR** While the distribution of the text representations of SONAR in the encoder input space and the first encoder layer is very similar to those of SeamlessM4T, the speech representations are mostly clustered into several smaller clusters for each language, as seen in the two t-SNE maps in Figure A.38. These smaller speech clusters are however not distributed across the t-SNE map without any structure like SeamlessM4T, as the clusters of one language are still closer to one other than to other language clusters, meaning that SONAR has a better understanding on how to extract the language of speech embeddings, even though SeamlessM4T and SONAR are initialized on the same models.

Nevertheless, the distribution of the SONAR text and speech representations follows the same path as SeamlessM4T. For text representations the merging begins in the sixth encoder layer (see (1) in Figure A.39) with linguistically similar languages, and is fully homogeneous in the 22nd layer (see (2) in Figure A.40). This also matches with the cross-lingual similarity increase between the sixth and 22nd encoder layers in Figure 5.5. In contrast to the distribution of the SeamlessM4T text representations, SONAR is fully capable of extracting the shared semantic meaning of all text inputs, as there are no text language clusters left in the t-SNE map of the 22nd encoder layer.

Unlike SeamlessM4T, the speech representations in SONAR are never divided into one distinct cluster for each language, instead they remain in multiple clusters for each language, as seen in the tenth encoder layer in Figure A.39, before linguistically similar languages start to join one another in the 14th layer (see (1) in Figure A.40). Several speech language clusters form bigger clusters with the depth of the encoder, just like SeamlessM4T, but they are not separated into the semantic meaning with only the encoder, as seen in the last encoder layer in Figure A.41. Regardless of the final encoder distribution of the speech representations, since similar languages are clustered together, an increase of the cross-lingual similarity within the speech modality from the 14th to the last encoder is shown in the cross-lingual SONAR analysis in Figure 5.5.

The t-SNE results of SONAR after the pooling show us that SONAR reaches its goal of producing language-agnostic embedding to a high extent (see (2) in Figure A.41), as representations, regardless of modality and language, are bundled together based on the semantic meaning (see (2) in Figure A.42). Since SONAR is the only model out of the three analyzed MLMs that reaches this distribution by producing representations independent from language and modality, SONAR also reaches the highest cross-modal and cross-lingual similarity scores, as seen in Figures 5.1 and 5.5.

The increases caused by the pooling in Figure 5.5 also correlates with our findings in this section. Since all text representations are already separated into semantic meaning in the last encoder layer (see (1) in Figure A.42), resembling the final embedding distribution, no increase in similarity is shown after the mean pooling in the cross-lingual intra-text analysis. Therefore, mean pooling is mainly used to transform the text representations to match the embedding size. In contrast, the learning pooling with MSE loss in the embedding space separates the speech clusters in the last encoder layer into semantic meaning, increasing the cross-lingual intra-speech similarity by emphasizing the semantic features in the speech representations.

**SALMONN** Same as the previous two models, all text embeddings are clustered into languages (see Figure A.43), since it is more easier to predict the language the input data is based on with text inputs than with speech inputs. Text clusters with similar linguistic features are either completely overlapping each other like Cantonese and Mandarin, or are connected like Bulgarian-Russian and English-French/Italian. Since embeddings of similar languages may have shared word tokens, the cross-lingual similarities of these languages are higher, supporting our finding in the cross-lingual similarity analysis within the text modality in Figure A.29.

Different to the input space t-SNE mappings of the previous two models, the distribution of the speech encoder outputs before the Q-Former build multiple smaller clusters for each language similar to SONAR, but are very randomly distributed across the t-SNE map. This observation is caused by the two encoders of SALMONN. While the Whisper encoder captures language features and produces similar speech representations for inputs of the same language, forming the language clusters in (1) of Figure A.43, the BEATs encoder outputs magnifies distortion of the speech representations, scattering the language clusters across the t-SNE map. The general distribution of the speech representations remains the same after the Q-Former, only being more aggregated than before (see (2) in Figure A.43). Since the Q-Former is used to downsize the varying speech sequence lengths to text with maintaining all important features, it does not greatly change the distribution and similarity of the representations, as seen previously in Figures 5.1 and 5.5.

The SALMONN speech and text representations after and including the second decoder layer follow a different path than SeamlessM4T and SONAR. The speech representations of SONAR never form distinct language clusters as the previous two models, instead the speech aggregation of multiple smaller clusters for each layer become denser with the depth of the decoder. In the second and the fourth layers, the smaller language clusters are still differentiable from others (see Figure A.44), however the speech representations begin to merge after the fourth layer until the distribution of the 12th decoder layer is reached (see (1) in Figure A.45). This is then maintained until the more homogeneous merge of speech representations is achieved in the last decoder layer (see (2) in Figure A.46). Since t-SNE maps similar representations with more closer to one another, this progression of the speech distribution resembles the course of the SALMONN cross-modal and cross-lingual similarity results (see Figures 5.1 and 5.5), with an increase in similarity from the fourth to the 12th decoder layer, maintaining the similarity of the 12th layer. Additionally, as the final distribution of speech representations lack of separation into languages or semantic meaning correlating to higher cross-modal and cross-lingual similarities, this observation also aligns with the decrease in similarity in the last decoder layer of all SALMONN analysis.

However, languages such as Amharic, Georgian, Khmer and Shona are more noticeable throughout the decoder layers. For example, Shona is not fully integrated with the rest of the speech aggregation in the last decoder layer. We assume that the reason behind this lies in the Whisper encoder not fully supporting these low resource languages, resulting in less accurate speech representations with a low similarity with other representations of other languages. Additionally, the languages of the less integrated speech clusters match the languages in cross-lingual analysis results of the last decoder layer in Figure A.26, with which all cross-lingual comparisons hold the lowest similarity.

The distribution of the SALMONN text representations undergoes the most drastic transformation from the decoder input space to the second decoder layer, since the distinct language clusters form a chain of text representations throughout the t-SNE map (see (1) in Figure A.44). These is no structure noticeable within the chain of text representations, as the representations of each language are evenly spread out across the continuum of the chain. We assume this is the reason behind the deep drop in the cross-lingual intra-text SALMONN analysis results, as shown in Figure 5.5, and therefore also in the cross-modal analysis result in Figure 5.1. Since the text representations are not gathered based on language nor semantic meaning, which are correlated to a high cross-lingual similarity as seen in the previous models, the cross-lingual similarity decreases after the second decoder layer.

However not all languages are integrated into the text representation chains, as seen at the far left end of the chain in Figure A.44. Since Vicuna has not been sufficiently trained on these languages (Amharic, Armenian, Cantonese, Chinese Mandarin, Georgian, Greek, Khmer, Thai and Tamil), their representations as a result do not accurately in capture linguistic and semantic features. The same observation is also evident in the cross-lingual intra-text SALMONN analysis of the first decoder layer, since comparisons including one of these aforementioned languages are noticeably lower than other comparisons (see (2) in Figure A.29).

After the fourth decoder layer, the chain of text representations divides into smaller chains for each language, until distinct, isolated clusters for each language, resembling distribution in the decoder input space, is reached in the final decoder layer (see (2) in Figure A.46). The SALMONN decoder is therefore able to recover from the low similarity in the earlier layers of the decoder, with accurately capturing the languages features of the text inputs. The clusters containing the similar languages Cantonese and Chinese Mandarin is an exception, as they are completely overlapping each other, which explains the high cross-lingual intra-text similarity of the Cantonese-Mandarin language pair shown in heatmap (2) of Figure A.30.

Since the both speech and text representations of SALMONN are not separated into their shared semantic meaning across languages, indicating a higher cross-modal and cross-lingual similarity as seen in the previous models, the SALMONN decoder reaches a lower final similarity in both cross-modal and cross-lingual similarity analysis results, as shown in Figures 5.1 and 5.5.

| | SeamlessM4T | SONAR | SALMONN |
|---|---|---|---|
| Common Observation Across Models | • separation into modalities visible throughout all layers (exception: SONAR)<br>→ Modality-specific features are evident in the representations of all layers, preventing the closure of the modality gap. | | |
| Distribution of Speech Representations | • embeddings are collected into one aggregation across languages<br>• distinct language clusters reached in the 14th encoder layer<br>• 18th to the last encoder layer: merging of clusters, beginning with linguistically similar languages<br>• separation into semantic meaning partly reached, aggregation across languages still visible in the last encoder layer | • embeddings in multiple clusters for each language<br>• one distinct cluster for each language is never formed<br>• from the 14th encoder layer: merging of clusters, beginning with linguistically similar languages<br>• separation into semantic meaning not reached with the encoder, aggregation across similar languages still visible in the last encoder layer | • embeddings in multiple clusters for each language, scattered across the t-SNE map<br>• one distinct cluster for each language is never formed<br>• forms homogeneous aggregation with the depth of the decoder<br>→ The lack of structure and separation aligns with the low cross-modal and cross-lingual similarities. |
| | → The merging and separation into semantic meaning correlates with the gradual increase in the cross-modal and cross-lingual similarity analysis results. | | |
| Distribution of Text Representations | • embeddings are clustered into languages<br>• 4th to 14th encoder layer: merging of clusters, beginning with linguistically similar languages<br>• separation into semantic meaning reached in the last encoder layer, with a few independent language clusters as exceptions | • embeddings are clustered into languages<br>• 6th to the 22th encoder layer: merging of clusters, beginning with linguistically similar languages<br>• separation into semantic meaning fully reached in the last encoder layer | • embeddings are clustered into languages with linguistically similar languages overlapping each other<br>• forms a chain in the second decoder layer<br>• 12th to the last decoder layer: chain is separated into languages until distinct language clusters are formed<br>→ It is assumed that the chain of text representations correlates with the drop in the cross-modal and cross-lingual intra text similarity, as representations lack of shared features.<br>→ The lack of structure and separation aligns with the low cross-modal and cross-lingual similarities. |
| | → The merging and separation into semantic meaning correlates with the gradual increase in cross-modal and cross-lingual similarity analysis results. | | |

| | SeamlessM4T | SONAR | SALMONN |
|---|---|---|---|
| Impact of ... | Length Adaptor:<br>• only slightly separates speech aggregation into language clusters<br>→ The length adaptor does not contribute to capturing shared semantic meaning, as it only downsizes speech representations with preserving features. | Pooling:<br>• the distribution of text representations remain the same<br>• aligns speech representations to the text representations clustered separated into semantic meaning<br>→ Learning pooling and MSE loss in the embedding space emphasizes semantic features of the speech representations, increasing both cross-modal and cross-lingual similarities. | Window-Level Q-Former<br>• no noticeable difference, only brings scattered speech clustered together<br>→ The Q-Former does not add to both cross-modal and cross-lingual similarities, as it is used as a connection module between encoder speech outputs and the text-based LLM. |

**Table 5.3.: Summary of the t-SNE Results.**

# 6. Conclusion

## 6.1. Answers to Research Questions

In this work, we analyzed the cross-modal and cross-lingual similarities between representations of MLMs across a set of languages, comparing them within a shared semantic space to better understand how MLMs handle their multimodal input. With the results of our analysis in Section 5, we can answer our research questions of Section 1.2 that built the basis of this work. In summary, the similarity between multimodal representations is influenced by a wide range of factors, including the model architecture, the task it is designed to perform, and the availability of language data.

**Research Question 1: How does the similarity of representations change with the depth of the model's layers?** First, we observed that in both cross-modal and cross-lingual settings, the SVCCA similarities increase with the depth of the model architecture. Thus, each model progressively captures shared features through each hidden layer, producing modality- and language-independent representations. As both features are present in the representations of almost encoder or decoder layers, the modality and language gap is never closed, with the highest cross-modal and cross-lingual still below 0.95.

**Research Question 2: How does the similarity of representations change with varying language resource levels?** Language resource levels also affect the similarity between multimodal representations. Since high-resource languages benefit from a large amount of high-quality training data, cross-modal and cross-lingual similarities are significantly higher for these languages. Therefore, it is important for MLMs to be sufficiently trained on languages to capture shared features between multimodal representations. However, it has also been observed that even for low-resource languages, which may have been trained on minimal data or are zero-shot languages, the similarities are still higher than a random baseline, meaning that the models are able to generalize beyond the data they have seen during training, emphasizing their robustness.

**Research Question 3: How do the similarities of representations differ in a cross-modal and cross-lingual setting?** Regarding the differences in cross-modal and cross-lingual settings, the model demonstrates varying levels of similarity depending on how many attributes — modality or language — differ. Higher similarities are reached when only one attribute differs in the comparison of the representations. This is because, in addition to the shared semantic meaning, another consistent attribute across the comparisons allows the representations to have more shared features, resulting in higher similarities. Therefore, when both modality and language differ, the similarities decrease, even if the semantic meaning between the compared representations is preserved. This highlights the limitations in the alignment capabilities of current MLMs.

**Research Question 4: How does the architecture of the model affect the similarity of representations?** The tasks on which the MLMs have been trained, as well as the components of their architecture, also greatly affect the similarity between speech and text representations. For models developed for multimodal translation or for achieving multimodal language-agnostic embeddings, such as SeamlessM4T and SONAR, high similarity between multimodal encoder representations of the same semantic meaning is crucial for their performance. These models therefore reach cross-modal similarities of over 0.9, while the decoder of instruction-following models like SALMONN, which do not prioritize high multimodal similarity, lack such high levels of similarity.

Higher similarities are also made possible by various speech sequence shortening methods, as MLMs generally have a shared embedding space where speech representations of varying lengths must be adapted. The length adaptor and Q-Former of SeamlessM4T and SALMONN marginally increase both cross-modal and cross-lingual similarities, as their primary goal is to downsize speech representations of varying lengths, rather than to increase similarity by extracting more shared features from the multimodal comparison. In contrast, the leaning pooling method of SONAR in combination with the MSE loss in embedding space increases both cross-modal and cross-lingual similarities, as they smooth out language- and modality-specific features, making the representations more abstract and similar through their shared semantic meaning.

## 6.2. Limitations of this Work

Our work is limited to the data used for the analysis, since the extracted representations only come from one multimodal and multilingual n-way parallel dataset: FLEURS. Although the FLEURS dataset covers different domains due to it being based on FLORES (see Section 4.2), it is often the case that one dataset does not fully capture the diversity of languages, dialects and speech patterns, due to all sentences being sourced from Wikimedia sources and the low number of speakers for each language (Goyal et al., 2022).

Additionally, this work focuses on the analyzing MLMs and their multimodal representation similarities, rather than evaluating the model architecture and its performance. While our findings help in understanding on how MLMs handle with multimodality, it does not make any implications on model improvements, as this would be out of the scope of this work. However, future research building upon the findings of this work could dive into enhancing model architectures or training strategies to further improve cross-modal alignment.

# Bibliography

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2016). *Neural Machine Translation by Jointly Learning to Align and Translate.* arXiv: `1409.0473 [cs.CL]`. URL: `https://arxiv.org/abs/1409.0473` (cit. on p. 3).

Brown, Tom et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems.* Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf` (cit. on pp. 1, 4).

Chen, Sanyuan et al. (2022). *BEATs: Audio Pre-Training with Acoustic Tokenizers.* arXiv: `2212.09058 [eess.AS]`. URL: `https://arxiv.org/abs/2212.09058` (cit. on p. 13).

Chung, Yu-An et al. (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)*, pp. 244–250. DOI: `10.1109/ASRU51503.2021.9688253` (cit. on p. 12).

Communication, Seamless, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, et al. (2023). *SeamlessM4T: Massively Multilingual & Multimodal Machine Translation.* arXiv: `2308.11596 [cs.CL]`. URL: `https://arxiv.org/abs/2308.11596` (cit. on pp. 15, 31).

Communication, Seamless, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, et al. (2023). *Seamless: Multilingual Expressive and Streaming Speech Translation.* arXiv: `2312.05187 [cs.CL]`. URL: `https://arxiv.org/abs/2312.05187` (cit. on pp. 11, 12, 31).

Conneau, Alexis et al. (2023). "FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech". In: *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805. DOI: `10.1109/SLT54892.2023.10023141` (cit. on p. 15).

Devlin, Jacob et al. (June 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).* Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: `10.18653/v1/N19-1423`. URL: `https://aclanthology.org/N19-1423` (cit. on p. 4).

Duquenne, Paul-Ambroise, Holger Schwenk, and Benoıt Sagot (Aug. 2023). *SONAR: Sentence-Level Multimodal and Language-Agnostic Representations.* DOI: `10.48550/arXiv.2308.11466`. arXiv: `2308.11466 [cs.CL]` (cit. on pp. 12, 13).

Girdhar, Rohit et al. (June 2023). "ImageBind: One Embedding Space To Bind Them All". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190 (cit. on p. 5).

Goyal, Naman et al. (May 2022). "The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation". In: *Transactions of the Association for Computational Linguistics* 10, pp. 522–538. ISSN: 2307-387X. DOI: `10.1162/tacl_a_00474`. eprint: `https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl\_a\_00474/2020699/tacl\_a\_00474.pdf`. URL: `https://doi.org/10.1162/tacl%5C_a%5C_00474` (cit. on pp. 15, 46).

Han, Jiaming et al. (June 2024). "OneLLM: One Framework to Align All Modalities with Language". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26584–26595 (cit. on p. 5).

Hinton, Geoffrey E and Sam Roweis (2002). "Stochastic Neighbor Embedding". In: *Advances in Neural Information Processing Systems.* Ed. by S. Becker, S. Thrun, and K. Obermayer. Vol. 15. MIT Press. URL: `https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf` (cit. on p. 9).

Hsu, Wei-Ning et al. (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. DOI: `10.1109/TASLP.2021.3122291` (cit. on p. 6).

Hu, Edward J. et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models.* arXiv: `2106.09685 [cs.CL]`. URL: `https://arxiv.org/abs/2106.09685` (cit. on p. 13).

Kaddour, Jean et al. (2023). *Challenges and Applications of Large Language Models.* arXiv: `2307.10169 [cs.CL]`. URL: `https://arxiv.org/abs/2307.10169` (cit. on p. 1).

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605. URL: `http://jmlr.org/papers/v9/vandermaaten08a.html` (cit. on pp. 5, 9).

Nguyen, Tu Anh et al. (2024). *SpiRit-LM: Interleaved Spoken and Written Language Model.* arXiv: `2402.05755 [cs.CL]`. URL: `https://arxiv.org/abs/2402.05755` (cit. on p. 5).

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (2019). *Representation Learning with Contrastive Predictive Coding.* arXiv: `1807.03748 [cs.LG]`. URL: `https://arxiv.org/abs/1807.03748` (cit. on p. 5).

OpenAI et al. (2024). *GPT-4 Technical Report.* arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774 (cit. on p. 4).

Radford, Alec et al. (23–29 Jul 2023). “Robust Speech Recognition via Large-Scale Weak Supervision”. In: *Proceedings of the 40th International Conference on Machine Learning.* Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html (cit. on pp. 13, 26).

Raghu, Maithra et al. (2017). “SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability”. In: *Advances in Neural Information Processing Systems.* Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. (cit. on pp. 5, 8, 17).

Seyssel, Maureen de et al. (Sept. 2022). “Probing phoneme, language and speaker information in unsupervised speech representations”. In: *Interspeech 2022.* interspeech$_2$022. ISCA. DOI: 10.21437/interspeech.2022-373. URL: http://dx.doi.org/10.21437/Interspeech.2022-373 (cit. on p. 5).

Sicherman, Amitay and Yossi Adi (2023). “Analysing Discrete Self Supervised Speech Representation For Spoken Language Modeling”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10097097 (cit. on p. 6).

Sun, Haoran et al. (2023). *Towards a Deep Understanding of Multilingual End-to-End Speech Translation.* arXiv: 2310.20456 [cs.CL]. URL: https://arxiv.org/abs/2310.20456 (cit. on p. 5).

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). “Sequence to Sequence Learning with Neural Networks”. In: *Advances in Neural Information Processing Systems.* Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf (cit. on p. 3).

Tang, Changli et al. (2024). *SALMONN: Towards Generic Hearing Abilities for Large Language Models.* arXiv: 2310.13289 [cs.SD]. URL: https://arxiv.org/abs/2310.13289 (cit. on pp. 5, 13, 14, 36).

Team, Gemini et al. (2024). *Gemini: A Family of Highly Capable Multimodal Models.* arXiv: 2312.11805 [cs.CL]. URL: https://arxiv.org/abs/2312.11805 (cit. on p. 5).

Team, NLLB et al. (2022). *No Language Left Behind: Scaling Human-Centered Machine Translation.* arXiv: 2207.04672 [cs.CL]. URL: https://arxiv.org/abs/2207.04672 (cit. on p. 12).

Tharwat, Alaa et al. (2017). “Linear discriminant analysis: A detailed tutorial”. In: *AI communications* 30.2, pp. 169–190 (cit. on p. 5).

Touvron, Hugo et al. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models.* arXiv: 2307.09288 [cs.CL]. URL: https://arxiv.org/abs/2307.09288 (cit. on pp. 1, 4, 13, 25, 33).

Vaswani, Ashish et al. (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems.* Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. (cit. on p. 3).

Wang, Gary et al. (2023). “Understanding Shared Speech-Text Representations”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095508 (cit. on p. 5).

Wang, Xiao et al. (2023). “Large-scale multi-modal pre-trained models: A comprehensive survey”. In: *Machine Intelligence Research* 20.4, pp. 447–482 (cit. on p. 4).

Yin, Shukang et al. (2024). *A Survey on Multimodal Large Language Models.* arXiv: 2306.13549 [cs.CV]. URL: https://arxiv.org/abs/2306.13549 (cit. on p. 4).

Zhang, Duzhen et al. (2024). *MM-LLMs: Recent Advances in MultiModal Large Language Models.* arXiv: 2401.13601 [cs.CL]. URL: https://arxiv.org/abs/2401.13601 (cit. on p. 4).

Zhao, Jinming et al. (2022). *M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation.* arXiv: 2207.00952 [cs.CL]. URL: https://arxiv.org/abs/2207.00952 (cit. on p. 12).

Zhao, Wayne Xin et al. (2023). *A Survey of Large Language Models.* arXiv: 2303.18223 [cs.CL]. URL: https://arxiv.org/abs/2303.18223 (cit. on p. 1).

Zheng, Lianmin et al. (2023). “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena”. In: *Advances in Neural Information Processing Systems.* Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., pp. 46595–46623. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf (cit. on p. 13).

# A. Appendix

## A.1. Target Dimensions for Cross-Modal Analysis

### A.1.1. SeamlessM4T

| | amh | arb | bul | cat | cmn | deu | ell | eng | est | fin | fra | hin | hye | ind | ita |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input | 87 | 90 | 94 | 87 | 85 | 94 | 96 | 93 | 93 | 86 | 92 | 92 | 92 | 91 | 87 |
| 1 | 143 | 141 | 141 | 142 | 138 | 144 | 138 | 140 | 145 | 133 | 139 | 133 | 142 | 135 | 135 |
| 2 | 142 | 140 | 141 | 144 | 139 | 141 | 139 | 141 | 144 | 135 | 142 | 134 | 143 | 136 | 137 |
| 4 | 157 | 155 | 157 | 159 | 156 | 156 | 155 | 156 | 158 | 152 | 157 | 151 | 158 | 155 | 153 |
| 6 | 162 | 161 | 166 | 165 | 158 | 165 | 162 | 164 | 166 | 158 | 165 | 156 | 160 | 160 | 163 |
| 8 | 157 | 155 | 159 | 156 | 154 | 155 | 155 | 157 | 157 | 150 | 155 | 154 | 156 | 155 | 156 |
| 10 | 162 | 163 | 168 | 166 | 157 | 165 | 164 | 168 | 166 | 160 | 166 | 161 | 164 | 162 | 164 |
| 12 | 174 | 175 | 179 | 177 | 173 | 175 | 178 | 179 | 179 | 173 | 177 | 175 | 176 | 176 | 176 |
| 14 | 167 | 166 | 168 | 165 | 162 | 164 | 168 | 169 | 168 | 163 | 166 | 165 | 168 | 167 | 164 |
| 16 | 167 | 161 | 163 | 159 | 160 | 159 | 162 | 165 | 162 | 159 | 161 | 162 | 164 | 163 | 158 |
| 18 | 171 | 167 | 170 | 166 | 166 | 166 | 169 | 172 | 169 | 166 | 168 | 169 | 170 | 168 | 166 |
| 20 | 178 | 176 | 179 | 176 | 175 | 174 | 178 | 180 | 179 | 175 | 176 | 177 | 179 | 178 | 176 |
| 22 | 182 | 182 | 184 | 182 | 180 | 179 | 184 | 185 | 183 | 179 | 181 | 182 | 184 | 183 | 182 |
| 23 | 184 | 184 | 186 | 185 | 182 | 182 | 186 | 187 | 186 | 183 | 184 | 185 | 186 | 186 | 185 |
| 24 | 186 | 187 | 189 | 187 | 183 | 184 | 189 | 189 | 188 | 185 | 187 | 187 | 188 | 188 | 187 |
| Adaptor | 192 | 193 | 195 | 193 | 189 | 191 | 194 | 195 | 193 | 191 | 193 | 193 | 194 | 194 | 193 |

| | jpn | kat | khm | kor | lit | mar | nld | pes | rus | sna | snd | tam | tha | tur | yue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input | 90 | 84 | 89 | 94 | 91 | 90 | 90 | 88 | 91 | 91 | 93 | 93 | 91 | 90 | 94 |
| 1 | 139 | 140 | 136 | 138 | 144 | 134 | 151 | 140 | 141 | 139 | 137 | 131 | 146 | 139 | 143 |
| 2 | 138 | 141 | 138 | 136 | 144 | 134 | 149 | 138 | 140 | 140 | 139 | 132 | 144 | 140 | 141 |
| 4 | 155 | 157 | 153 | 151 | 160 | 152 | 161 | 155 | 155 | 158 | 157 | 150 | 156 | 158 | 154 |
| 6 | 159 | 162 | 159 | 157 | 164 | 156 | 169 | 161 | 164 | 160 | 162 | 153 | 161 | 162 | 160 |
| 8 | 153 | 157 | 156 | 152 | 158 | 156 | 160 | 155 | 157 | 157 | 161 | 152 | 156 | 155 | 160 |
| 10 | 161 | 164 | 159 | 160 | 166 | 162 | 170 | 161 | 166 | 161 | 165 | 156 | 160 | 162 | 158 |
| 12 | 173 | 177 | 172 | 173 | 177 | 173 | 179 | 174 | 179 | 173 | 176 | 170 | 174 | 173 | 176 |
| 14 | 165 | 168 | 163 | 164 | 167 | 166 | 168 | 166 | 167 | 163 | 165 | 163 | 163 | 166 | 165 |
| 16 | 161 | 164 | 163 | 162 | 162 | 164 | 162 | 164 | 161 | 165 | 165 | 161 | 161 | 163 | 163 |
| 18 | 167 | 169 | 167 | 168 | 169 | 170 | 169 | 169 | 168 | 170 | 171 | 167 | 166 | 169 | 170 |
| 20 | 175 | 178 | 175 | 176 | 178 | 178 | 177 | 177 | 178 | 174 | 177 | 176 | 176 | 177 | 178 |
| 22 | 180 | 184 | 179 | 181 | 183 | 183 | 182 | 182 | 183 | 176 | 181 | 181 | 181 | 182 | 183 |
| 23 | 182 | 186 | 181 | 184 | 185 | 185 | 185 | 184 | 186 | 178 | 183 | 183 | 183 | 184 | 185 |
| 24 | 186 | 188 | 184 | 186 | 188 | 187 | 187 | 187 | 188 | 180 | 185 | 185 | 185 | 187 | 187 |
| Adaptor | 190 | 194 | 189 | 191 | 194 | 193 | 193 | 192 | 194 | 188 | 192 | 192 | 192 | 192 | 192 |

**Table A.1.: SeamlessM4T Target Dimensions for Speech Representation Sets.** For each language (header) and encoder layer (left column). 'Input' refers to the encoder input embeddings of SeamlessM4T and 'Adaptor' refers to the speech representations after the length adaptor.

| | amh | arb | bul | cat | cmn | deu | ell | eng | est | fin | fra | hin | hye | ind | ita |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input | 191 | 194 | 193 | 194 | 184 | 196 | 189 | 197 | 198 | 197 | 194 | 191 | 188 | 196 | 196 |
| 1 | 189 | 189 | 190 | 192 | 180 | 194 | 186 | 194 | 194 | 194 | 192 | 188 | 185 | 192 | 194 |
| 2 | 186 | 185 | 187 | 189 | 177 | 190 | 183 | 191 | 191 | 191 | 189 | 185 | 183 | 189 | 190 |
| 4 | 185 | 184 | 185 | 186 | 175 | 188 | 182 | 187 | 188 | 188 | 186 | 184 | 183 | 186 | 187 |
| 6 | 188 | 188 | 188 | 189 | 177 | 190 | 186 | 190 | 190 | 189 | 189 | 188 | 187 | 188 | 190 |
| 8 | 190 | 190 | 190 | 191 | 181 | 192 | 189 | 192 | 191 | 190 | 191 | 190 | 189 | 190 | 192 |
| 10 | 191 | 191 | 191 | 191 | 184 | 192 | 190 | 193 | 192 | 191 | 192 | 192 | 191 | 191 | 193 |
| 12 | 192 | 192 | 192 | 192 | 185 | 193 | 191 | 193 | 193 | 191 | 193 | 193 | 191 | 192 | 193 |
| 14 | 192 | 191 | 192 | 192 | 185 | 193 | 191 | 193 | 193 | 191 | 193 | 194 | 191 | 192 | 193 |
| 16 | 192 | 191 | 193 | 193 | 185 | 193 | 191 | 193 | 193 | 191 | 193 | 194 | 192 | 193 | 194 |
| 18 | 192 | 191 | 193 | 193 | 185 | 193 | 191 | 194 | 193 | 191 | 193 | 194 | 192 | 193 | 194 |
| 20 | 192 | 191 | 193 | 193 | 185 | 193 | 191 | 194 | 194 | 191 | 194 | 195 | 192 | 193 | 194 |
| 22 | 193 | 192 | 194 | 194 | 186 | 194 | 192 | 195 | 195 | 192 | 195 | 195 | 193 | 194 | 195 |
| 23 | 194 | 192 | 195 | 195 | 187 | 195 | 193 | 195 | 196 | 193 | 195 | 196 | 194 | 195 | 196 |
| 24 | 197 | 197 | 197 | 197 | 191 | 197 | 197 | 197 | 198 | 196 | 198 | 198 | 197 | 198 | 198 |

| | jpn | kat | khm | kor | lit | mar | nld | pes | rus | sna | snd | tam | tha | tur | yue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input | 192 | 193 | 190 | 192 | 197 | 194 | 196 | 193 | 196 | 197 | 191 | 192 | 191 | 198 | 186 |
| 1 | 188 | 190 | 186 | 189 | 194 | 190 | 194 | 189 | 192 | 194 | 188 | 189 | 186 | 194 | 180 |
| 2 | 186 | 187 | 184 | 187 | 191 | 187 | 191 | 187 | 189 | 190 | 185 | 187 | 184 | 191 | 178 |
| 4 | 184 | 186 | 182 | 185 | 187 | 186 | 188 | 185 | 186 | 187 | 184 | 186 | 182 | 188 | 174 |
| 6 | 186 | 188 | 185 | 187 | 188 | 189 | 190 | 187 | 189 | 188 | 187 | 189 | 185 | 189 | 173 |
| 8 | 188 | 190 | 187 | 189 | 189 | 190 | 192 | 189 | 191 | 190 | 188 | 190 | 187 | 190 | 174 |
| 10 | 190 | 191 | 189 | 190 | 190 | 191 | 193 | 191 | 193 | 191 | 190 | 191 | 189 | 191 | 175 |
| 12 | 191 | 193 | 190 | 190 | 190 | 193 | 193 | 192 | 193 | 192 | 190 | 192 | 190 | 192 | 176 |
| 14 | 191 | 193 | 190 | 190 | 191 | 193 | 193 | 192 | 194 | 193 | 190 | 193 | 190 | 192 | 176 |
| 16 | 192 | 193 | 190 | 189 | 191 | 193 | 194 | 192 | 194 | 193 | 190 | 193 | 190 | 192 | 176 |
| 18 | 192 | 193 | 190 | 189 | 190 | 193 | 193 | 192 | 194 | 193 | 189 | 193 | 190 | 192 | 177 |
| 20 | 192 | 193 | 191 | 189 | 191 | 194 | 194 | 193 | 194 | 193 | 189 | 193 | 190 | 193 | 177 |
| 22 | 193 | 194 | 192 | 190 | 192 | 195 | 195 | 194 | 195 | 194 | 190 | 194 | 191 | 194 | 178 |
| 23 | 194 | 195 | 192 | 190 | 192 | 195 | 196 | 195 | 196 | 195 | 190 | 195 | 192 | 194 | 179 |
| 24 | 197 | 197 | 196 | 196 | 196 | 198 | 198 | 197 | 198 | 197 | 196 | 197 | 196 | 197 | 185 |

**Table A.2.: SeamlessM4T Target Dimensions for Text Representation Sets.** For each language (header) and encoder layer (left column). 'Input' refers to the encoder input embeddings of SeamlessM4T.

## A.1.2. SONAR

| | amh | arb | bul | cat | cmn | deu | ell | eng | est | fin | fra | hin | hye | ind | ita |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input | 89 | 90 | 95 | 87 | 85 | 95 | 91 | 92 | 84 | 92 | 77 | 92 | 90 | 87 | 92 |
| 1 | 146 | 152 | 149 | 148 | 149 | 150 | 144 | 150 | 141 | 147 | 149 | 143 | 143 | 143 | 144 |
| 2 | 146 | 145 | 153 | 143 | 144 | 153 | 143 | 146 | 140 | 152 | 147 | 139 | 137 | 146 | 152 |
| 4 | 154 | 155 | 159 | 156 | 155 | 159 | 153 | 157 | 149 | 160 | 152 | 148 | 148 | 155 | 158 |
| 6 | 158 | 158 | 168 | 166 | 155 | 167 | 162 | 162 | 152 | 167 | 160 | 153 | 157 | 164 | 160 |
| 8 | 158 | 162 | 159 | 157 | 158 | 161 | 154 | 156 | 146 | 159 | 154 | 157 | 158 | 154 | 159 |
| 10 | 147 | 150 | 153 | 149 | 147 | 151 | 153 | 151 | 143 | 153 | 147 | 148 | 147 | 149 | 149 |
| 12 | 171 | 174 | 177 | 174 | 167 | 172 | 175 | 176 | 170 | 174 | 171 | 173 | 172 | 172 | 173 |
| 14 | 178 | 177 | 182 | 180 | 175 | 179 | 182 | 182 | 177 | 180 | 178 | 178 | 178 | 180 | 177 |
| 16 | 171 | 172 | 174 | 172 | 170 | 173 | 176 | 174 | 170 | 173 | 172 | 172 | 172 | 173 | 172 |
| 18 | 166 | 176 | 165 | 164 | 173 | 168 | 165 | 168 | 168 | 166 | 176 | 174 | 170 | 167 | 176 |
| 20 | 175 | 180 | 173 | 169 | 177 | 174 | 173 | 177 | 177 | 171 | 182 | 179 | 177 | 172 | 180 |
| 22 | 183 | 183 | 181 | 177 | 178 | 179 | 179 | 185 | 183 | 178 | 186 | 183 | 182 | 178 | 182 |
| 23 | 189 | 188 | 188 | 184 | 184 | 185 | 185 | 190 | 188 | 184 | 192 | 189 | 189 | 185 | 187 |
| 24 | 191 | 189 | 190 | 186 | 187 | 186 | 188 | 192 | 190 | 186 | 192 | 191 | 191 | 186 | 189 |
| Pooling | 201 | 198 | 202 | 202 | 200 | 202 | 203 | 202 | 201 | 201 | 202 | 199 | 201 | 202 | 200 |

| | jpn | kat | khm | kor | lit | mar | nld | pes | rus | sna | snd | tam | tha | tur | yue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input | 94 | 89 | 92 | 90 | 89 | 91 | 87 | 91 | 94 | 92 | 93 | 94 | 84 | 89 | 95 |
| 1 | 146 | 141 | 153 | 146 | 144 | 157 | 146 | 148 | 145 | 137 | 142 | 146 | 148 | 145 | 150 |
| 2 | 145 | 139 | 149 | 149 | 147 | 156 | 147 | 139 | 152 | 141 | 139 | 144 | 142 | 141 | 155 |
| 4 | 149 | 147 | 156 | 156 | 154 | 163 | 152 | 151 | 158 | 147 | 146 | 151 | 147 | 152 | 159 |
| 6 | 149 | 147 | 166 | 158 | 157 | 171 | 155 | 163 | 161 | 147 | 151 | 157 | 147 | 154 | 160 |
| 8 | 153 | 149 | 158 | 158 | 159 | 165 | 152 | 158 | 163 | 148 | 155 | 158 | 149 | 151 | 161 |
| 10 | 144 | 147 | 152 | 148 | 150 | 156 | 146 | 150 | 152 | 143 | 145 | 149 | 147 | 147 | 150 |
| 12 | 168 | 168 | 176 | 173 | 173 | 178 | 171 | 175 | 176 | 169 | 168 | 174 | 169 | 170 | 172 |
| 14 | 173 | 176 | 181 | 177 | 177 | 182 | 177 | 182 | 179 | 175 | 175 | 178 | 176 | 176 | 179 |
| 16 | 167 | 172 | 173 | 173 | 172 | 176 | 173 | 174 | 172 | 170 | 171 | 174 | 171 | 171 | 174 |
| 18 | 175 | 170 | 164 | 173 | 174 | 169 | 173 | 164 | 178 | 171 | 175 | 175 | 168 | 171 | 177 |
| 20 | 179 | 174 | 173 | 178 | 178 | 176 | 178 | 174 | 181 | 177 | 177 | 178 | 174 | 177 | 181 |
| 22 | 182 | 184 | 181 | 182 | 182 | 181 | 184 | 181 | 183 | 179 | 182 | 183 | 185 | 181 | 182 |
| 23 | 189 | 193 | 187 | 189 | 189 | 187 | 189 | 188 | 187 | 184 | 188 | 189 | 193 | 187 | 186 |
| 24 | 190 | 194 | 188 | 190 | 190 | 188 | 191 | 189 | 189 | 187 | 189 | 191 | 194 | 189 | 189 |
| Pooling | 200 | 200 | 201 | 200 | 201 | 202 | 201 | 202 | 196 | 199 | 199 | 199 | 200 | 201 | 199 |

**Table A.3.: SONAR Target Dimensions for Speech Representation Sets.** For each language (header) and encoder layer (left column). 'Input' refers to the encoder input embeddings of SONAR and 'Pooling' refers to the speech representations after learning pooling.

| | amh | arb | bul | cat | cmn | deu | ell | eng | est | fin | fra | hin | hye | ind | ita |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input | 191 | 192 | 192 | 195 | 184 | 197 | 197 | 198 | 198 | 194 | 190 | 194 | 197 | 196 | 193 |
| 1 | 189 | 189 | 190 | 193 | 181 | 195 | 196 | 196 | 196 | 193 | 187 | 191 | 195 | 195 | 189 |
| 2 | 188 | 189 | 190 | 192 | 180 | 194 | 194 | 195 | 194 | 193 | 187 | 190 | 193 | 193 | 187 |
| 4 | 188 | 188 | 189 | 191 | 180 | 191 | 192 | 193 | 192 | 190 | 188 | 189 | 191 | 191 | 187 |
| 6 | 191 | 191 | 191 | 192 | 182 | 193 | 192 | 194 | 193 | 191 | 191 | 191 | 191 | 192 | 188 |
| 8 | 193 | 192 | 193 | 193 | 186 | 194 | 193 | 194 | 193 | 193 | 194 | 193 | 192 | 193 | 191 |
| 10 | 193 | 193 | 193 | 192 | 189 | 194 | 193 | 194 | 192 | 193 | 194 | 193 | 192 | 193 | 192 |
| 12 | 194 | 193 | 193 | 193 | 190 | 194 | 194 | 194 | 193 | 193 | 194 | 194 | 193 | 193 | 192 |
| 14 | 194 | 193 | 194 | 193 | 191 | 194 | 194 | 194 | 193 | 194 | 194 | 194 | 192 | 194 | 193 |
| 16 | 194 | 193 | 194 | 193 | 191 | 194 | 193 | 194 | 193 | 194 | 194 | 194 | 192 | 194 | 193 |
| 18 | 194 | 193 | 193 | 193 | 191 | 194 | 193 | 193 | 192 | 193 | 193 | 194 | 191 | 193 | 193 |
| 20 | 191 | 191 | 191 | 190 | 188 | 191 | 190 | 191 | 190 | 190 | 191 | 192 | 190 | 190 | 191 |
| 22 | 195 | 196 | 195 | 195 | 194 | 196 | 196 | 195 | 195 | 195 | 196 | 196 | 196 | 195 | 195 |
| 23 | 198 | 196 | 197 | 196 | 196 | 199 | 199 | 197 | 197 | 198 | 197 | 198 | 197 | 198 | 198 |
| 24 | 203 | 200 | 203 | 202 | 202 | 203 | 203 | 203 | 202 | 203 | 202 | 203 | 203 | 203 | 203 |
| Pooling | 203 | 200 | 203 | 202 | 202 | 203 | 203 | 203 | 202 | 203 | 202 | 203 | 203 | 203 | 203 |

| | jpn | kat | khm | kor | lit | mar | nld | pes | rus | sna | snd | tam | tha | tur | yue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input | 193 | 188 | 197 | 190 | 195 | 197 | 194 | 195 | 191 | 197 | 191 | 193 | 186 | 199 | 185 |
| 1 | 190 | 185 | 195 | 188 | 193 | 195 | 191 | 193 | 189 | 195 | 189 | 190 | 183 | 196 | 182 |
| 2 | 189 | 185 | 194 | 188 | 192 | 194 | 190 | 192 | 189 | 194 | 189 | 189 | 184 | 195 | 182 |
| 4 | 188 | 185 | 191 | 188 | 190 | 192 | 189 | 190 | 188 | 192 | 190 | 189 | 185 | 192 | 182 |
| 6 | 189 | 187 | 192 | 191 | 192 | 193 | 191 | 191 | 190 | 193 | 191 | 191 | 187 | 193 | 184 |
| 8 | 191 | 190 | 193 | 193 | 193 | 194 | 192 | 193 | 192 | 193 | 193 | 193 | 191 | 193 | 187 |
| 10 | 192 | 191 | 193 | 193 | 193 | 193 | 193 | 193 | 193 | 193 | 193 | 194 | 192 | 193 | 189 |
| 12 | 193 | 193 | 193 | 193 | 193 | 193 | 194 | 193 | 193 | 193 | 193 | 194 | 192 | 193 | 191 |
| 14 | 193 | 193 | 193 | 194 | 194 | 194 | 194 | 193 | 193 | 194 | 194 | 194 | 193 | 193 | 192 |
| 16 | 193 | 194 | 193 | 194 | 194 | 194 | 194 | 193 | 194 | 194 | 194 | 194 | 193 | 194 | 192 |
| 18 | 193 | 193 | 193 | 193 | 194 | 193 | 193 | 193 | 193 | 193 | 194 | 194 | 193 | 193 | 192 |
| 20 | 190 | 191 | 191 | 191 | 191 | 191 | 191 | 191 | 190 | 191 | 191 | 192 | 191 | 191 | 189 |
| 22 | 195 | 195 | 195 | 196 | 196 | 196 | 195 | 196 | 196 | 195 | 196 | 196 | 195 | 195 | 194 |
| 23 | 197 | 198 | 197 | 197 | 197 | 198 | 198 | 199 | 198 | 198 | 196 | 197 | 197 | 198 | 197 |
| 24 | 202 | 202 | 203 | 202 | 202 | 203 | 203 | 204 | 201 | 202 | 202 | 201 | 203 | 203 | 202 |
| Pooling | 202 | 202 | 203 | 202 | 202 | 203 | 203 | 204 | 201 | 202 | 202 | 201 | 203 | 203 | 202 |

**Table A.4.: SONAR Target Dimensions for Text Representation Sets.** For each language (header) and encoder layer (left column). 'Input' refers to the encoder input embeddings of SONAR and 'Pooling' refers to the text representations after mean pooling.

### A.1.3. SALMONN

| | amh | arb | bul | cat | cmn | deu | ell | eng | est | fin | fra | hin | hye | ind | ita |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Encoder | 168 | 166 | 169 | 165 | 162 | 161 | 168 | 159 | 168 | 158 | 166 | 164 | 167 | 163 | 162 |
| Input | 164 | 163 | 164 | 163 | 160 | 159 | 163 | 163 | 163 | 159 | 165 | 160 | 163 | 160 | 161 |
| 1 | 164 | 163 | 164 | 163 | 160 | 159 | 163 | 163 | 163 | 159 | 165 | 161 | 163 | 161 | 161 |
| 2 | 164 | 163 | 164 | 163 | 160 | 158 | 163 | 163 | 162 | 158 | 165 | 160 | 163 | 160 | 160 |
| 4 | 169 | 169 | 170 | 168 | 166 | 164 | 169 | 169 | 168 | 164 | 170 | 165 | 169 | 166 | 165 |
| 6 | 177 | 179 | 181 | 179 | 178 | 177 | 181 | 181 | 180 | 176 | 181 | 176 | 180 | 177 | 177 |
| 8 | 185 | 186 | 187 | 186 | 186 | 184 | 188 | 188 | 187 | 184 | 188 | 184 | 187 | 185 | 185 |
| 10 | 190 | 192 | 193 | 192 | 192 | 191 | 193 | 194 | 193 | 191 | 194 | 190 | 192 | 191 | 192 |
| 12 | 194 | 195 | 197 | 196 | 196 | 195 | 197 | 197 | 196 | 195 | 197 | 194 | 196 | 195 | 196 |
| 14 | 195 | 197 | 198 | 198 | 198 | 197 | 198 | 199 | 198 | 197 | 198 | 196 | 197 | 197 | 197 |
| 16 | 196 | 198 | 199 | 198 | 198 | 197 | 199 | 199 | 199 | 198 | 199 | 197 | 198 | 198 | 198 |
| 18 | 197 | 199 | 200 | 199 | 199 | 198 | 200 | 200 | 200 | 199 | 200 | 198 | 199 | 199 | 199 |
| 20 | 198 | 200 | 201 | 200 | 200 | 199 | 201 | 201 | 200 | 200 | 201 | 199 | 200 | 200 | 200 |
| 22 | 199 | 201 | 201 | 201 | 201 | 200 | 201 | 202 | 201 | 200 | 201 | 199 | 200 | 201 | 201 |
| 24 | 199 | 201 | 202 | 201 | 201 | 200 | 201 | 202 | 201 | 201 | 202 | 200 | 201 | 201 | 201 |
| 26 | 198 | 200 | 201 | 201 | 200 | 199 | 201 | 201 | 201 | 200 | 201 | 199 | 200 | 200 | 201 |
| 28 | 196 | 198 | 199 | 200 | 199 | 197 | 199 | 200 | 199 | 199 | 199 | 198 | 198 | 199 | 199 |
| 30 | 195 | 198 | 198 | 199 | 198 | 196 | 198 | 199 | 198 | 198 | 198 | 197 | 197 | 198 | 198 |
| 31 | 194 | 197 | 198 | 198 | 197 | 195 | 197 | 198 | 197 | 197 | 197 | 196 | 196 | 197 | 197 |
| 32 | 191 | 193 | 194 | 194 | 193 | 191 | 194 | 195 | 192 | 192 | 194 | 191 | 192 | 192 | 192 |

| | jpn | kat | khm | kor | lit | mar | nld | pes | rus | sna | snd | tam | tha | tur | yue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Encoder | 160 | 166 | 160 | 162 | 160 | 165 | 167 | 160 | 166 | 159 | 167 | 162 | 166 | 167 | 167 |
| Input | 158 | 163 | 158 | 160 | 161 | 160 | 164 | 158 | 162 | 159 | 163 | 159 | 160 | 162 | 161 |
| 1 | 158 | 163 | 158 | 160 | 161 | 160 | 164 | 159 | 162 | 159 | 163 | 159 | 160 | 162 | 161 |
| 2 | 158 | 163 | 158 | 160 | 161 | 160 | 164 | 158 | 162 | 158 | 162 | 159 | 160 | 162 | 161 |
| 4 | 164 | 168 | 163 | 166 | 167 | 164 | 171 | 163 | 168 | 162 | 167 | 163 | 165 | 168 | 167 |
| 6 | 176 | 177 | 173 | 178 | 178 | 174 | 182 | 173 | 179 | 170 | 176 | 173 | 176 | 178 | 179 |
| 8 | 184 | 185 | 182 | 185 | 185 | 182 | 188 | 182 | 187 | 178 | 183 | 181 | 184 | 185 | 186 |
| 10 | 190 | 190 | 189 | 191 | 192 | 188 | 194 | 189 | 193 | 185 | 190 | 188 | 190 | 192 | 192 |
| 12 | 195 | 194 | 193 | 196 | 195 | 193 | 197 | 194 | 196 | 190 | 194 | 192 | 193 | 196 | 196 |
| 14 | 197 | 196 | 195 | 197 | 197 | 195 | 199 | 196 | 198 | 192 | 195 | 194 | 195 | 198 | 197 |
| 16 | 198 | 196 | 196 | 198 | 198 | 196 | 199 | 197 | 199 | 193 | 196 | 195 | 196 | 198 | 198 |
| 18 | 199 | 198 | 197 | 199 | 199 | 197 | 200 | 198 | 200 | 195 | 198 | 197 | 197 | 199 | 199 |
| 20 | 200 | 198 | 198 | 200 | 199 | 198 | 201 | 199 | 201 | 196 | 199 | 198 | 198 | 200 | 200 |
| 22 | 200 | 199 | 199 | 201 | 200 | 199 | 202 | 200 | 201 | 197 | 199 | 198 | 199 | 201 | 200 |
| 24 | 201 | 200 | 200 | 201 | 200 | 199 | 202 | 200 | 201 | 198 | 200 | 199 | 199 | 202 | 201 |
| 26 | 200 | 199 | 199 | 201 | 199 | 199 | 201 | 200 | 201 | 197 | 199 | 198 | 198 | 201 | 200 |
| 28 | 199 | 197 | 197 | 199 | 198 | 197 | 200 | 198 | 199 | 195 | 197 | 197 | 196 | 200 | 198 |
| 30 | 198 | 196 | 196 | 199 | 197 | 196 | 199 | 198 | 199 | 194 | 196 | 196 | 195 | 199 | 197 |
| 31 | 197 | 195 | 195 | 198 | 196 | 195 | 198 | 197 | 198 | 194 | 195 | 195 | 194 | 198 | 197 |
| 32 | 193 | 192 | 189 | 194 | 192 | 189 | 194 | 191 | 194 | 189 | 190 | 190 | 190 | 194 | 192 |

**Table A.5.: SALMONN Target Dimensions for Speech Representation Sets.** For each language (header) and decoder layer (left column). 'Encoder' refers to the encoder speech outputs before the Q-Former and 'Input' refers to the decoder input embeddings after the Q-Former.

| | amh | arb | bul | cat | cmn | deu | ell | eng | est | fin | fra | hin | hye | ind | ita |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Embedding | 50 | 40 | 194 | 204 | 184 | 206 | 35 | 208 | 198 | 198 | 205 | 45 | 37 | 196 | 205 |
| 1 | 117 | 111 | 191 | 200 | 182 | 202 | 112 | 205 | 195 | 195 | 202 | 119 | 113 | 193 | 202 |
| 2 | 24 | 23 | 53 | 64 | 66 | 69 | 29 | 66 | 62 | 66 | 72 | 33 | 23 | 60 | 75 |
| 4 | 90 | 118 | 131 | 140 | 132 | 139 | 123 | 143 | 137 | 138 | 144 | 125 | 107 | 135 | 145 |
| 6 | 138 | 159 | 169 | 171 | 169 | 172 | 161 | 175 | 163 | 169 | 176 | 159 | 148 | 166 | 175 |
| 8 | 158 | 175 | 181 | 182 | 182 | 183 | 175 | 185 | 173 | 180 | 186 | 175 | 166 | 179 | 185 |
| 10 | 168 | 183 | 187 | 188 | 188 | 189 | 183 | 191 | 180 | 187 | 191 | 183 | 175 | 185 | 190 |
| 12 | 176 | 188 | 192 | 191 | 193 | 192 | 189 | 195 | 186 | 191 | 194 | 188 | 183 | 189 | 194 |
| 14 | 178 | 189 | 193 | 193 | 194 | 194 | 190 | 196 | 187 | 192 | 195 | 189 | 185 | 190 | 195 |
| 16 | 177 | 190 | 194 | 194 | 196 | 195 | 190 | 198 | 189 | 193 | 197 | 190 | 186 | 192 | 196 |
| 18 | 182 | 192 | 196 | 196 | 198 | 197 | 192 | 200 | 191 | 194 | 199 | 193 | 189 | 195 | 198 |
| 20 | 183 | 193 | 197 | 197 | 199 | 199 | 192 | 202 | 192 | 195 | 200 | 193 | 189 | 196 | 199 |
| 22 | 186 | 195 | 198 | 199 | 199 | 200 | 194 | 203 | 193 | 196 | 201 | 195 | 192 | 197 | 200 |
| 24 | 188 | 196 | 199 | 199 | 199 | 200 | 195 | 204 | 195 | 197 | 201 | 195 | 193 | 198 | 200 |
| 26 | 191 | 197 | 200 | 200 | 200 | 201 | 196 | 204 | 196 | 198 | 202 | 197 | 195 | 199 | 201 |
| 28 | 192 | 198 | 200 | 200 | 200 | 202 | 197 | 205 | 197 | 198 | 202 | 197 | 196 | 200 | 201 |
| 30 | 193 | 199 | 201 | 201 | 201 | 202 | 198 | 205 | 198 | 199 | 202 | 198 | 196 | 200 | 201 |
| 31 | 196 | 201 | 203 | 203 | 202 | 204 | 200 | 206 | 201 | 201 | 204 | 200 | 198 | 203 | 203 |
| 32 | 195 | 201 | 202 | 199 | 200 | 202 | 200 | 201 | 197 | 200 | 201 | 199 | 197 | 201 | 201 |

| | jpn | kat | khm | kor | lit | mar | nld | pes | rus | sna | snd | tam | tha | tur | yue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Embedding | 183 | 32 | 64 | 133 | 196 | 49 | 203 | 36 | 202 | 192 | 53 | 37 | 53 | 194 | 179 |
| 1 | 183 | 114 | 125 | 151 | 192 | 121 | 200 | 118 | 198 | 190 | 125 | 119 | 116 | 191 | 180 |
| 2 | 68 | 26 | 45 | 54 | 59 | 34 | 71 | 27 | 61 | 68 | 28 | 34 | 37 | 61 | 63 |
| 4 | 140 | 121 | 110 | 136 | 134 | 126 | 141 | 122 | 136 | 140 | 115 | 113 | 116 | 135 | 132 |
| 6 | 171 | 156 | 145 | 169 | 160 | 158 | 174 | 159 | 172 | 162 | 153 | 150 | 151 | 162 | 169 |
| 8 | 183 | 170 | 163 | 182 | 171 | 171 | 184 | 175 | 184 | 170 | 169 | 168 | 168 | 175 | 182 |
| 10 | 189 | 177 | 171 | 188 | 178 | 179 | 190 | 183 | 190 | 176 | 177 | 176 | 177 | 183 | 189 |
| 12 | 193 | 184 | 178 | 192 | 184 | 184 | 193 | 189 | 193 | 182 | 184 | 183 | 184 | 188 | 193 |
| 14 | 194 | 186 | 179 | 193 | 186 | 185 | 195 | 190 | 194 | 183 | 186 | 184 | 186 | 190 | 195 |
| 16 | 196 | 187 | 177 | 194 | 188 | 186 | 196 | 191 | 196 | 184 | 187 | 185 | 186 | 191 | 196 |
| 18 | 198 | 190 | 181 | 196 | 191 | 189 | 198 | 192 | 197 | 187 | 190 | 188 | 189 | 193 | 197 |
| 20 | 199 | 191 | 181 | 197 | 192 | 190 | 199 | 193 | 199 | 188 | 190 | 188 | 189 | 194 | 198 |
| 22 | 200 | 193 | 184 | 198 | 194 | 192 | 201 | 195 | 200 | 188 | 192 | 191 | 191 | 196 | 199 |
| 24 | 200 | 194 | 186 | 198 | 195 | 193 | 201 | 196 | 200 | 190 | 194 | 192 | 193 | 197 | 199 |
| 26 | 201 | 196 | 189 | 200 | 196 | 195 | 202 | 197 | 201 | 191 | 195 | 194 | 195 | 198 | 200 |
| 28 | 202 | 197 | 191 | 201 | 197 | 196 | 202 | 198 | 202 | 192 | 196 | 195 | 196 | 199 | 200 |
| 30 | 203 | 197 | 192 | 202 | 198 | 196 | 202 | 199 | 202 | 193 | 195 | 196 | 197 | 200 | 201 |
| 31 | 204 | 198 | 194 | 203 | 201 | 198 | 204 | 200 | 204 | 195 | 196 | 198 | 199 | 202 | 201 |
| 32 | 203 | 198 | 194 | 203 | 198 | 197 | 202 | 200 | 203 | 191 | 192 | 197 | 199 | 200 | 200 |

**Table A.6.: SALMONN Target Dimensions for Text Representation Sets.** For each language (header) and decoder layer (left column). ’Embedding’ refers to the Vicuna text embeddings.

## A.2. Target Dimensions for Cross-Lingual Analysis

| | SeamlessM4T | | SONAR | | SALMONN | |
|---|---|---|---|---|---|---|
| Layer | Speech | Text | Speech | Text | Speech | Text |
| Encoder/Embedding | | | | | 133 | 156 |
| Input | 91 | 155 | 90 | 155 | 133 | |
| 1 | 117 | 153 | 122 | 153 | 133 | 160 |
| 2 | 117 | 151 | 125 | 152 | 133 | 130 |
| 4 | 129 | 149 | 127 | 150 | 137 | 134 |
| 6 | 132 | 150 | 135 | 151 | 144 | 141 |
| 8 | 129 | 151 | 131 | 151 | 148 | 145 |
| 10 | 135 | 152 | 124 | 151 | 152 | 149 |
| 12 | 142 | 153 | 139 | 151 | 154 | 151 |
| 14 | 134 | 153 | 143 | 152 | 155 | 153 |
| 16 | 133 | 153 | 138 | 152 | 156 | 154 |
| 18 | 137 | 153 | 141 | 152 | 156 | 156 |
| 20 | 143 | 154 | 144 | 150 | 157 | 157 |
| 22 | 143 | 154 | 147 | 153 | 157 | 158 |
| 23 | 148 | 155 | 150 | 155 | | |
| 24 | 149 | 156 | 150 | 158 | 158 | 159 |
| Adaptor/Pooling | 154 | 156 | 158 | 158 | | |
| 26 | | | | | 157 | 159 |
| 28 | | | | | 156 | 160 |
| 30 | | | | | 155 | 160 |
| 31 | | | | | 155 | 161 |
| 32 | | | | | 152 | 159 |

**Table A.7.: Target Dimensions for Cross-Lingual Analysis.** For the cross-lingual analysis we used one target dimension for each layer and for all languages, leaving at least 90% of the variance behind. 'Encoder/Embedding' refers to the encoder speech outputs and text embeddings of SALMONN. 'Input' refers to the encoder input embeddings of SeamlessM4T and SONAR, for SALMONN it refers to the encoder speech outputs after the Q-Former. 'Adaptor/Pooling' refers to the representations after the length adaptor of SeamlessM4T and pooling of SONAR.

## A.3. Cross-Lingual Analysis Results

### A.3.1. SeamlessM4T - Cross-Modal Similarities

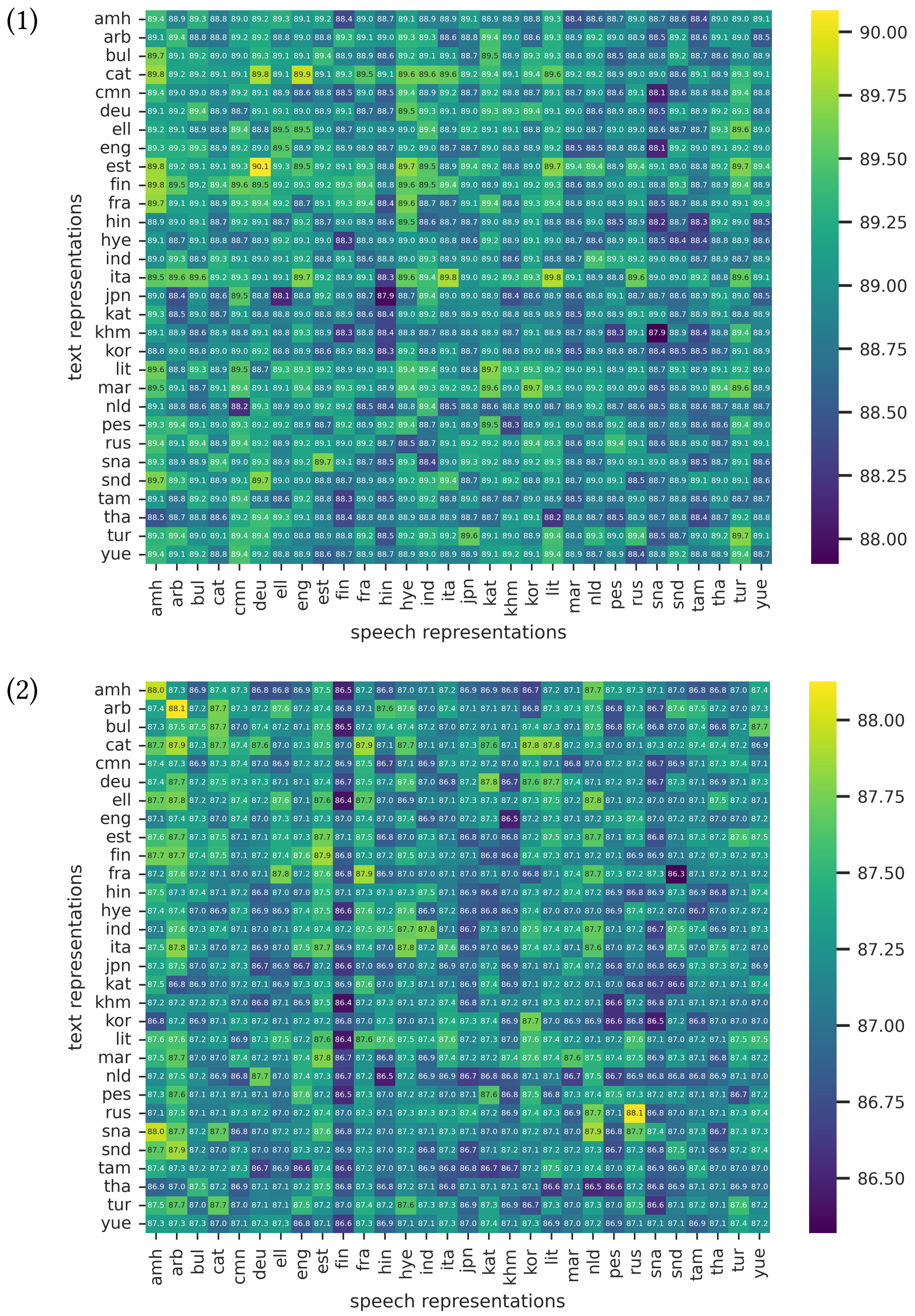


**Figure A.1.: SeamlessM4T Cross-Lingual & Cross-Modal Similarity Analysis Results 1.** With the results of (1) the input embeddings and (2) the first encoder layer.

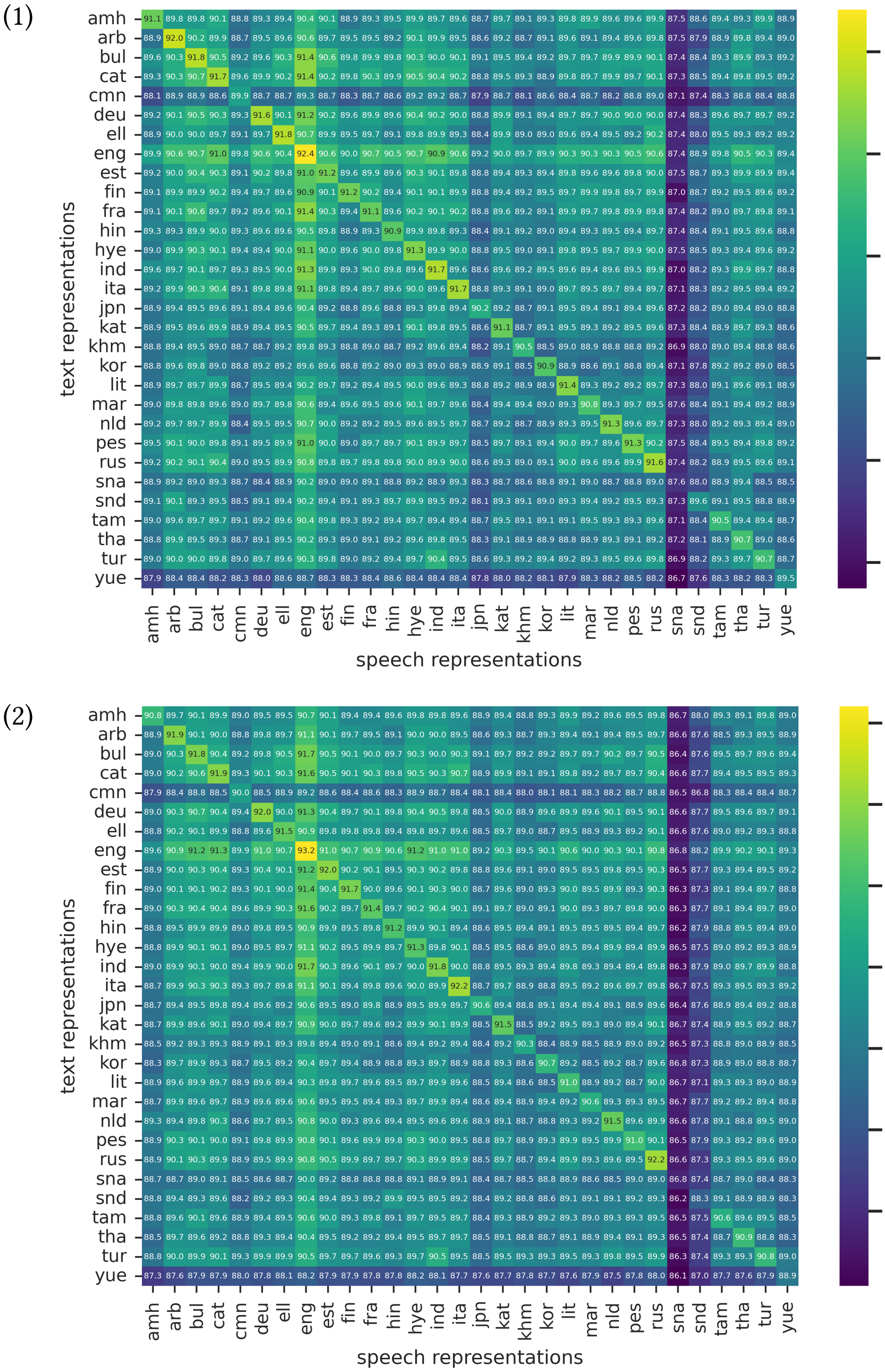


**Figure A.2.: SeamlessM4T Cross-Lingual & Cross-Modal Similarity Analysis Results 2.** With the results (1) of the last encoder layer and (2) from after the length adaptor.

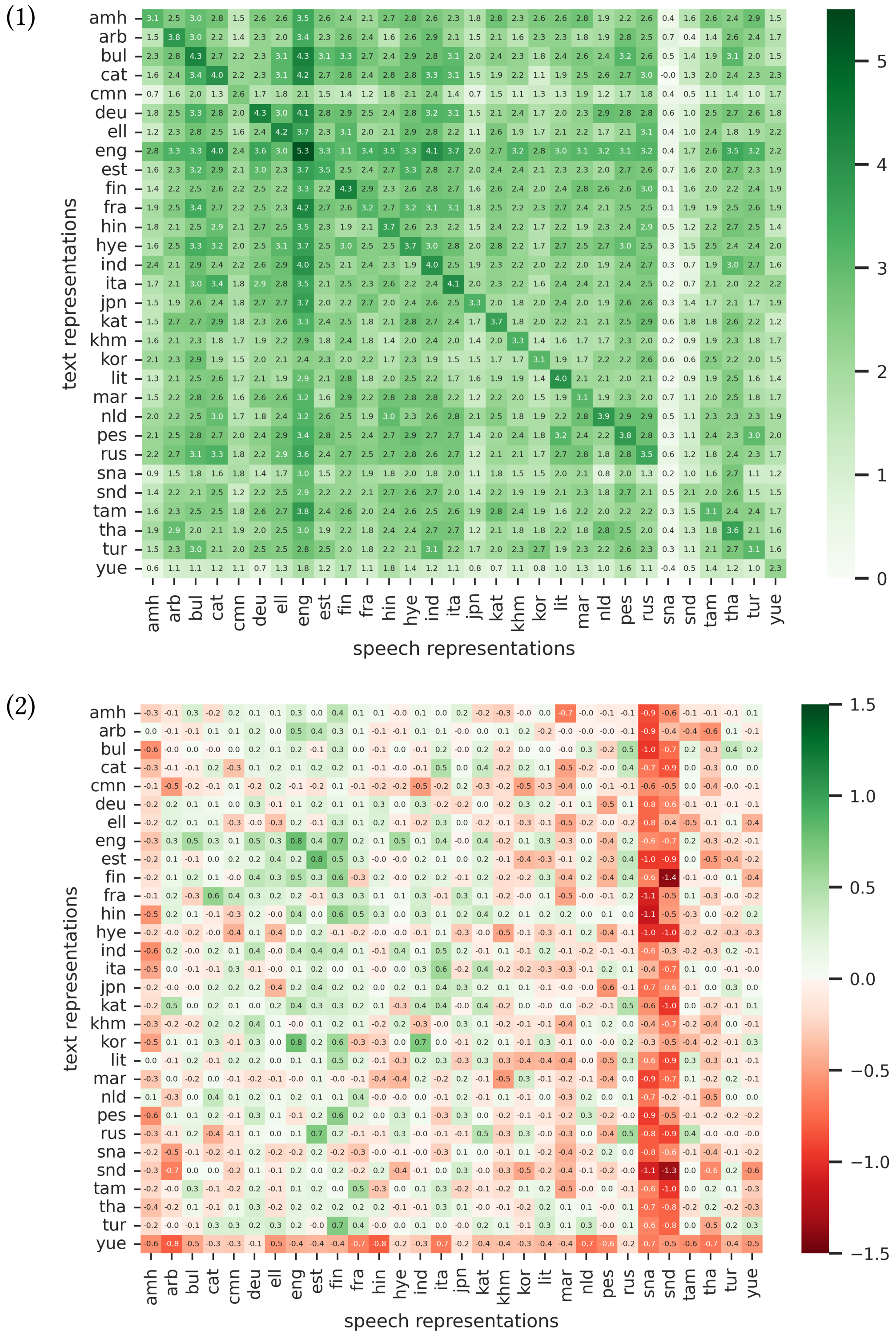


**Figure A.3.: SeamlessM4T Cross-Lingual & Cross-Modal Similarity Differences.** With the differences (1) between the first and last encoder layer and (2) between the last encoder layer and after the length adaptor.

## A.3.2. SeamlessM4T - Intra-Speech Similarities

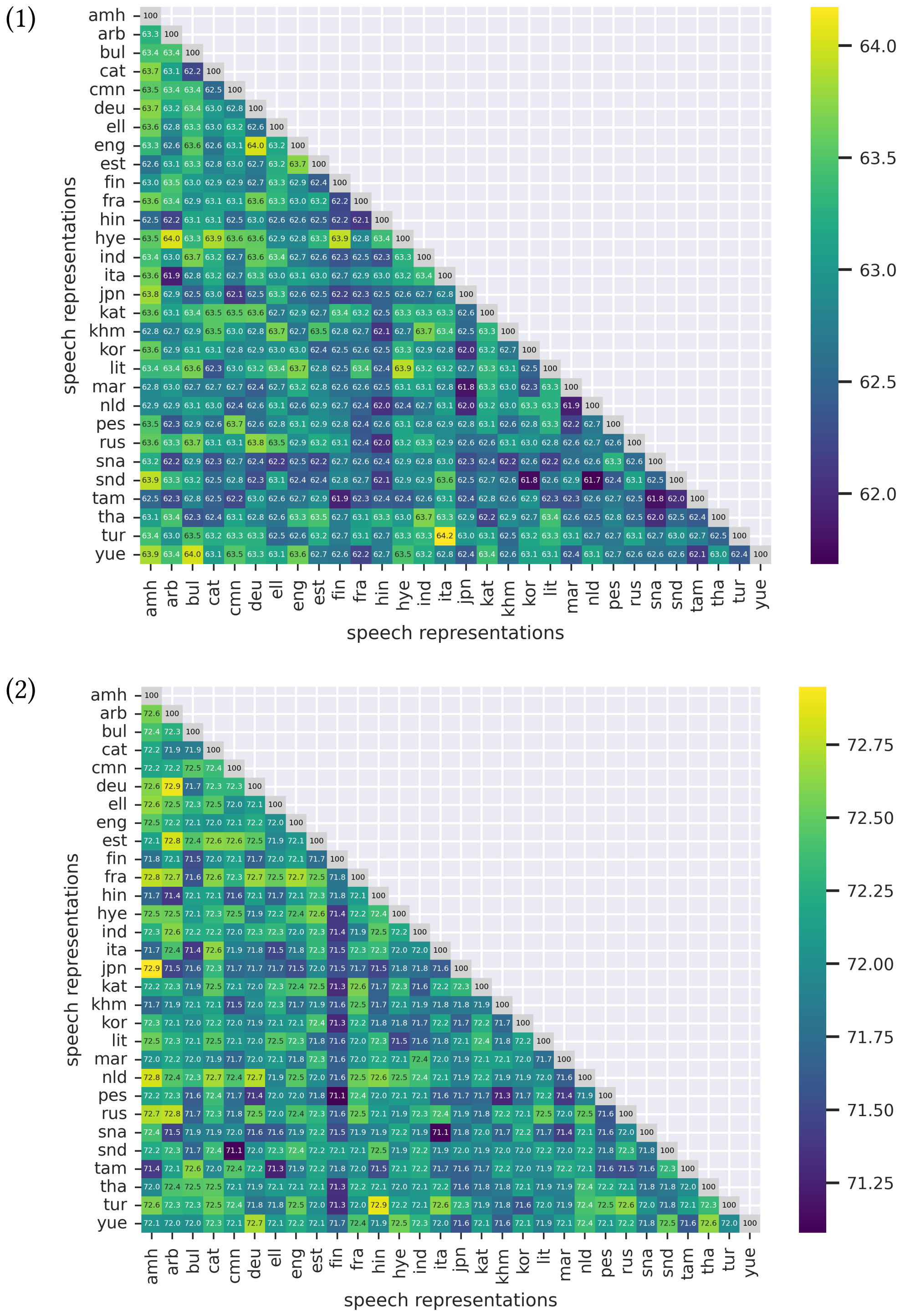


**Figure A.4.: SeamlessM4T Cross-Lingual & Intra-Speech Similarity Analysis Results 1.** With the results of (1) the input embeddings and (2) the first encoder layer.

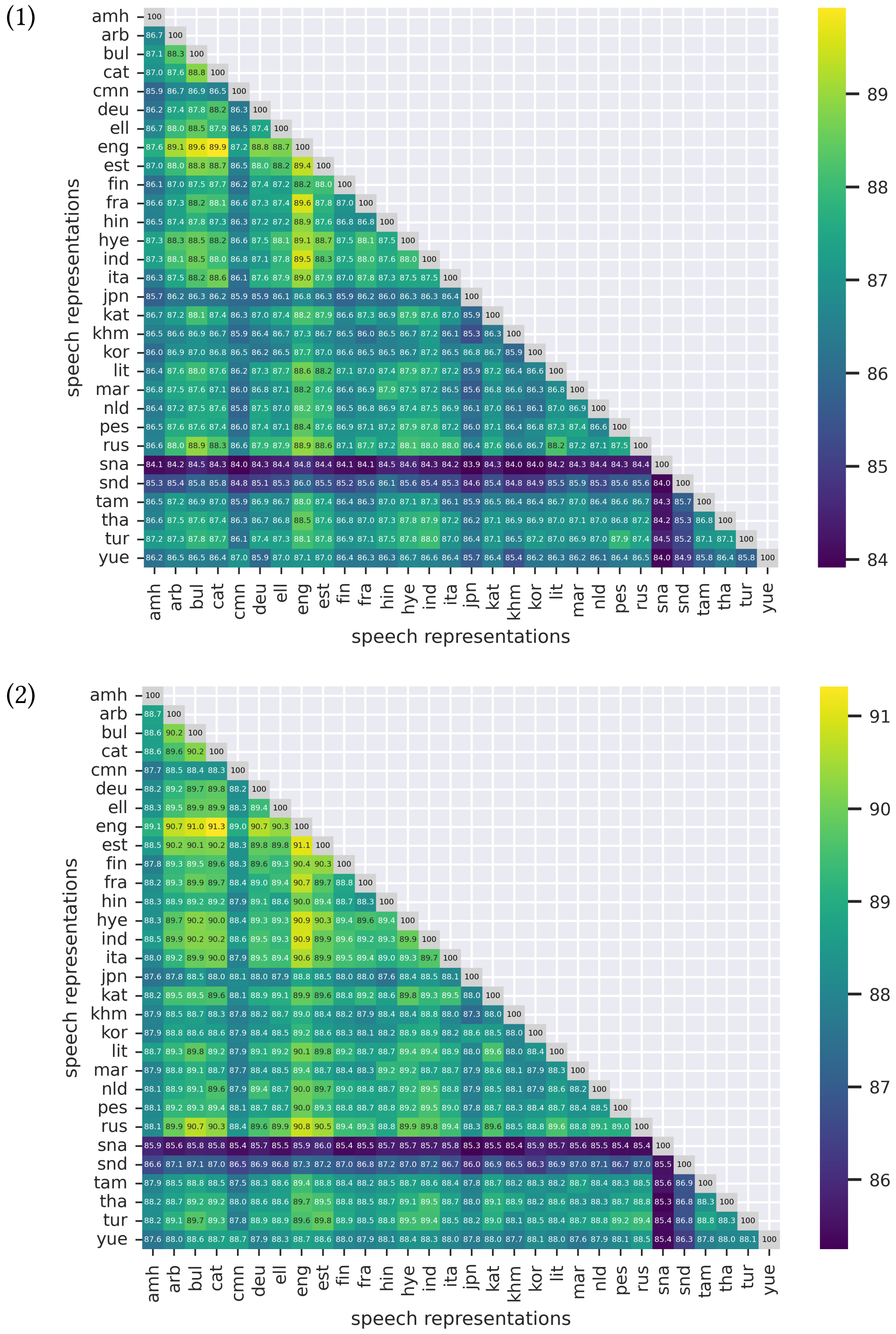


**Figure A.5.: SeamlessM4T Cross-Lingual & Intra-Speech Similarity Analysis Results 2.** With the results (1) of the last encoder layer and (2) from after the length adaptor.

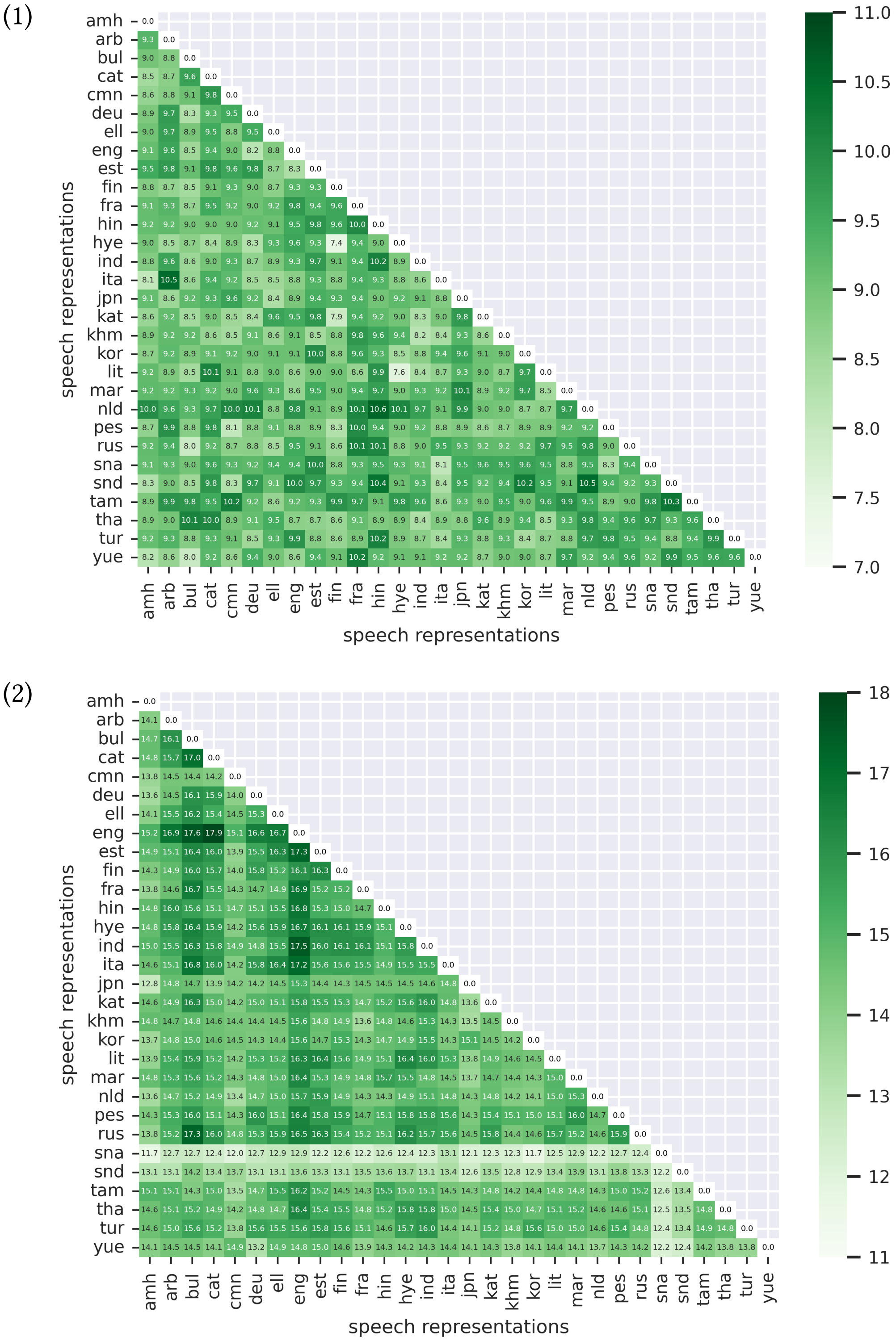


**Figure A.6.: SeamlessM4T Cross-Lingual & Intra-Speech Similarity Differences 1.** With the differences (1) between the input embeddings and the first encoder layer and (2) between the first and last encoder layer.

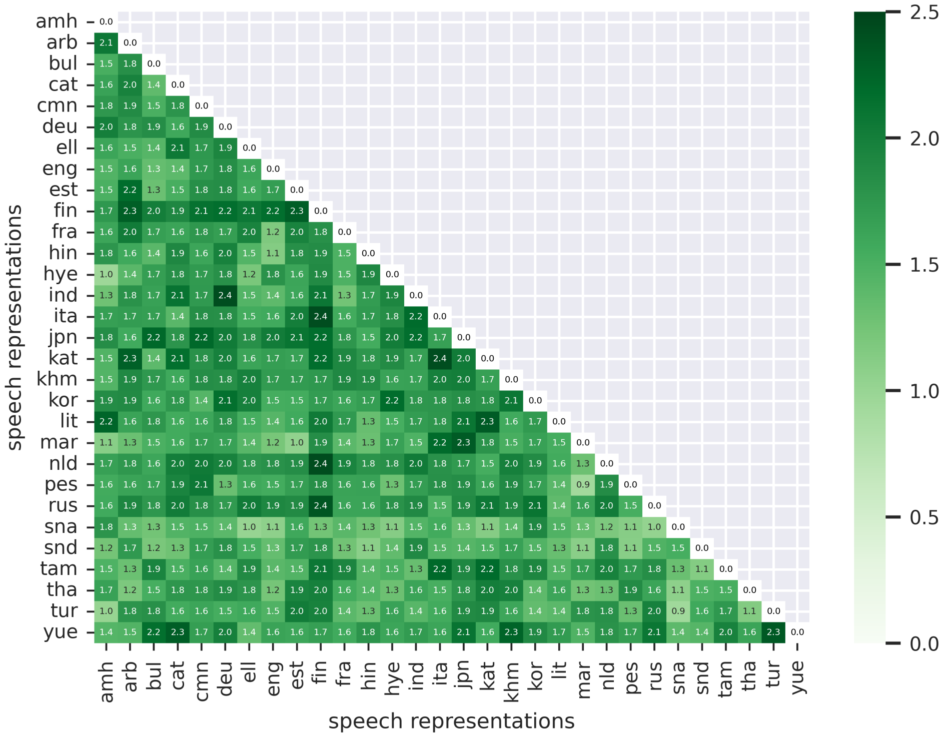
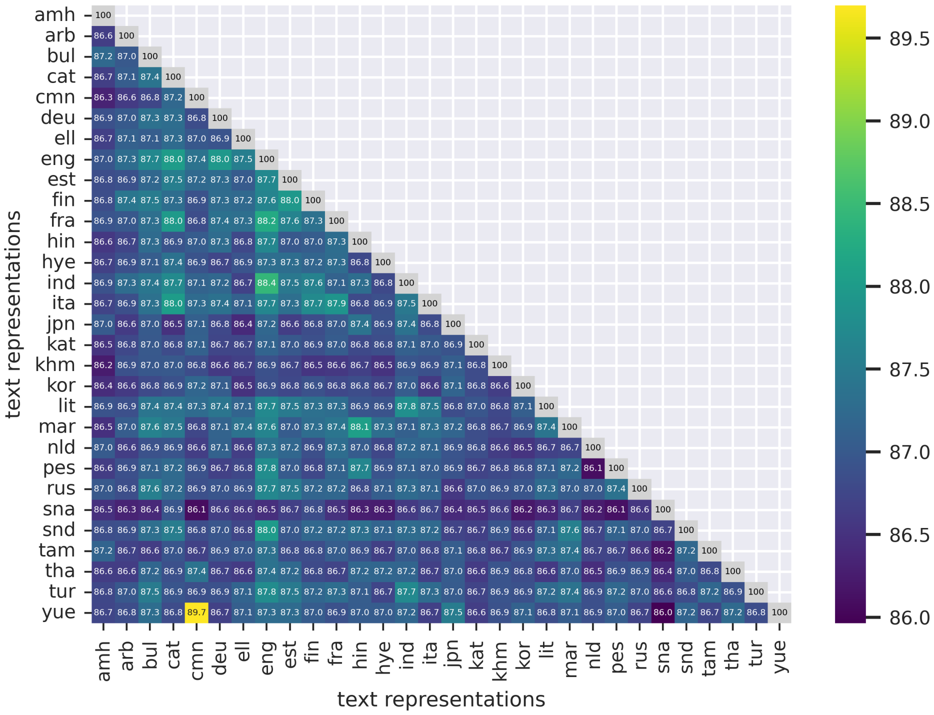


**Figure A.7.: SeamlessM4T Cross-Lingual Speech Similarity Differences 2.** Differences between the last encoder layer and after the length adaptor.

### A.3.3. SeamlessM4T - Intra-Text Similarities

(1)



(2)

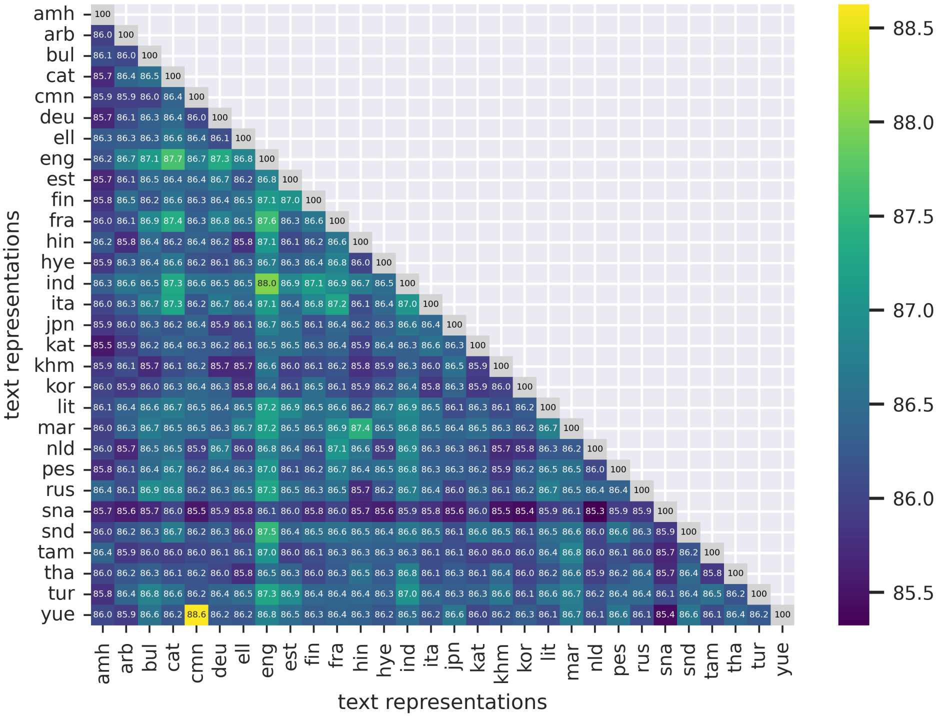


**Figure A.8.: SeamlessM4T Cross-Lingual & Intra-Text Similarity Analysis Results 1.** With the results of (1) the input embeddings and (2) the first encoder layer.

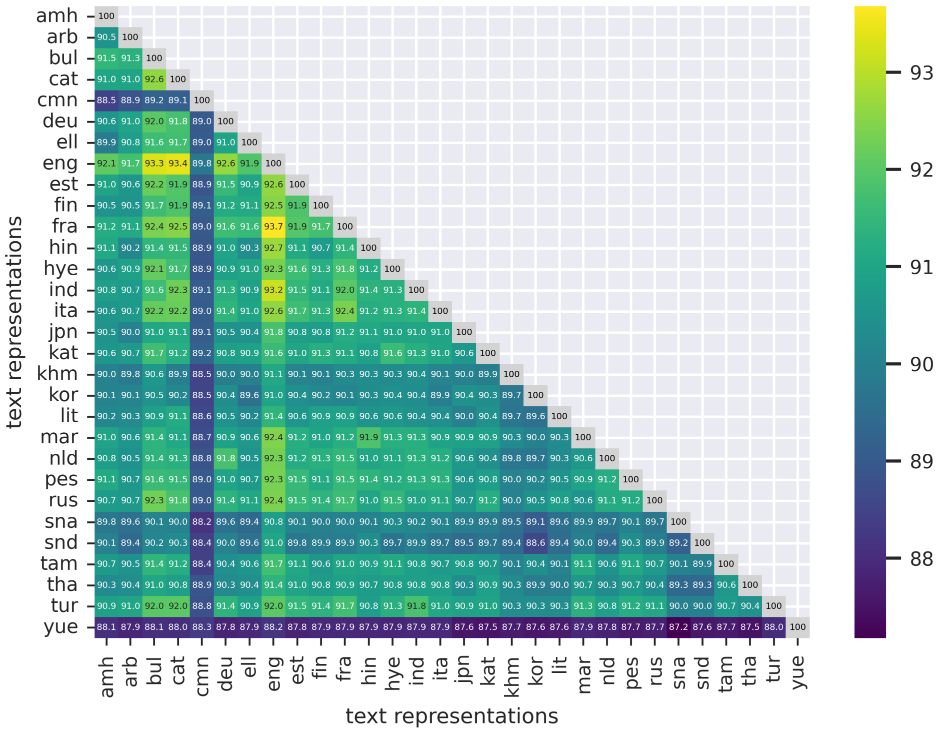


**Figure A.9.: SeamlessM4T Cross-Lingual & Intra-Text Similarity Analysis Results 2.** With the results of the last encoder layer.

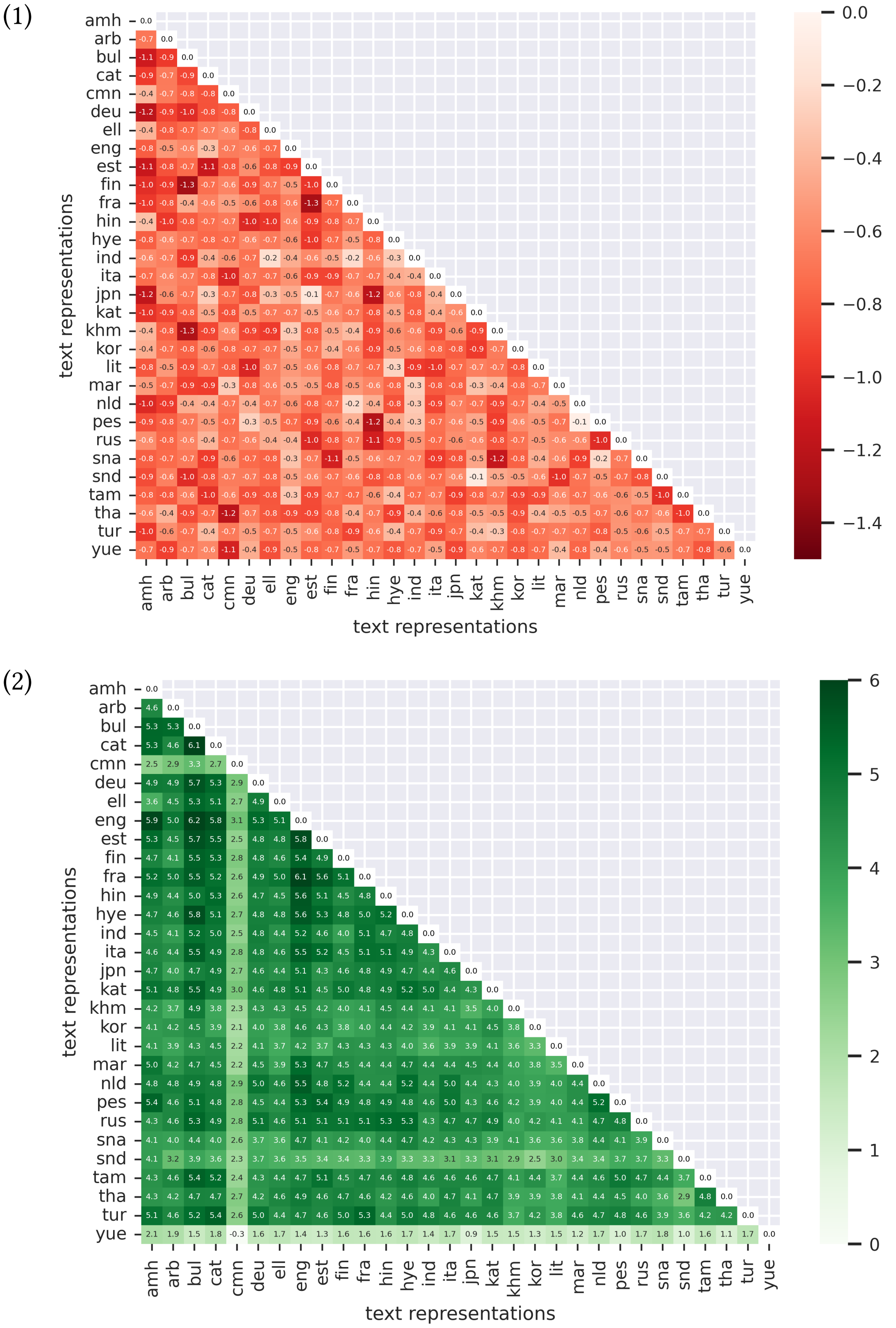


**Figure A.10.: SeamlessM4T Cross-Lingual & Intra-Text Similarity Differences.** With the differences (1) between the input embeddings and the first encoder layer and (2) between the first and last encoder layer.

### A.3.4. SONAR - Cross-Modal Similarities

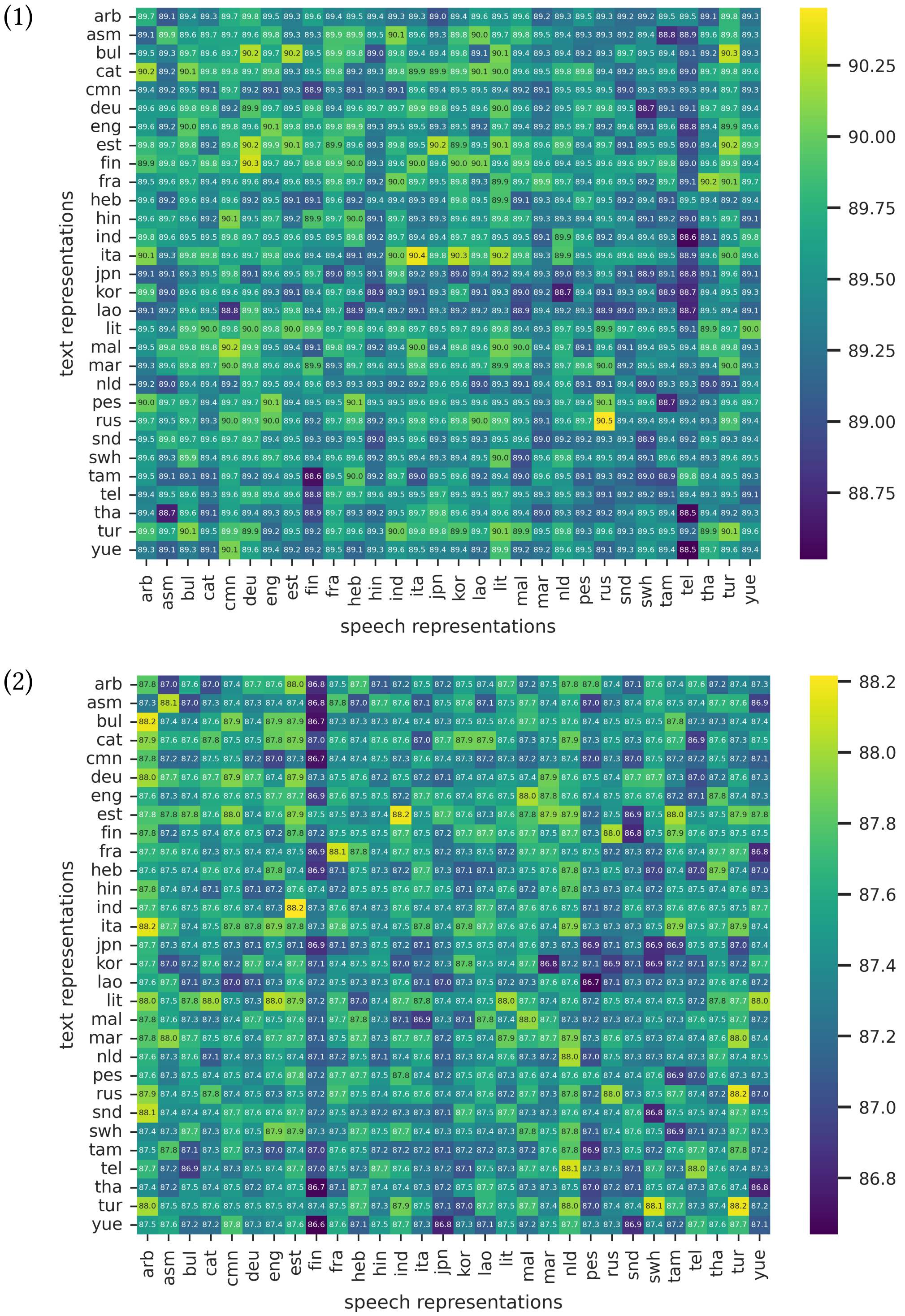


**Figure A.11.: SONAR Cross-Lingual & Cross-Modal Similarity Analysis Results 1.** With the results of (1) the input embeddings and (2) the first encoder layer.

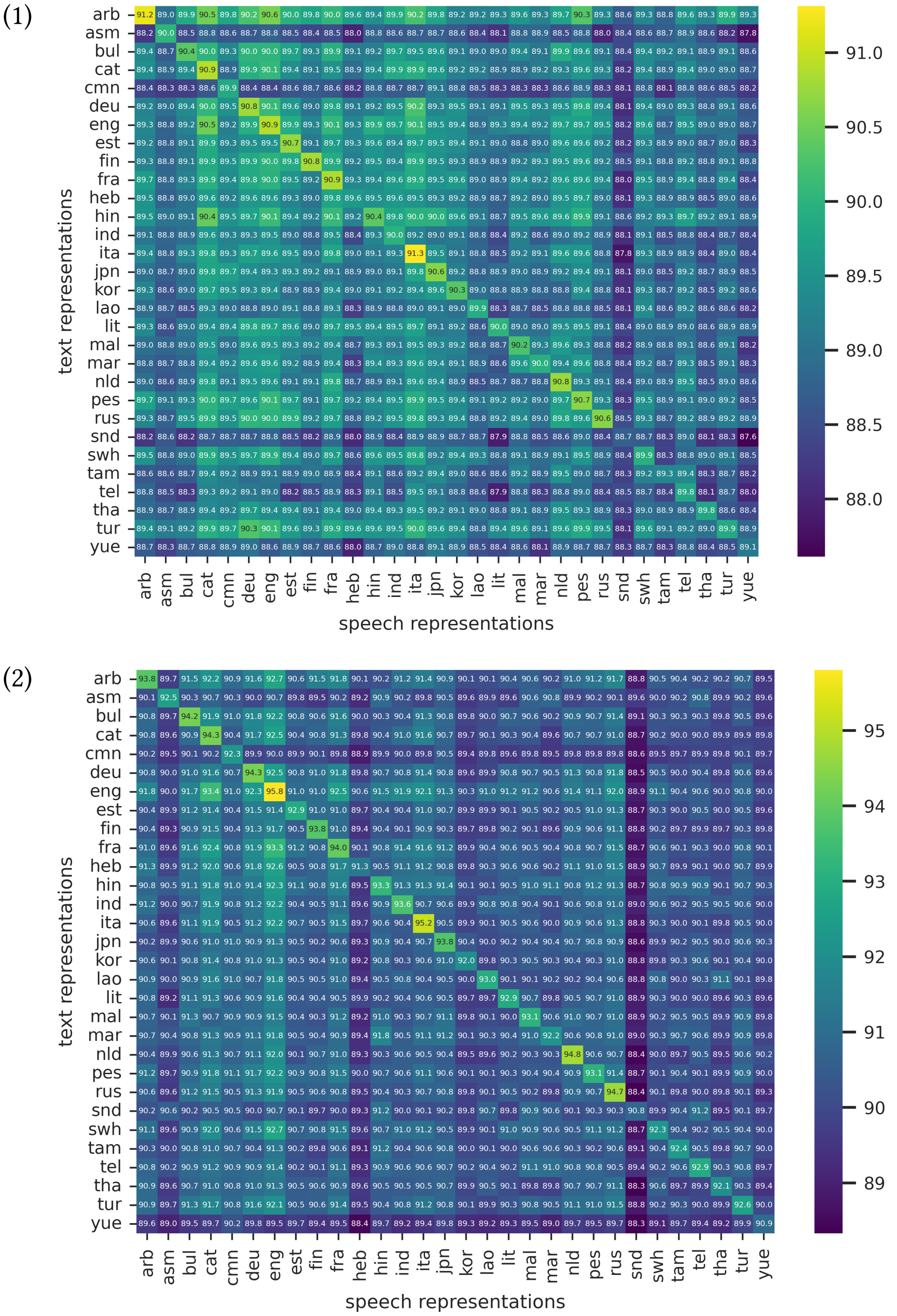


**Figure A.12.: SONAR Cross-Lingual & Cross-Modal Similarity Analysis Results 2.** With the results of (1) the last encoder layer and (2) the final language-agnostic embeddings.

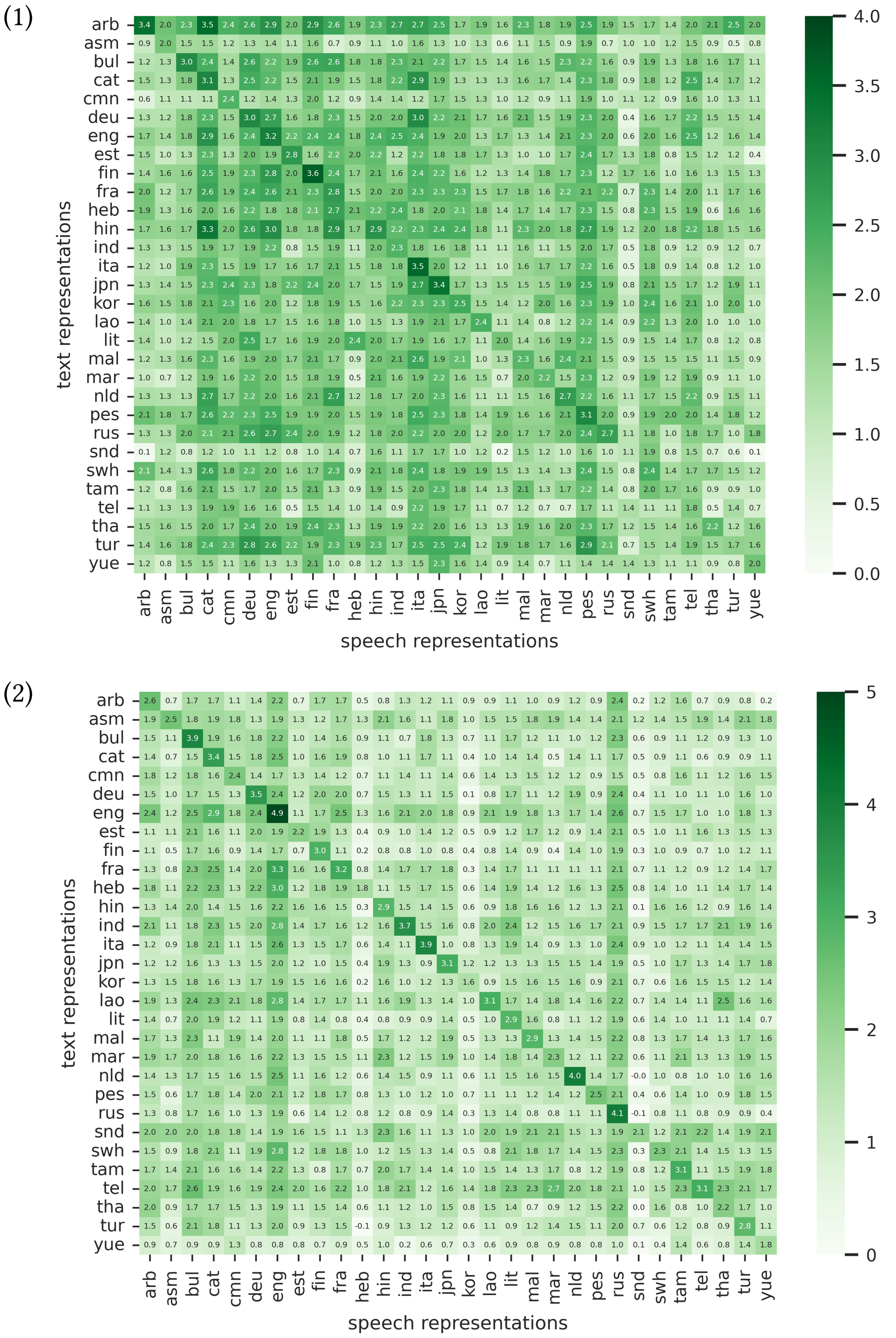


**Figure A.13.: SONAR Cross-Lingual & Cross-Modal Similarity Differences.** With the differences (1) between the first and last encoder layer and (2) between the last encoder layer and the final language-agnostic embeddings.

### A.3.5. SONAR - Intra-Speech Similarities

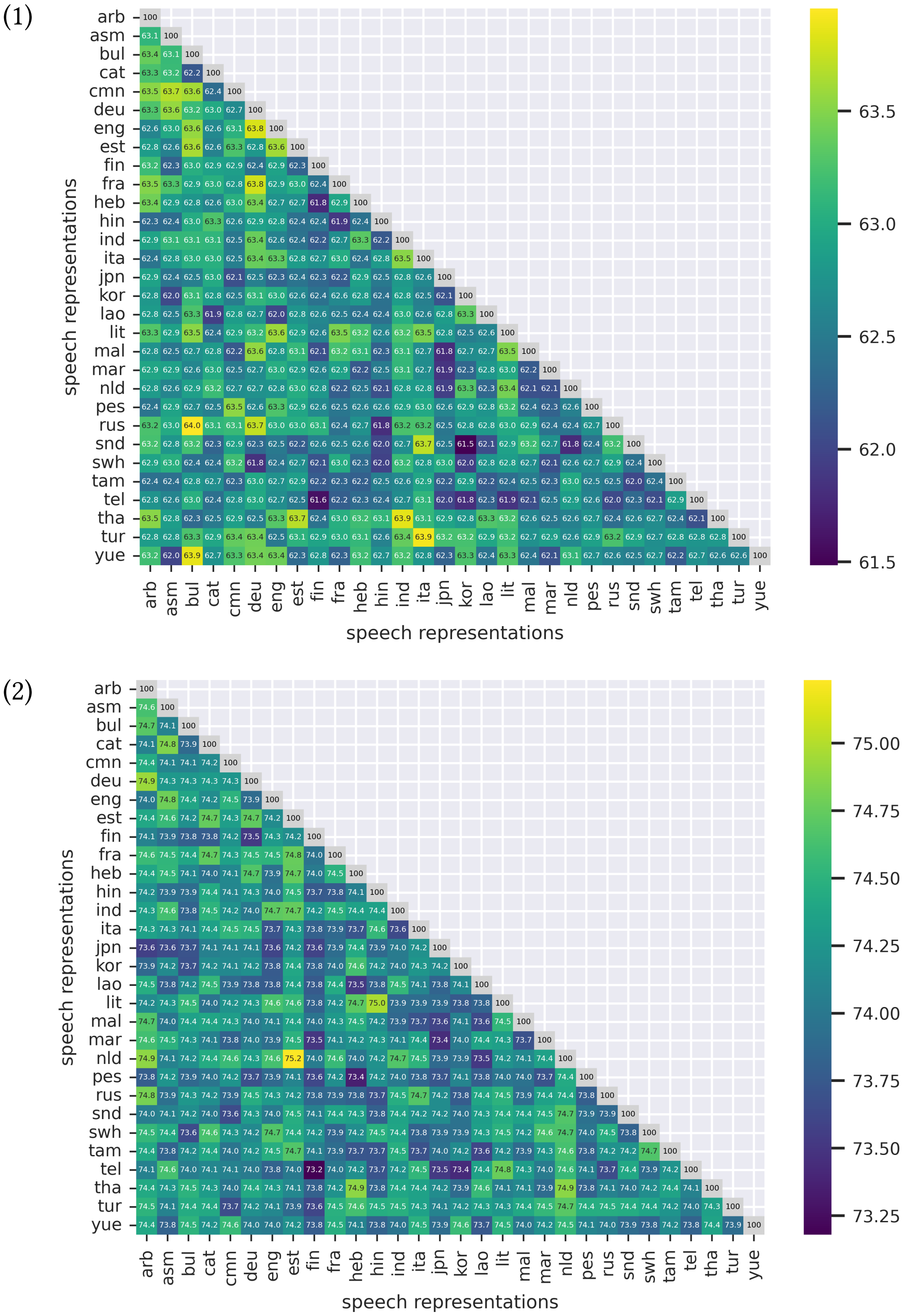


**Figure A.14.: SONAR Cross-Lingual & Intra-Speech Similarity Analysis Results 1.** With the results of (1) the input embeddings and (2) the first encoder layer.

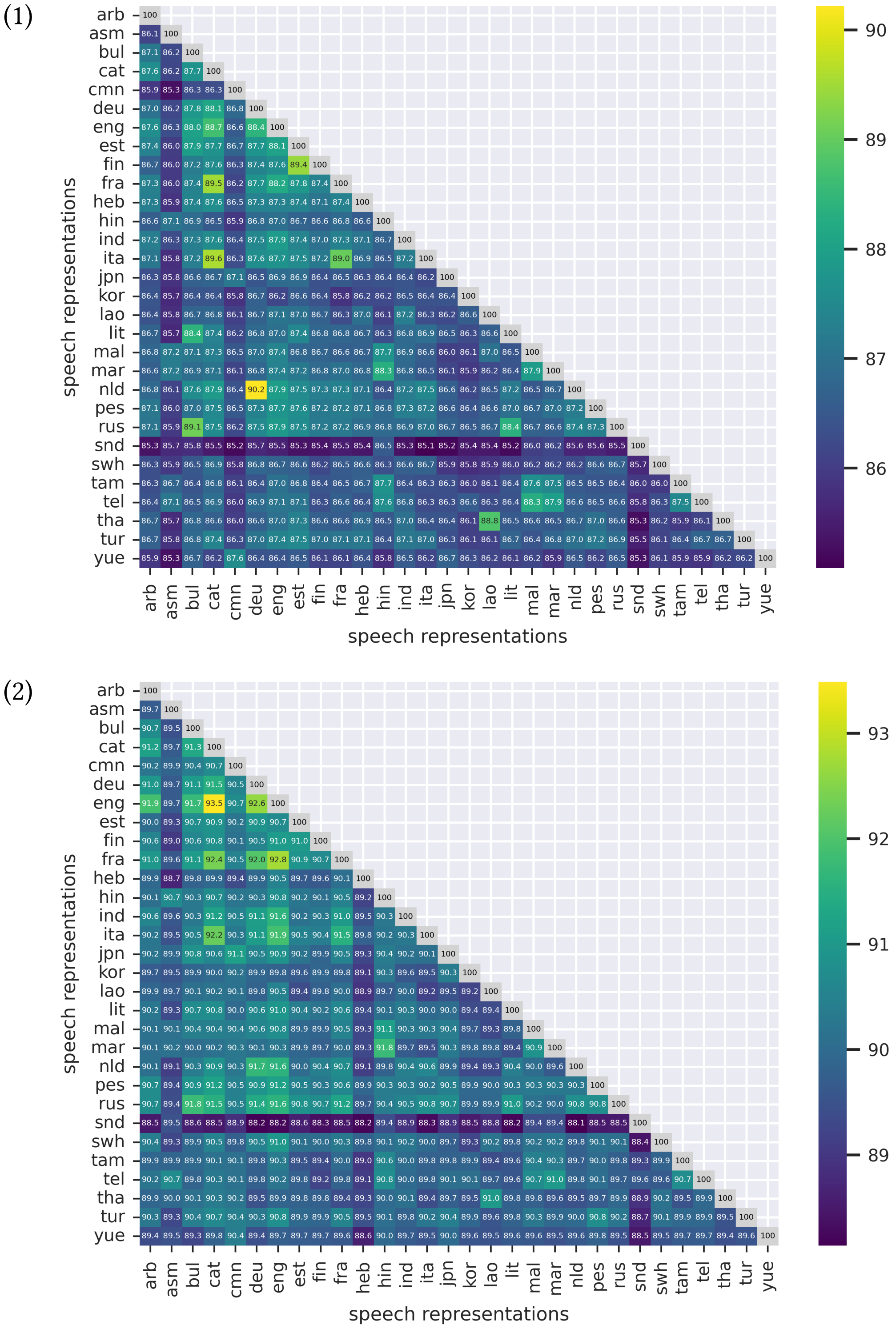


**Figure A.15.: SONAR Cross-Lingual & Intra-Speech Similarity Analysis Results 2.** With the results (1) of the last encoder layer and (2) of the final language-agnostic embeddings.

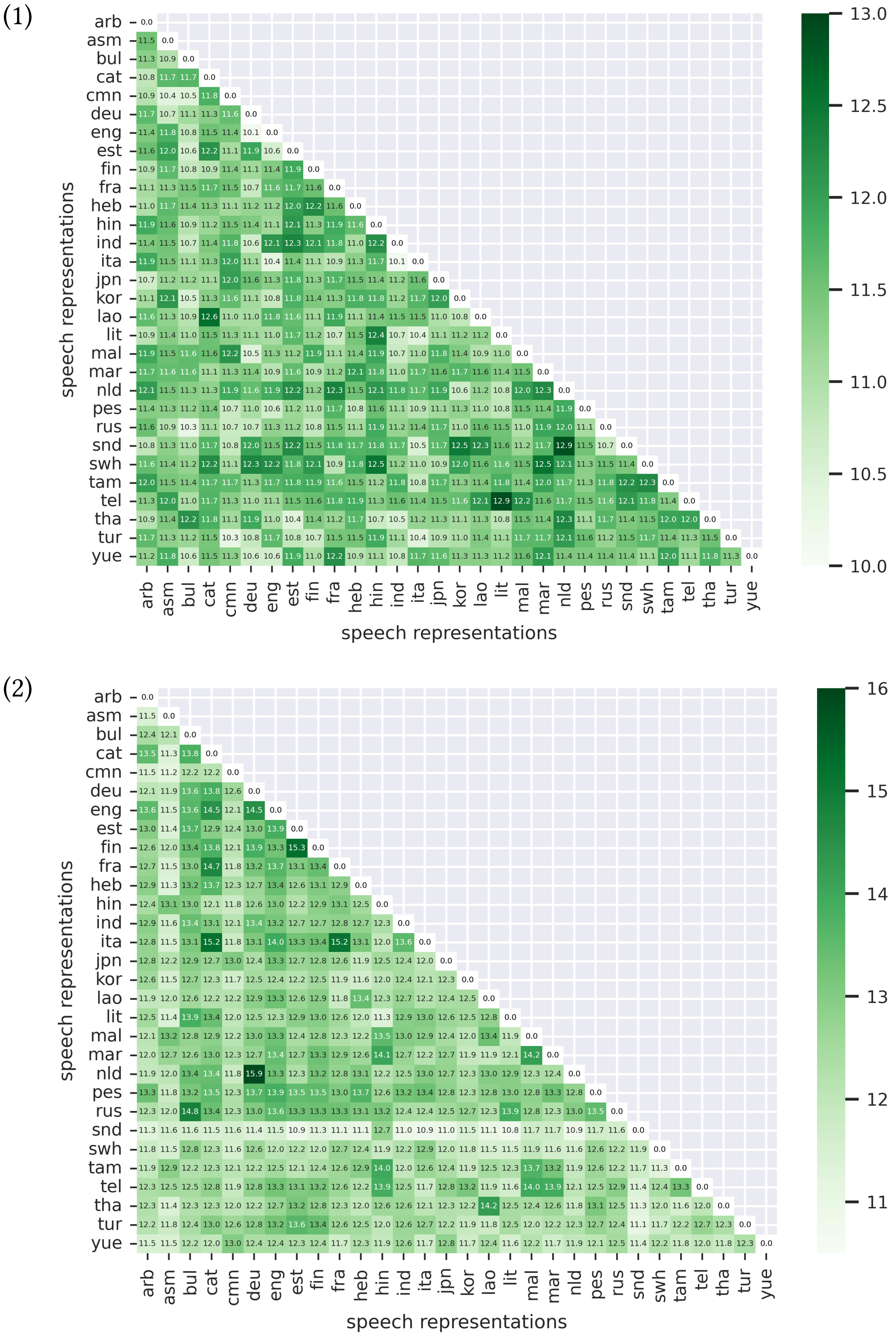


**Figure A.16.: SONAR Cross-Lingual & Intra-Speech Similarity Differences 1.** With the differences (1) between the input embeddings and the first encoder layer and (2) between the first and last encoder layer.

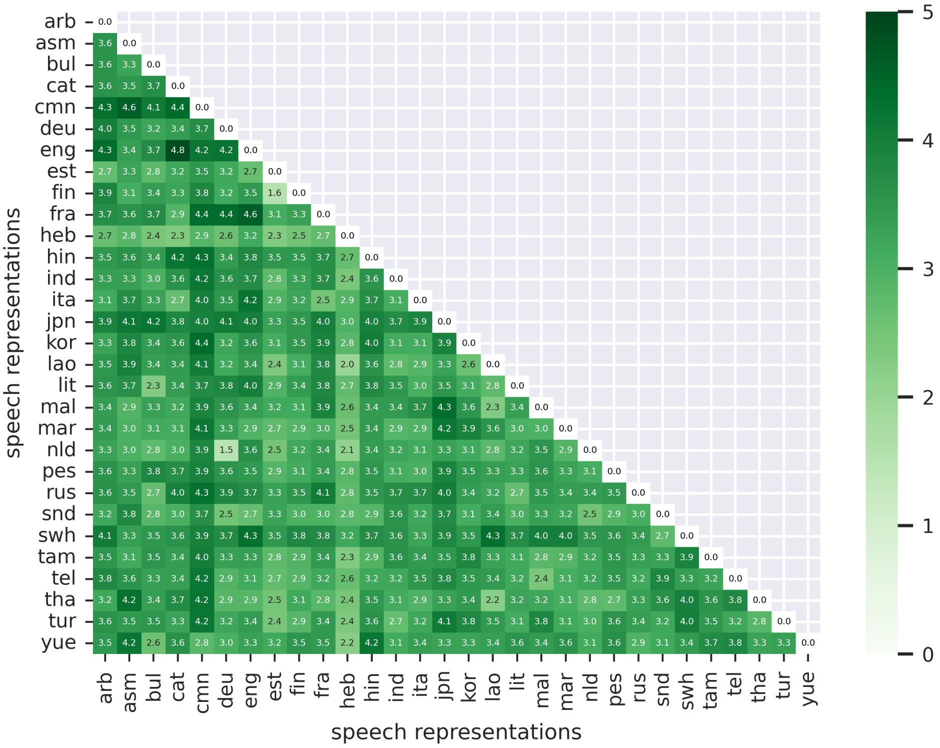


**Figure A.17.: SONAR Cross-Lingual & Intra-Speech Similarity Differences 2.** With the differences between the last encoder layer and the final language-agnostic embeddings.

### A.3.6. SONAR - Intra-Text Similarities

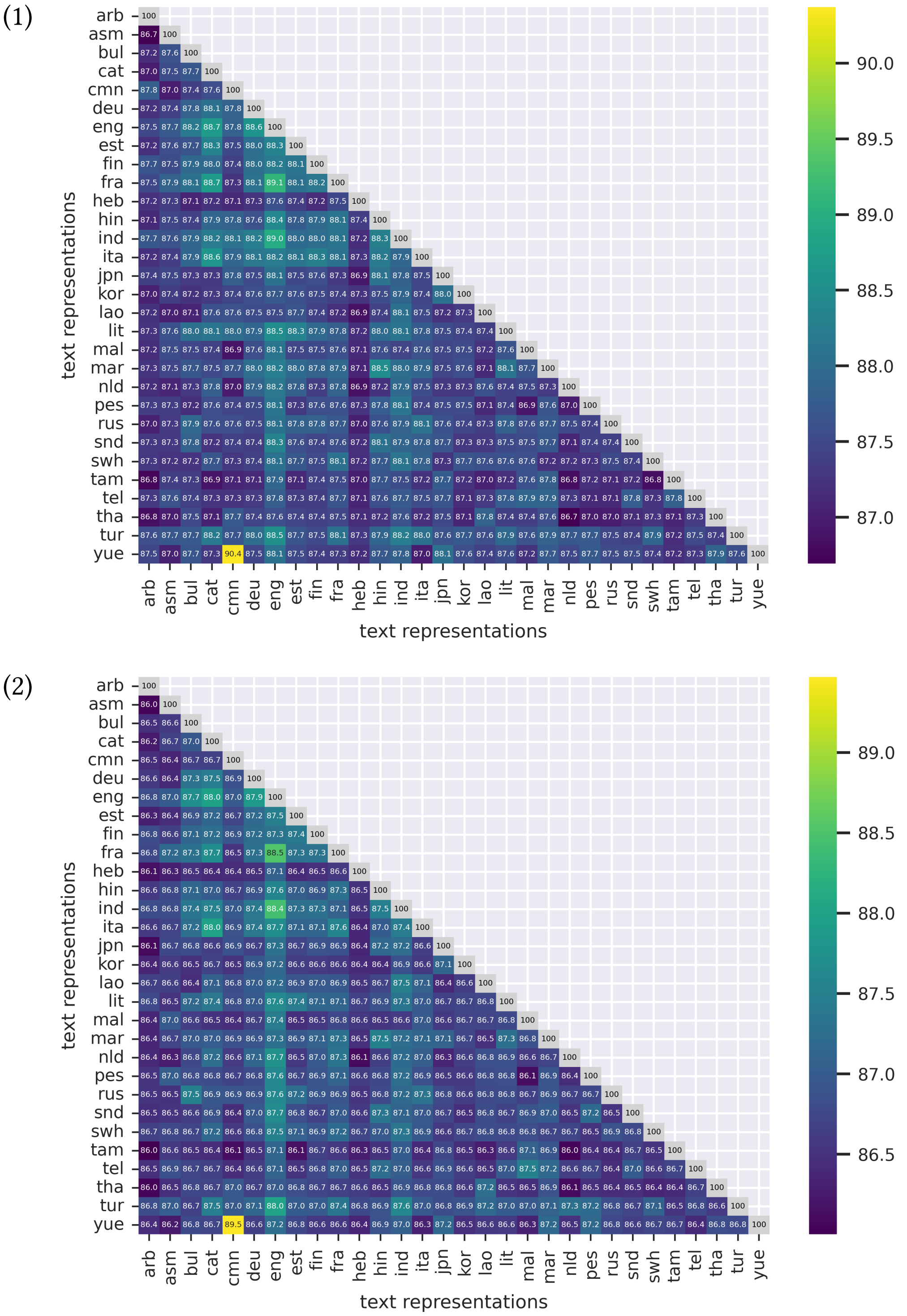
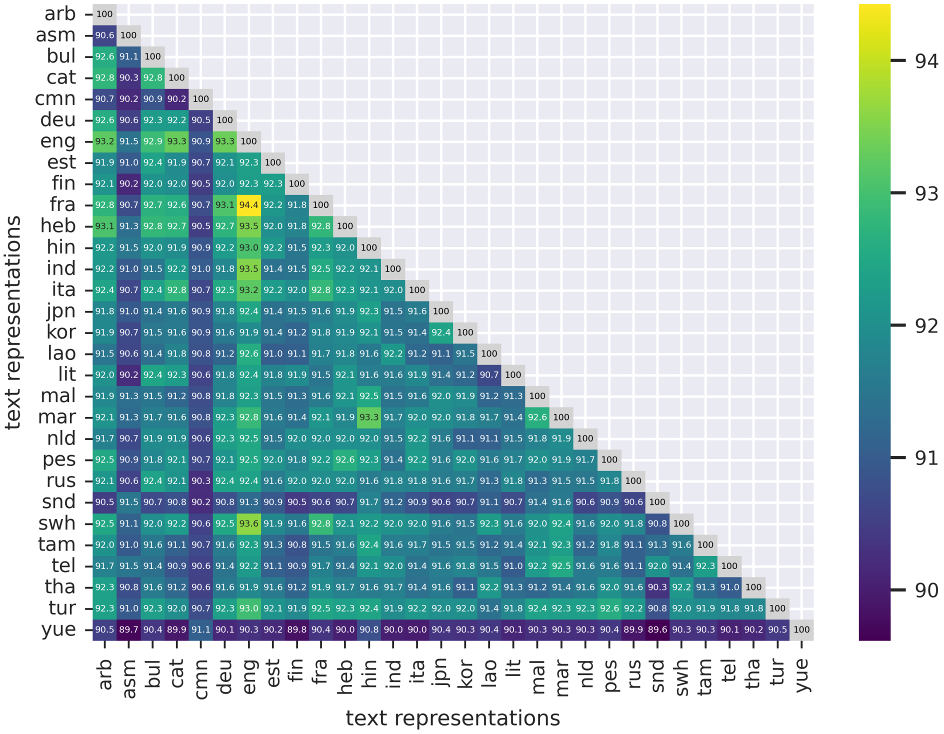


**Figure A.18.: SONAR Cross-Lingual & Intra-Text Similarity Analysis Results 1.** With the results of (1) the input embeddings and (2) the first encoder layer.



| | arb | asm | bul | cat | cmn | deu | eng | est | fin | fra | heb | hin | ind | ita | jpn | kor | lao | lit | mal | mar | nld | pes | rus | snd | swh | tam | tel | tha | tur | yue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| arb | 100 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| asm | 90.6 | 100 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| bul | 92.6 | 91.1 | 100 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| cat | 92.8 | 90.3 | 92.8 | 100 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| cmn | 90.7 | 90.2 | 90.9 | 90.2 | 100 | | | | | | | | | | | | | | | | | | | | | | | | | |
| deu | 92.6 | 90.6 | 92.3 | 92.2 | 90.5 | 100 | | | | | | | | | | | | | | | | | | | | | | | | |
| eng | 93.2 | 91.5 | 92.9 | 93.3 | 90.9 | 93.3 | 100 | | | | | | | | | | | | | | | | | | | | | | | |
| est | 91.9 | 91.0 | 92.4 | 91.9 | 90.7 | 92.1 | 92.3 | 100 | | | | | | | | | | | | | | | | | | | | | | |
| fin | 92.1 | 90.2 | 92.0 | 92.0 | 90.5 | 92.0 | 92.3 | 92.3 | 100 | | | | | | | | | | | | | | | | | | | | | |
| fra | 92.8 | 90.7 | 92.7 | 92.6 | 90.7 | 93.1 | 94.4 | 92.2 | 91.8 | 100 | | | | | | | | | | | | | | | | | | | | |
| heb | 93.1 | 91.3 | 92.8 | 92.7 | 90.5 | 92.7 | 93.5 | 92.0 | 91.8 | 92.8 | 100 | | | | | | | | | | | | | | | | | | | |
| hin | 92.2 | 91.5 | 92.0 | 91.9 | 90.9 | 92.2 | 93.0 | 92.2 | 91.5 | 92.3 | 92.0 | 100 | | | | | | | | | | | | | | | | | | |
| ind | 92.2 | 91.0 | 91.5 | 92.2 | 91.0 | 91.8 | 93.5 | 91.4 | 91.5 | 92.5 | 92.2 | 92.1 | 100 | | | | | | | | | | | | | | | | | |
| ita | 92.4 | 90.7 | 92.4 | 92.8 | 90.7 | 92.5 | 93.2 | 92.2 | 92.0 | 92.8 | 92.3 | 92.1 | 92.0 | 100 | | | | | | | | | | | | | | | | |
| jpn | 91.8 | 91.0 | 91.4 | 91.6 | 90.9 | 91.8 | 92.4 | 91.4 | 91.5 | 91.6 | 91.9 | 92.3 | 91.5 | 91.6 | 100 | | | | | | | | | | | | | | | |
| kor | 91.9 | 90.7 | 91.5 | 91.6 | 90.9 | 91.6 | 91.9 | 91.4 | 91.2 | 91.8 | 91.9 | 92.1 | 91.5 | 91.4 | 92.4 | 100 | | | | | | | | | | | | | | |
| lao | 91.5 | 90.6 | 91.4 | 91.8 | 90.8 | 91.2 | 92.6 | 91.0 | 91.1 | 91.7 | 91.8 | 91.6 | 92.2 | 91.2 | 91.1 | 91.5 | 100 | | | | | | | | | | | | | |
| lit | 92.0 | 90.2 | 92.4 | 92.3 | 90.6 | 91.8 | 92.4 | 91.8 | 91.9 | 91.5 | 92.1 | 91.6 | 91.6 | 91.9 | 91.4 | 91.2 | 90.7 | 100 | | | | | | | | | | | | |
| mal | 91.9 | 91.3 | 91.5 | 91.2 | 90.8 | 91.8 | 92.3 | 91.5 | 91.3 | 91.6 | 92.1 | 92.5 | 91.5 | 91.6 | 92.0 | 91.9 | 91.2 | 91.3 | 100 | | | | | | | | | | | |
| mar | 92.1 | 91.3 | 91.7 | 91.6 | 90.8 | 92.3 | 92.8 | 91.6 | 91.4 | 92.1 | 91.9 | 93.3 | 91.7 | 92.0 | 92.0 | 91.8 | 91.7 | 91.4 | 92.6 | 100 | | | | | | | | | | |
| nld | 91.7 | 90.7 | 91.9 | 91.9 | 90.6 | 92.3 | 92.5 | 91.5 | 92.0 | 92.0 | 92.0 | 92.0 | 91.5 | 92.2 | 91.6 | 91.1 | 91.1 | 91.5 | 91.8 | 91.9 | 100 | | | | | | | | | |
| pes | 92.5 | 90.9 | 91.8 | 92.1 | 90.7 | 92.1 | 92.5 | 92.0 | 91.8 | 92.2 | 92.6 | 92.3 | 91.4 | 92.2 | 91.6 | 92.0 | 91.6 | 91.7 | 92.0 | 91.9 | 91.7 | 100 | | | | | | | | |
| rus | 92.1 | 90.6 | 92.4 | 92.1 | 90.3 | 92.4 | 92.4 | 91.6 | 92.0 | 92.0 | 92.0 | 91.6 | 91.8 | 91.8 | 91.6 | 91.7 | 91.3 | 91.8 | 91.3 | 91.5 | 91.5 | 91.8 | 100 | | | | | | | |
| snd | 90.5 | 91.5 | 90.7 | 90.8 | 90.2 | 90.8 | 91.3 | 90.9 | 90.5 | 90.6 | 90.7 | 91.7 | 91.2 | 90.9 | 90.6 | 90.7 | 91.1 | 90.7 | 91.4 | 91.6 | 90.6 | 90.9 | 90.6 | 100 | | | | | | |
| swh | 92.5 | 91.1 | 92.0 | 92.2 | 90.6 | 92.5 | 93.6 | 91.9 | 91.6 | 92.8 | 92.1 | 92.2 | 92.0 | 92.0 | 91.6 | 91.5 | 92.3 | 91.6 | 92.0 | 92.4 | 91.6 | 92.0 | 91.8 | 90.8 | 100 | | | | | |
| tam | 92.0 | 91.0 | 91.6 | 91.1 | 90.7 | 91.6 | 92.3 | 91.3 | 90.8 | 91.5 | 91.6 | 92.4 | 91.6 | 91.7 | 91.5 | 91.5 | 91.2 | 91.4 | 92.1 | 92.3 | 91.2 | 91.8 | 91.1 | 91.3 | 91.6 | 100 | | | | |
| tel | 91.7 | 91.5 | 91.4 | 90.9 | 90.6 | 91.4 | 92.2 | 91.1 | 90.9 | 91.7 | 91.4 | 92.1 | 92.0 | 91.4 | 91.6 | 91.8 | 91.5 | 91.0 | 92.2 | 92.5 | 91.6 | 91.6 | 91.1 | 92.0 | 91.4 | 92.3 | 100 | | | |
| tha | 92.3 | 90.8 | 91.6 | 91.2 | 90.6 | 91.6 | 91.9 | 91.6 | 91.2 | 91.9 | 91.7 | 91.6 | 91.7 | 91.4 | 91.6 | 91.1 | 92.2 | 91.3 | 91.2 | 91.4 | 91.6 | 92.0 | 91.6 | 90.3 | 92.2 | 91.3 | 91.0 | 100 | | |
| tur | 92.3 | 91.0 | 92.3 | 92.0 | 90.7 | 92.3 | 93.0 | 92.1 | 91.9 | 92.5 | 92.3 | 92.4 | 91.9 | 92.2 | 92.0 | 92.0 | 91.4 | 91.8 | 92.4 | 92.3 | 92.3 | 92.6 | 92.2 | 90.8 | 92.0 | 91.9 | 91.8 | 91.8 | 100 | |
| yue | 90.5 | 89.7 | 90.4 | 89.9 | 91.1 | 90.1 | 90.3 | 90.2 | 89.8 | 90.4 | 90.0 | 90.8 | 90.0 | 90.0 | 90.4 | 90.3 | 90.4 | 90.1 | 90.3 | 90.3 | 90.3 | 90.4 | 89.9 | 89.6 | 90.3 | 90.3 | 90.1 | 90.2 | 90.5 | 100 |

**Figure A.19.: SONAR Cross-Lingual & Intra-Text Similarity Analysis Results 2.** With the results of the last encoder layer and of the final language-agnostic embeddings, as the mean pooling does not change the similarities.

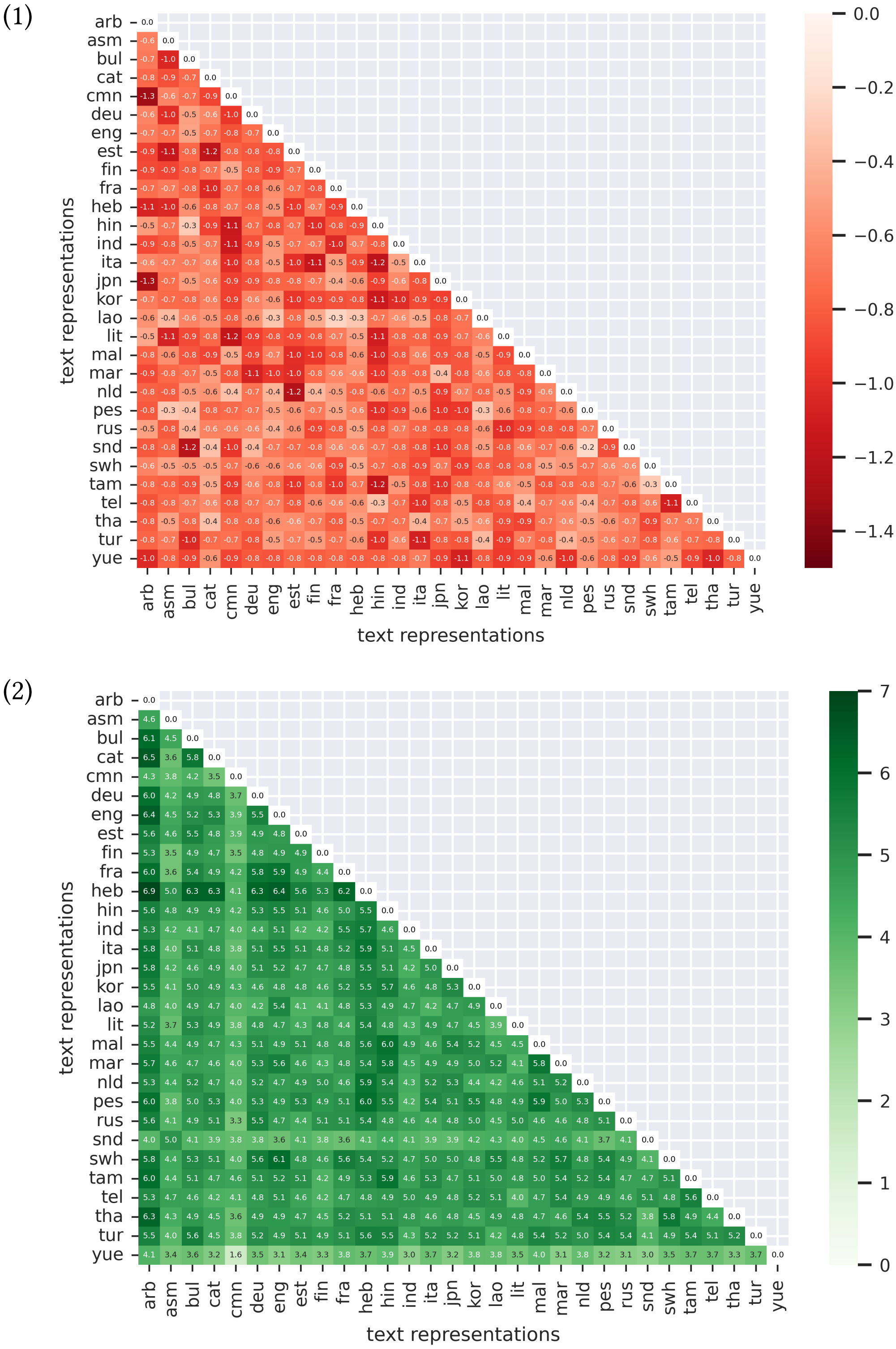


**Figure A.20.: SONAR Cross-Lingual & Intra-Text Similarity Differences.** With the differences (1) between the input embeddings and the first encoder layer and (2) between the first and last encoder layer.

### A.3.7. SALMONN - Cross-Modal Similarities

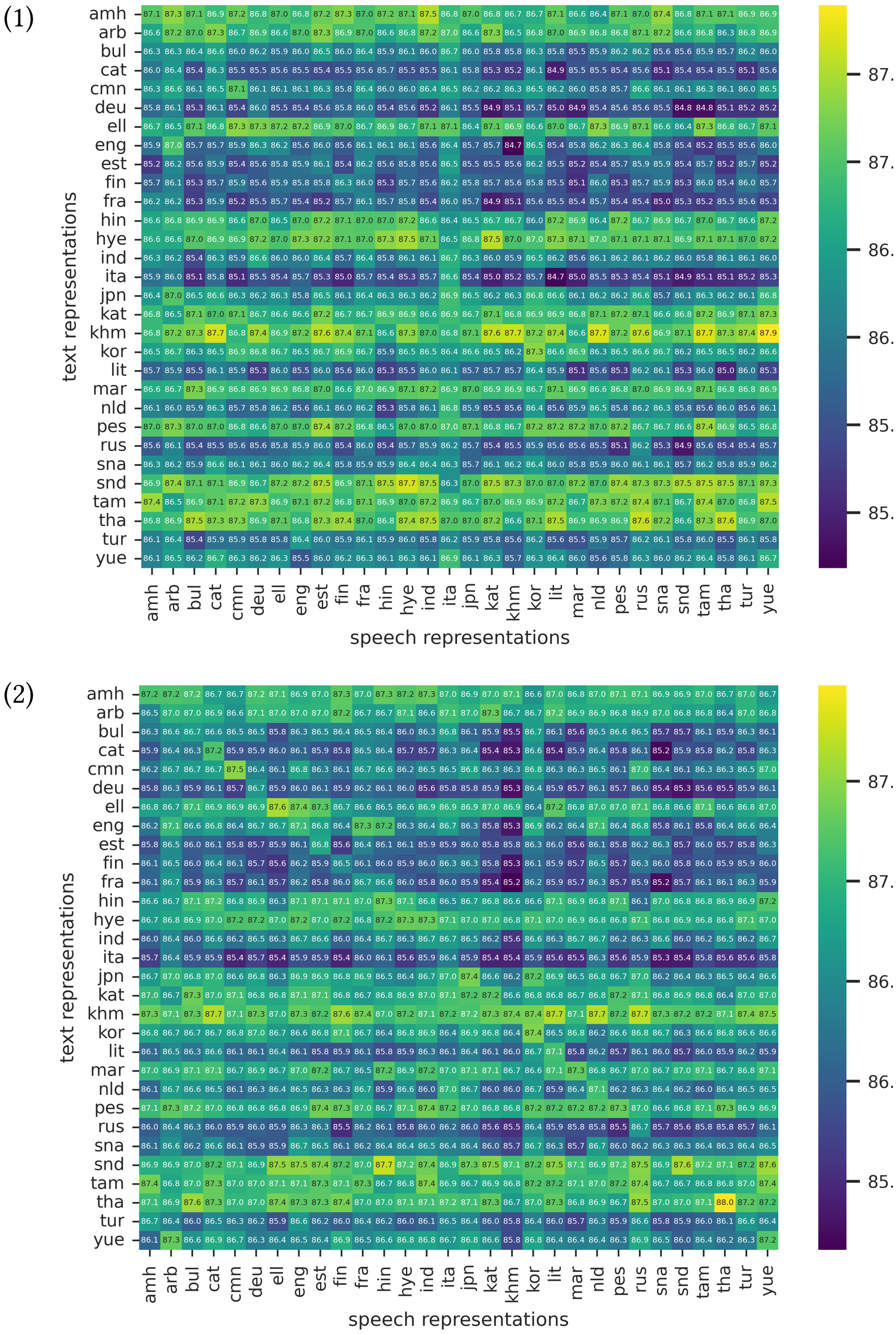


**Figure A.21.: SALMONN Cross-Lingual & Cross-Modal Similarity Analysis Results 1.** With the results of (1) the encoder outputs before and (2) after the Q-Former. The text embeddings are the same for both heatmaps.

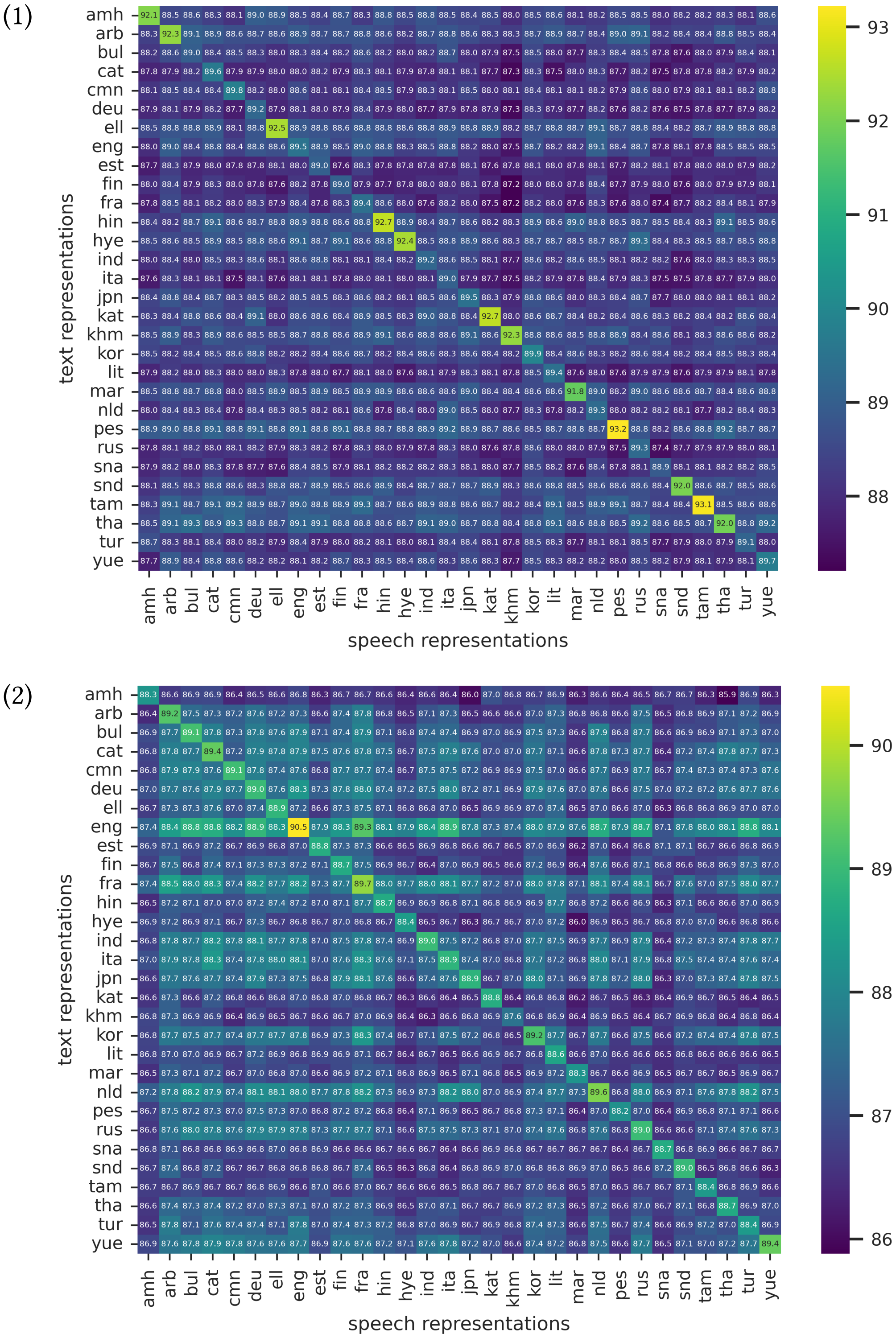
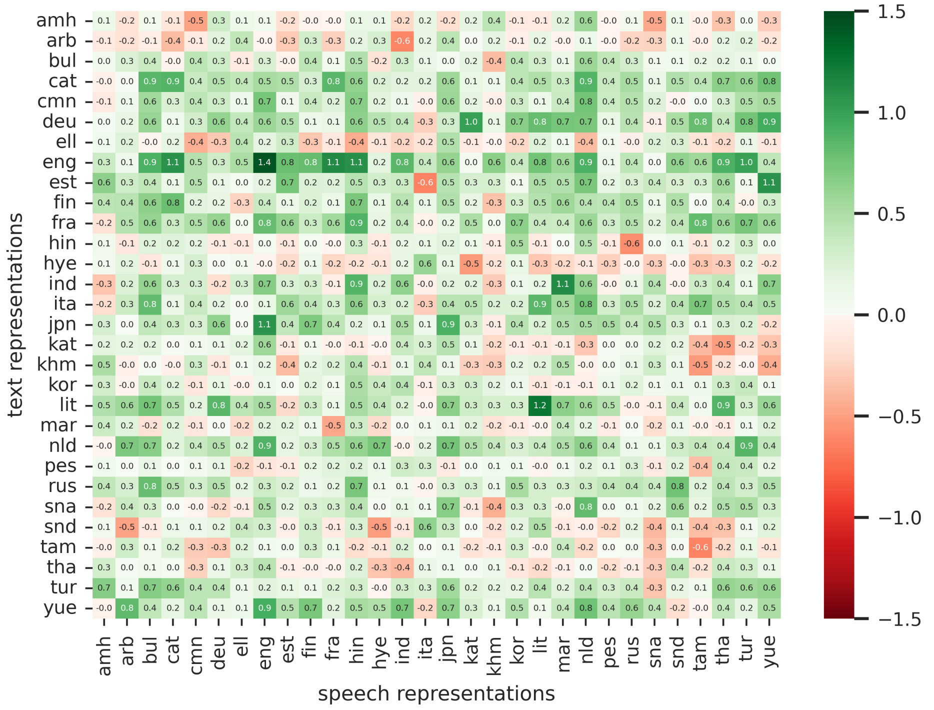


**Figure A.22.: SALMONN Cross-Lingual & Cross-Modal Similarity Analysis Results 2.** With the results of (1) the first and (2) the last decoder layer.

(1)



(2)

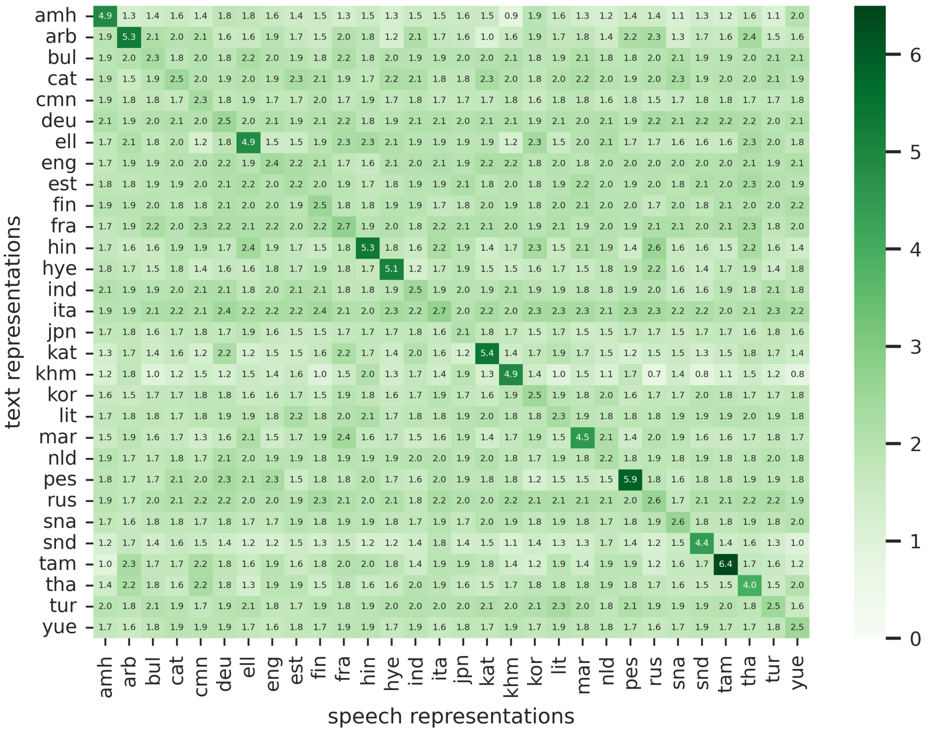


**Figure A.23.: SALMONN Cross-Lingual & Cross-Modal Similarity Differences 1.** With the differences (1) between the embeddings before and after the Q-Former and (2) between the input embeddings after the Q-Former and first decoder layer.

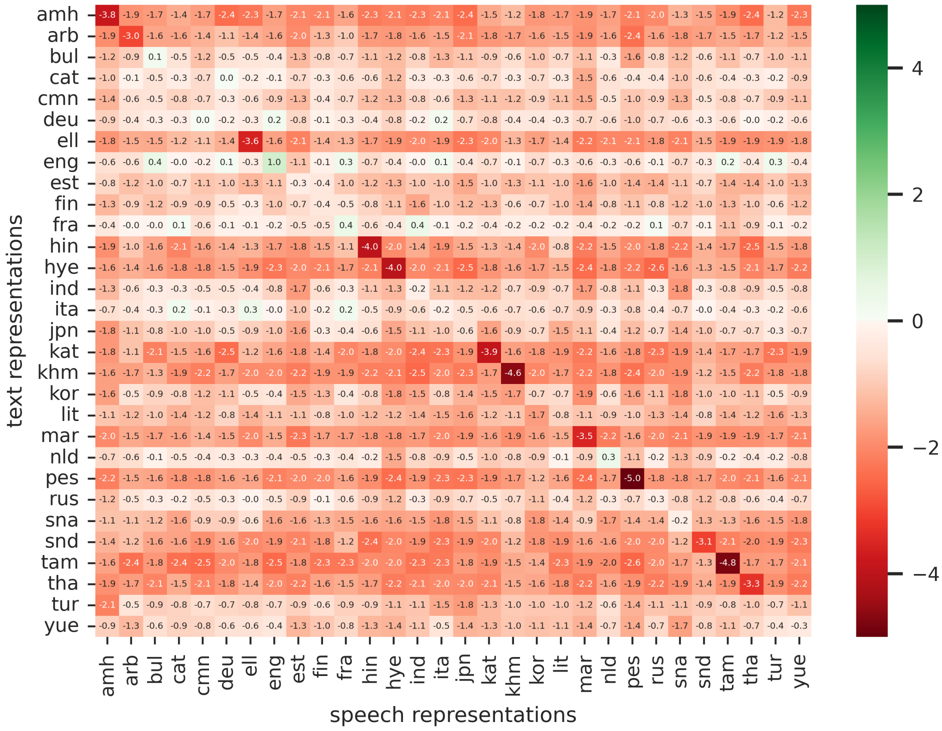


**Figure A.24.: SALMONN Cross-Lingual & Cross-Modal Similarity Differences 2.** With the differences between the first and last decoder layer.

### A.3.8. SALMONN - Intra-Speech Similarities

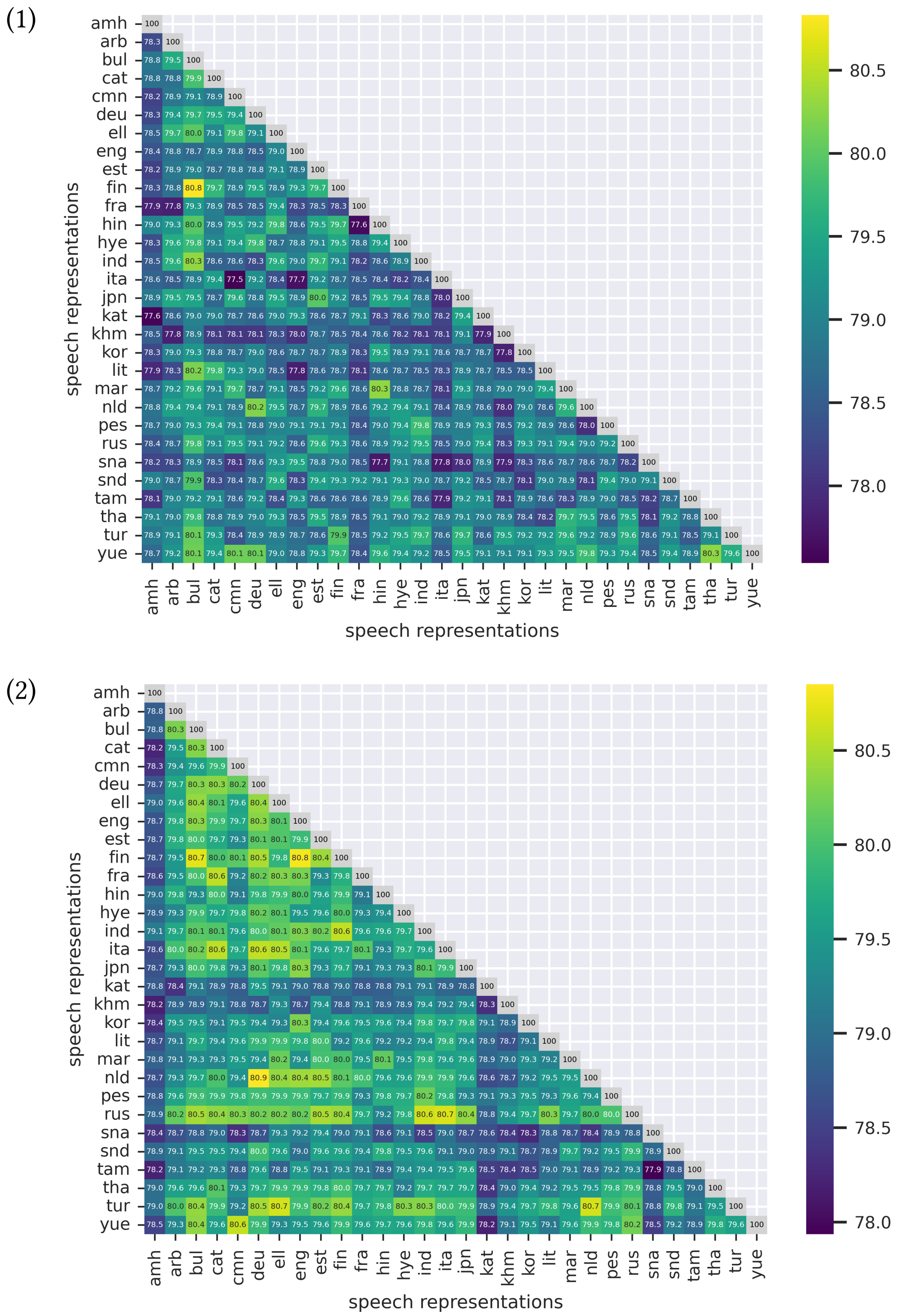


**Figure A.25.: SALMONN Cross-Lingual & Intra-Speech Similarity Analysis Results 1.** With the results of (1) the encoder outputs before and (2) after the Q-Former.

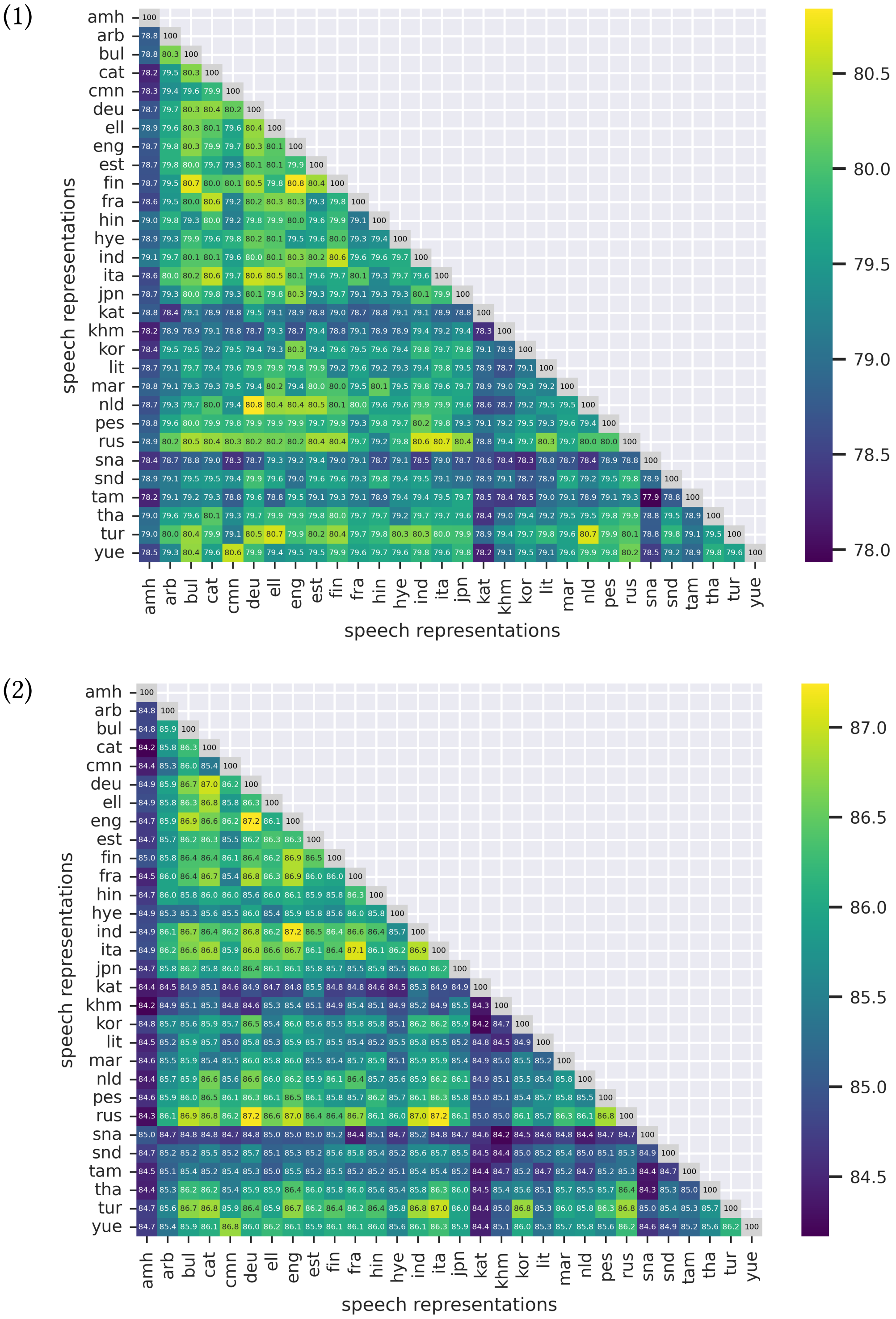


**Figure A.26.: SALMONN Cross-Lingual & Intra-Speech Similarity Analysis Results 2.** With the results of (1) the first and (2) the last decoder layer.

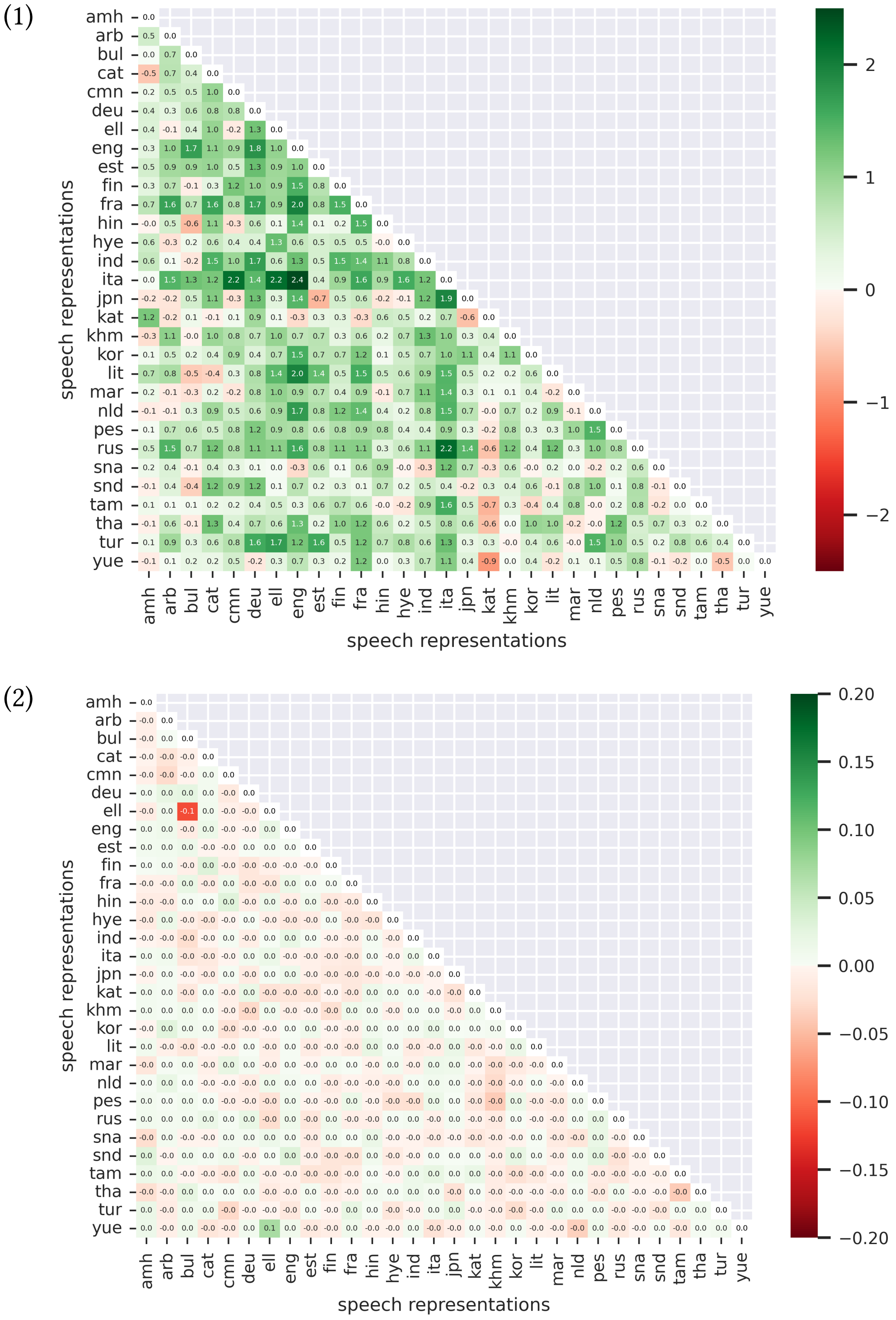


**Figure A.27.: SALMONN Cross-Lingual & Intra-Speech Similarity Differences 1.** With the differences (1) between the encoder embeddings before and after the Q-Former and (2) between the input auditory embeddings after the Q-Former and first decoder layer.

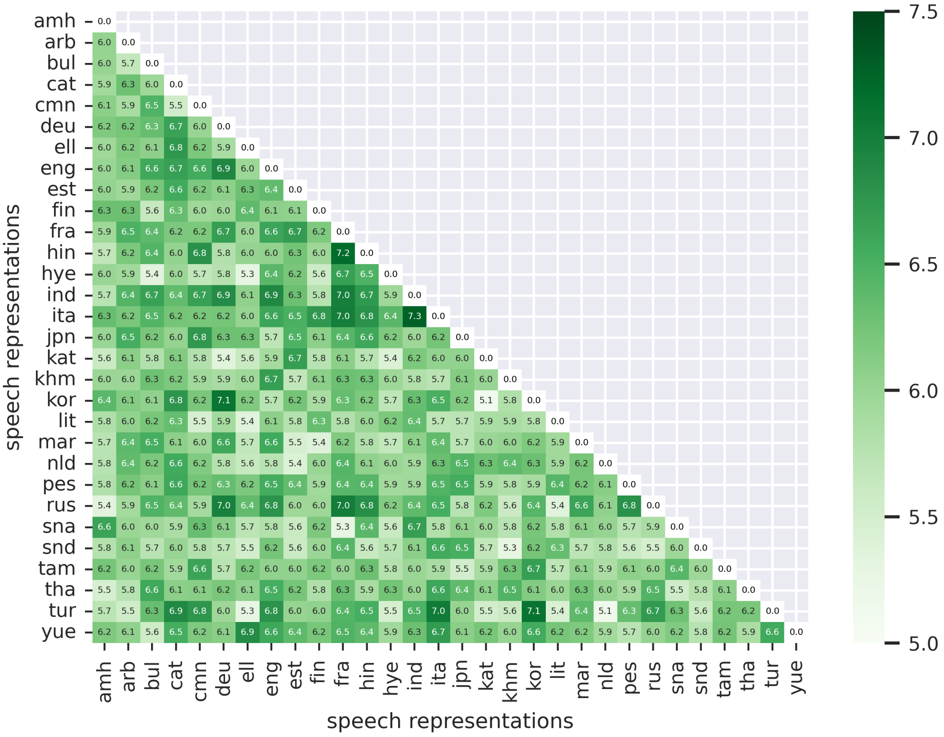


**Figure A.28.: SALMONN Cross-Lingual Speech Similarity Differences 2.** With the differences between the first and last decoder layer.

### A.3.9. SALMONN - Intra-Text Similarities

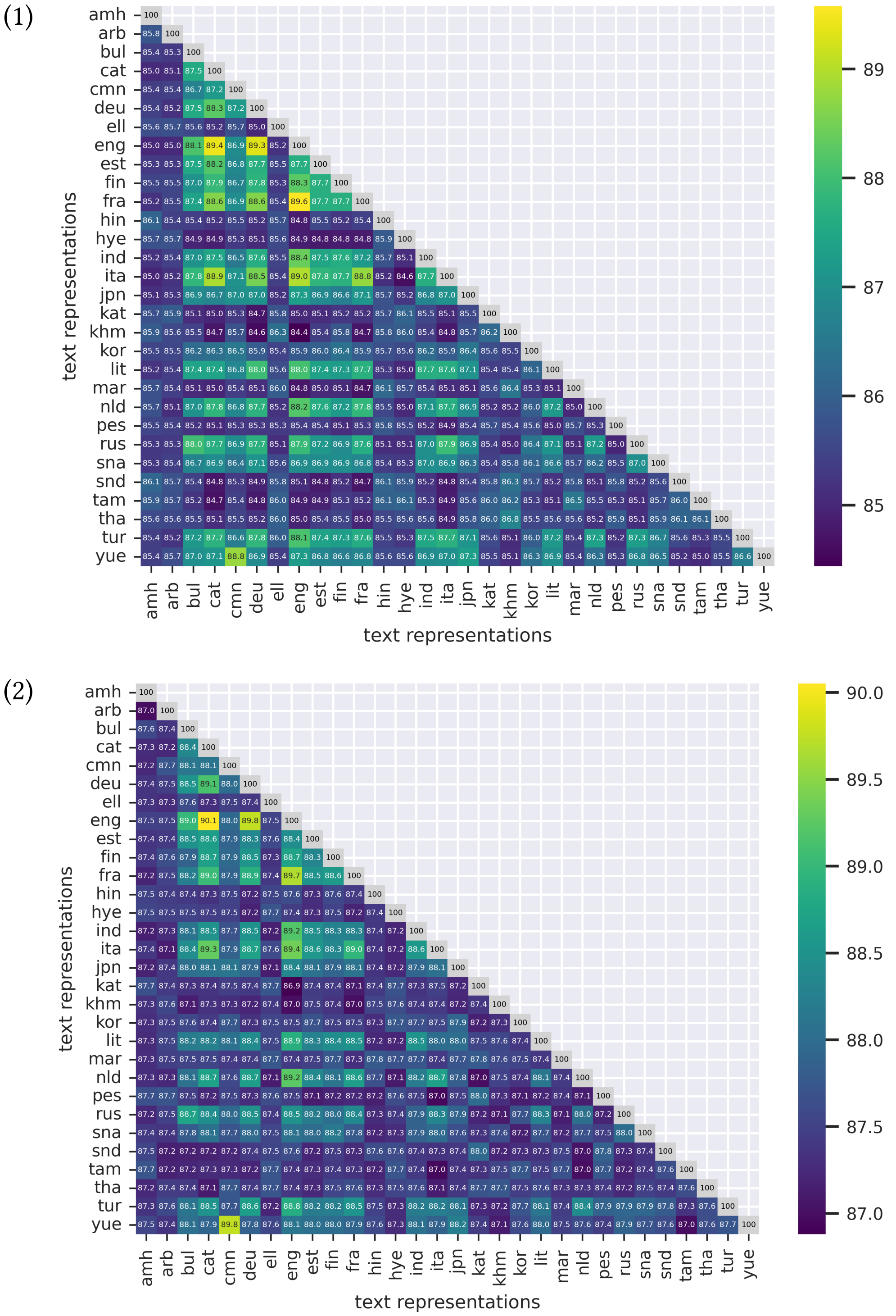


**Figure A.29.: SALMONN Cross-Lingual & Intra-Text Similarity Analysis Results 1.** With the results of (1) the input text embeddings and (2) the first decoder layer.

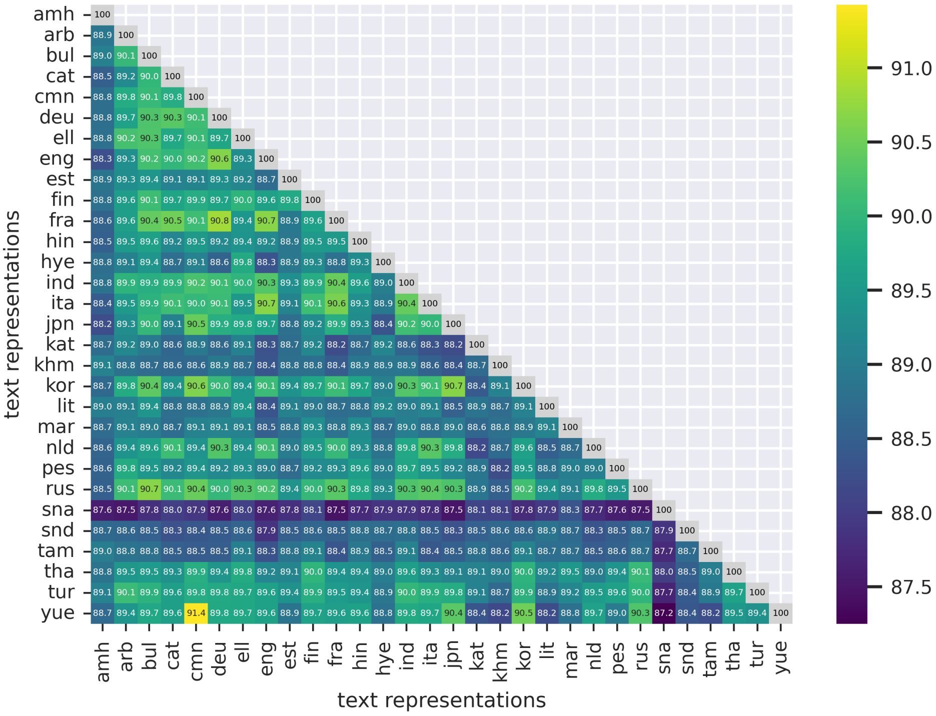


**Figure A.30.: SALMONN Cross-Lingual & Intra-Text Similarity Analysis Results 2.** With the results of the last decoder layer.

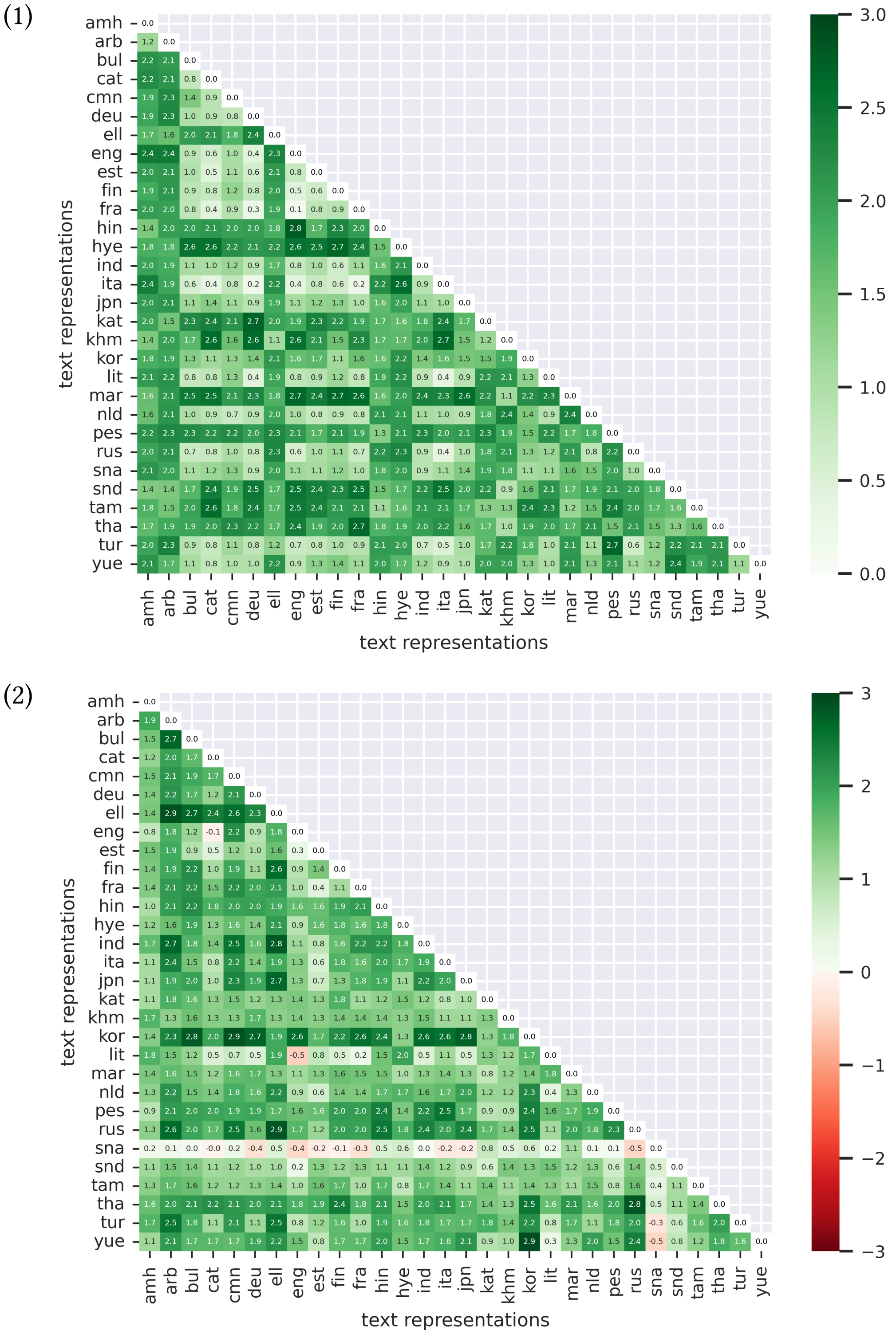


**Figure A.31.: SALMONN Cross-Lingual & Intra-Text Similarity Differences.** With the differences (1) between the input text embeddings and the first decoder layer and (2) between the first and last decoder layer.

# A.4. T-SNE Results

## A.4.1. SeamlessM4T

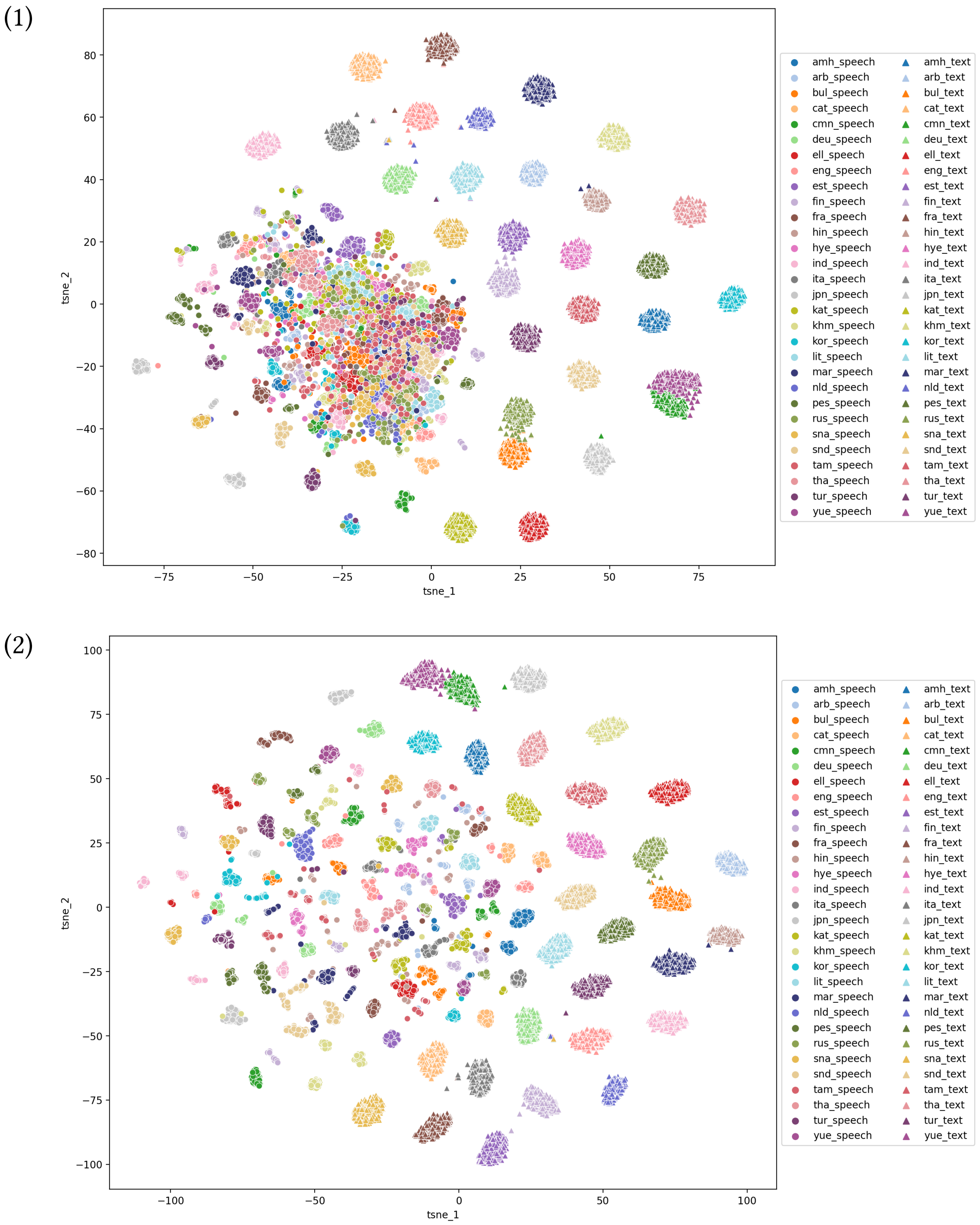


**Figure A.32.: SeamlessM4T t-SNE Results 1.** With the results of (1) the input embeddings and (2) the first encoder layer.

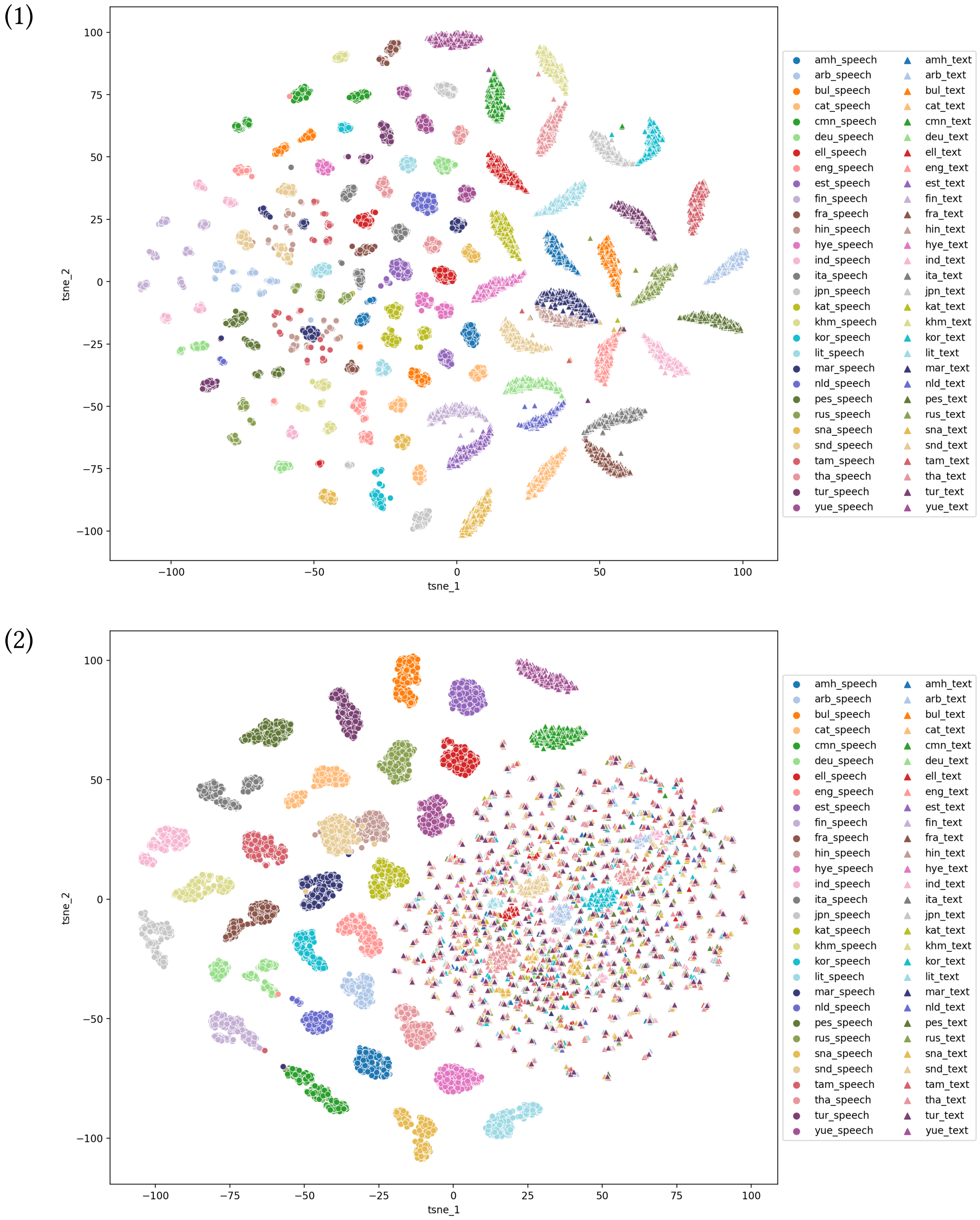


**Figure A.33.: SeamlessM4T t-SNE Results 2.** With the results of (1) the fourth and (2) the tenth encoder layer.

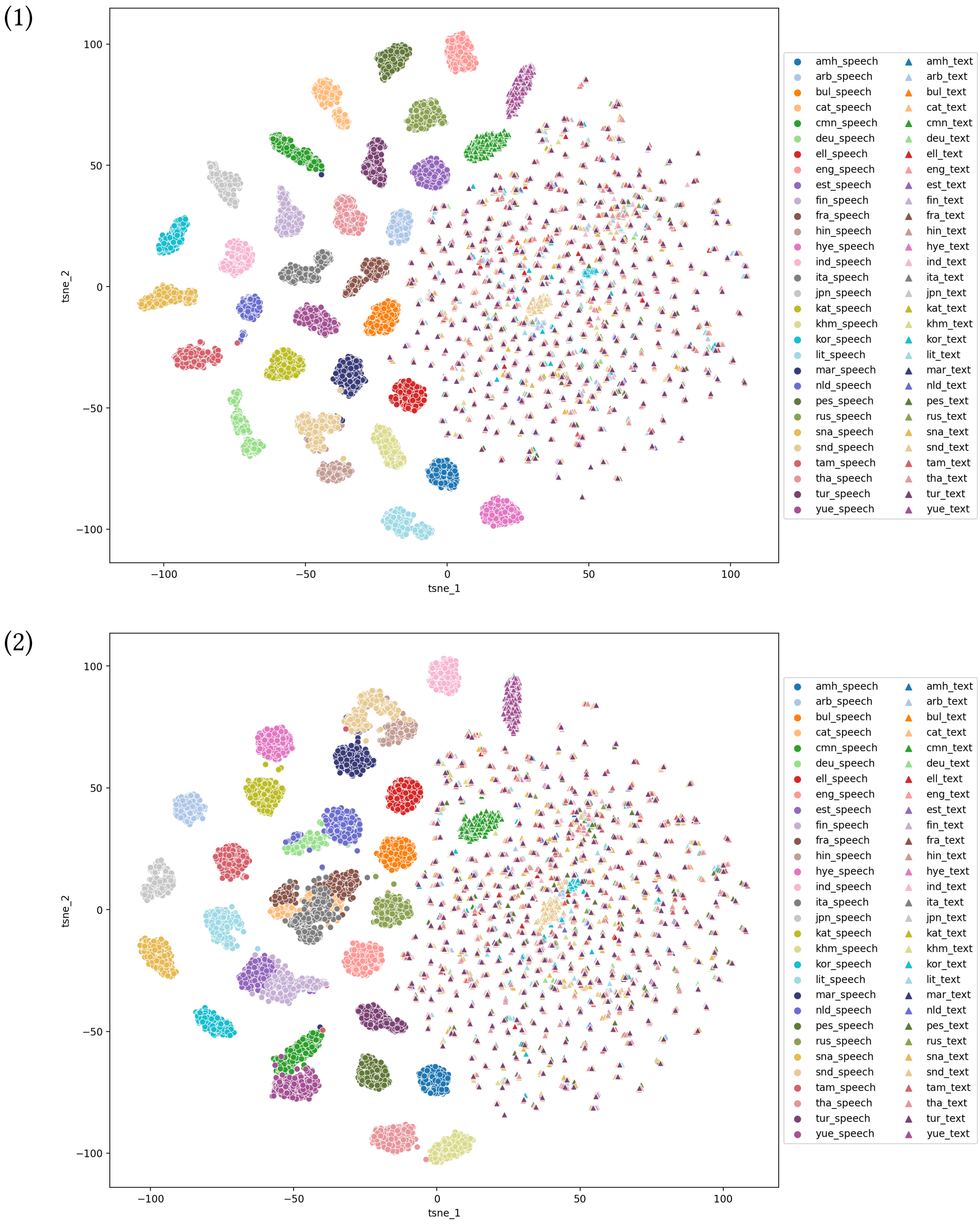


**Figure A.34.: SeamlessM4T t-SNE Results 3.** With the results of (1) the 14th and (2) the 18th encoder layer.

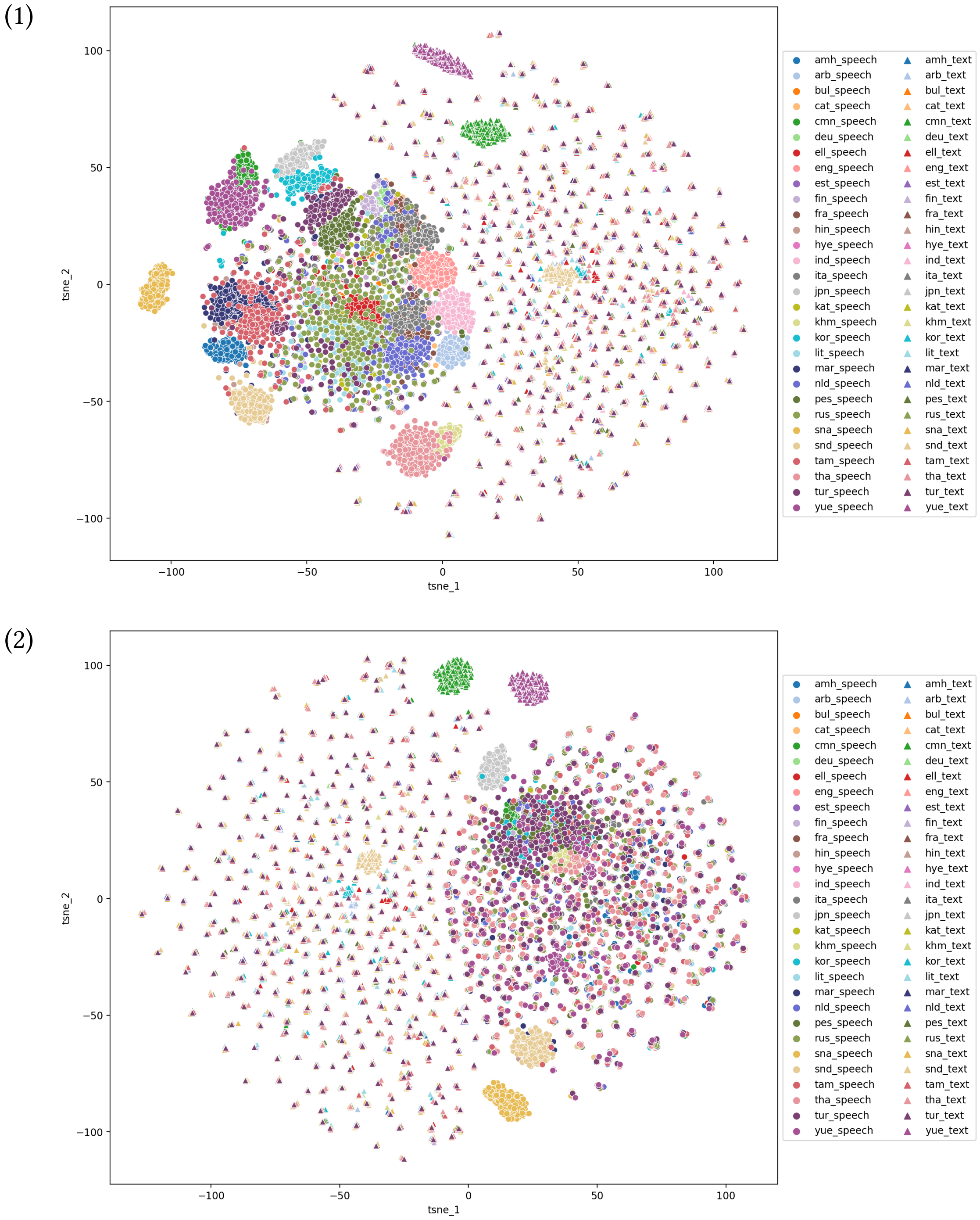


**Figure A.35.: SeamlessM4T t-SNE Results 4.** With the results of (1) the 22th and (2) the 24th encoder layer.

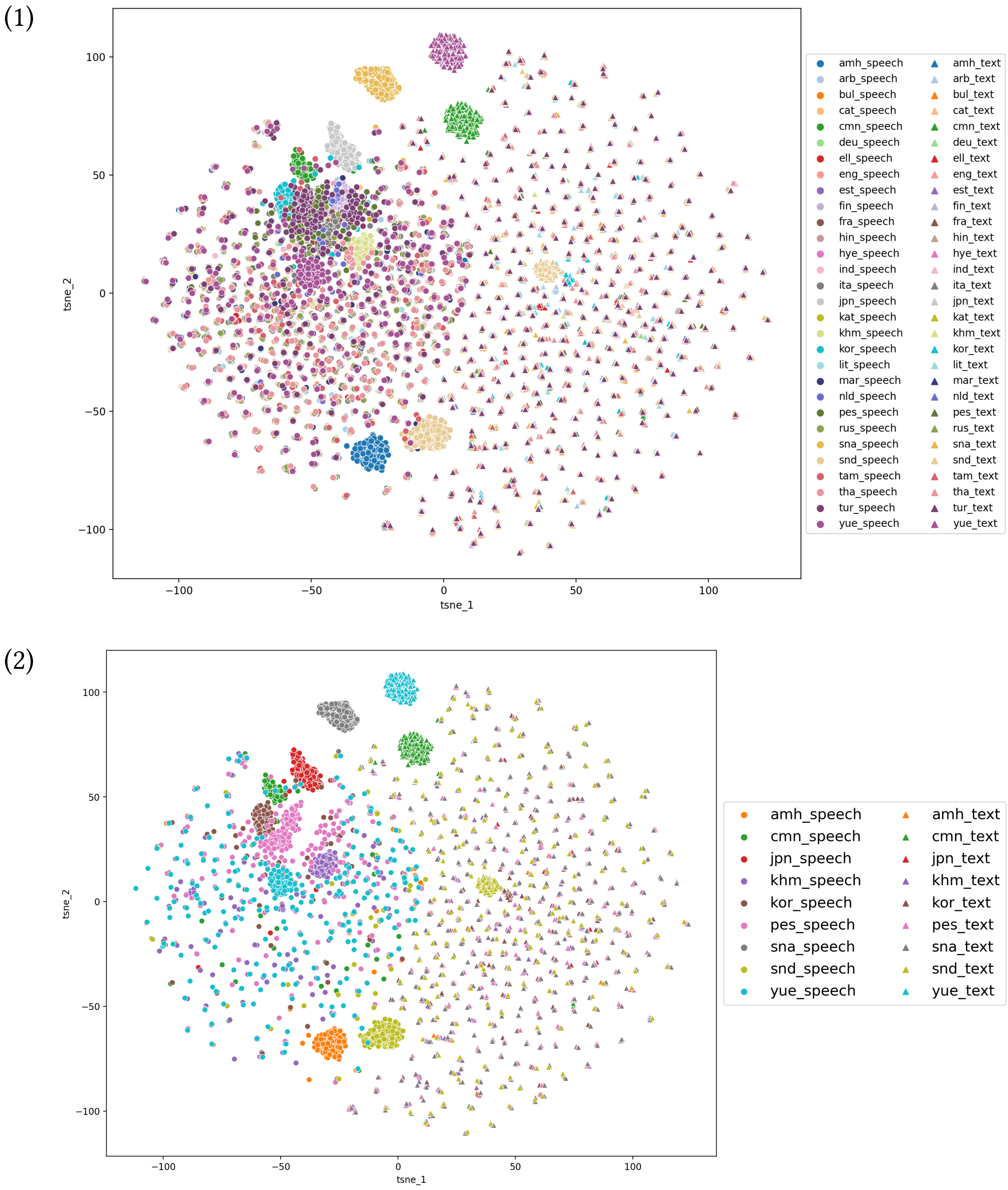


**Figure A.36.: SeamlessM4T t-SNE Results 5.** With the results of (1) after the length adaptor and (2) after the length adaptor with only the clusters visible.

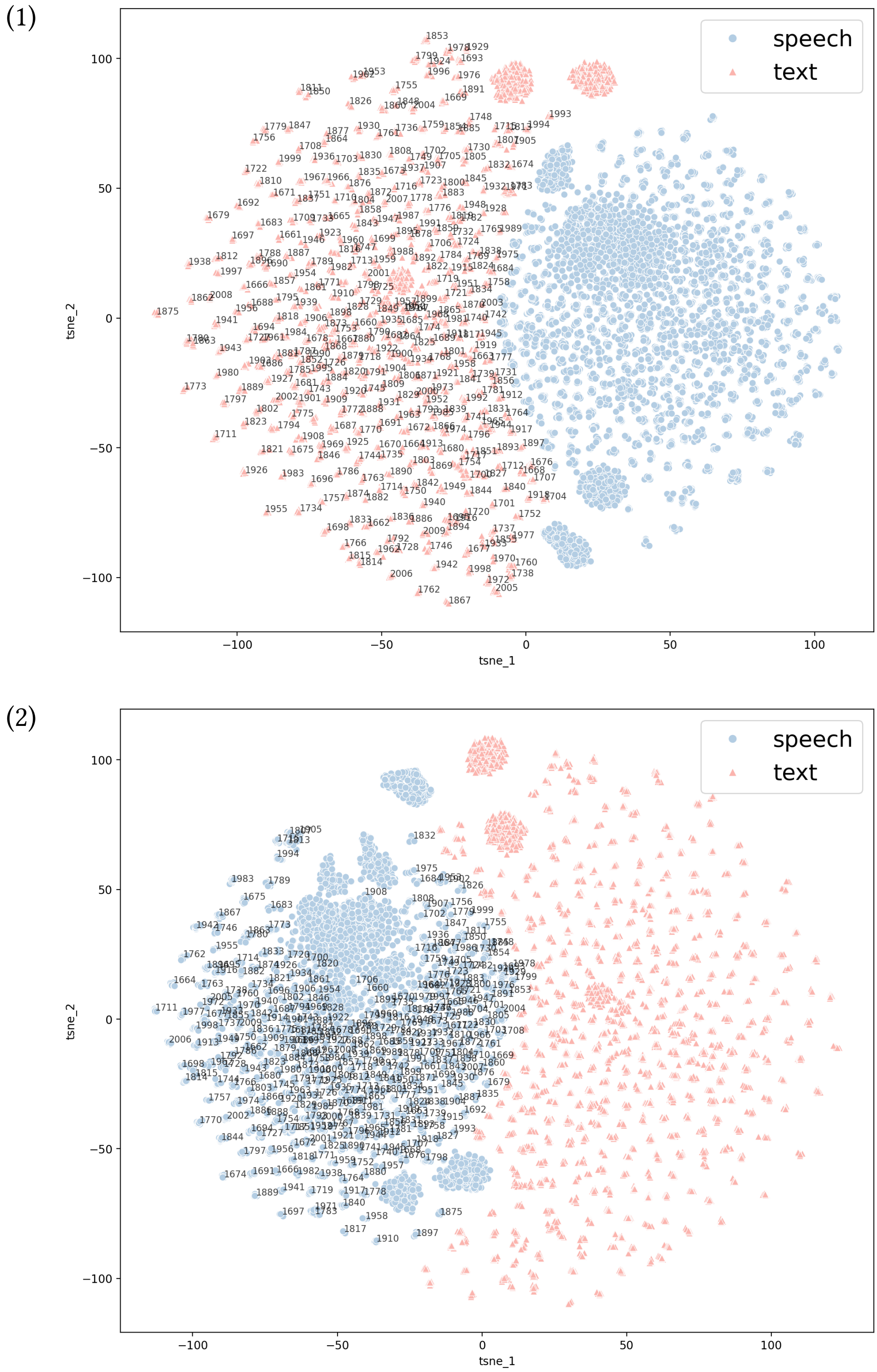


**Figure A.37.: SeamlessM4T t-SNE Results 6.** With the results of (1) last encoder layer and (2) after the length adaptor with only the ids of the input sentences visible. Sentences with the same semantic meaning have the same id in FLEURS across all languages, meaning that the representations of each bundle is based on the same semantic meaning.

### A.4.2. SONAR

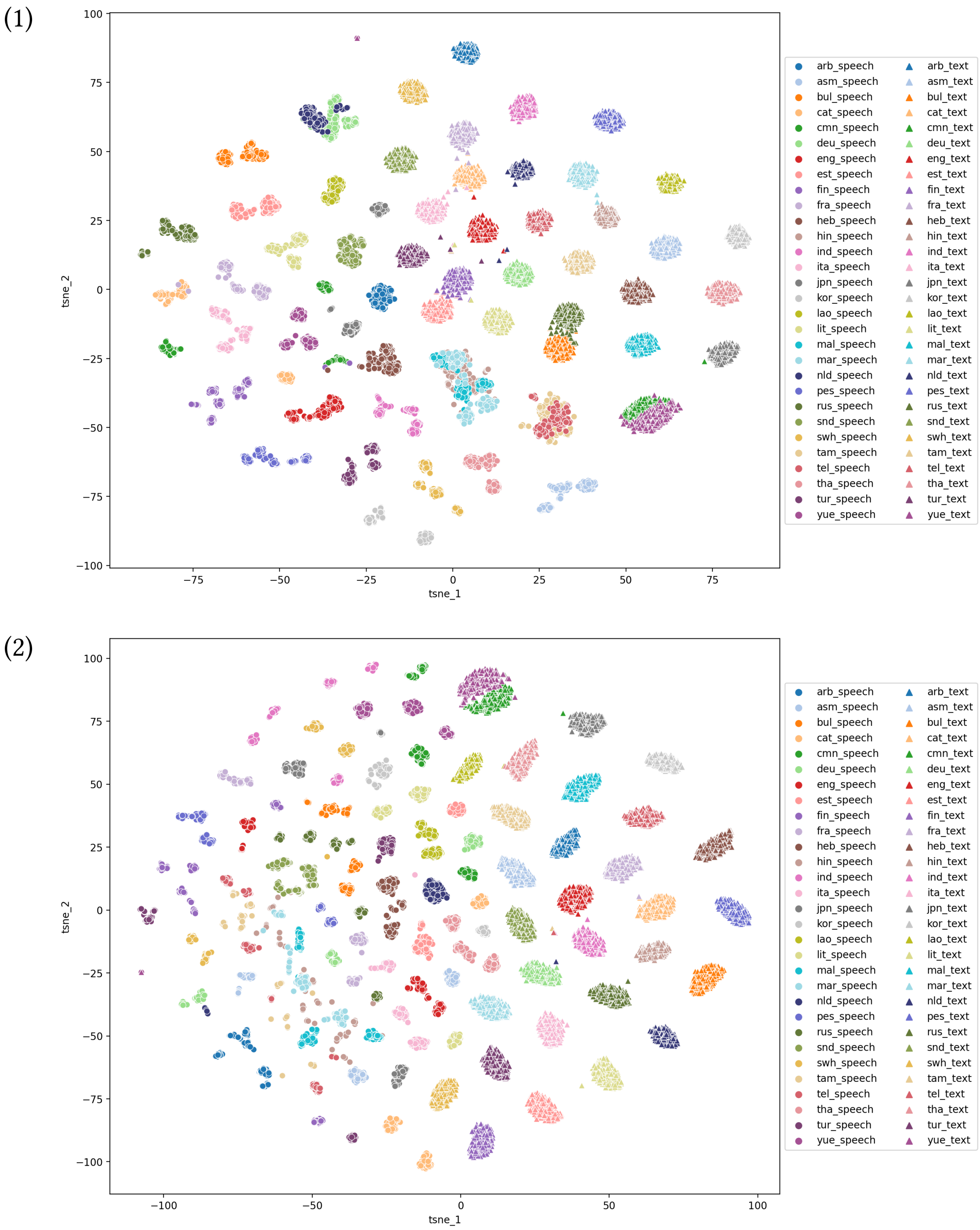


**Figure A.38.: SONAR t-SNE Results 1.** With the results of (1) the input embeddings and (2) the first encoder layer.

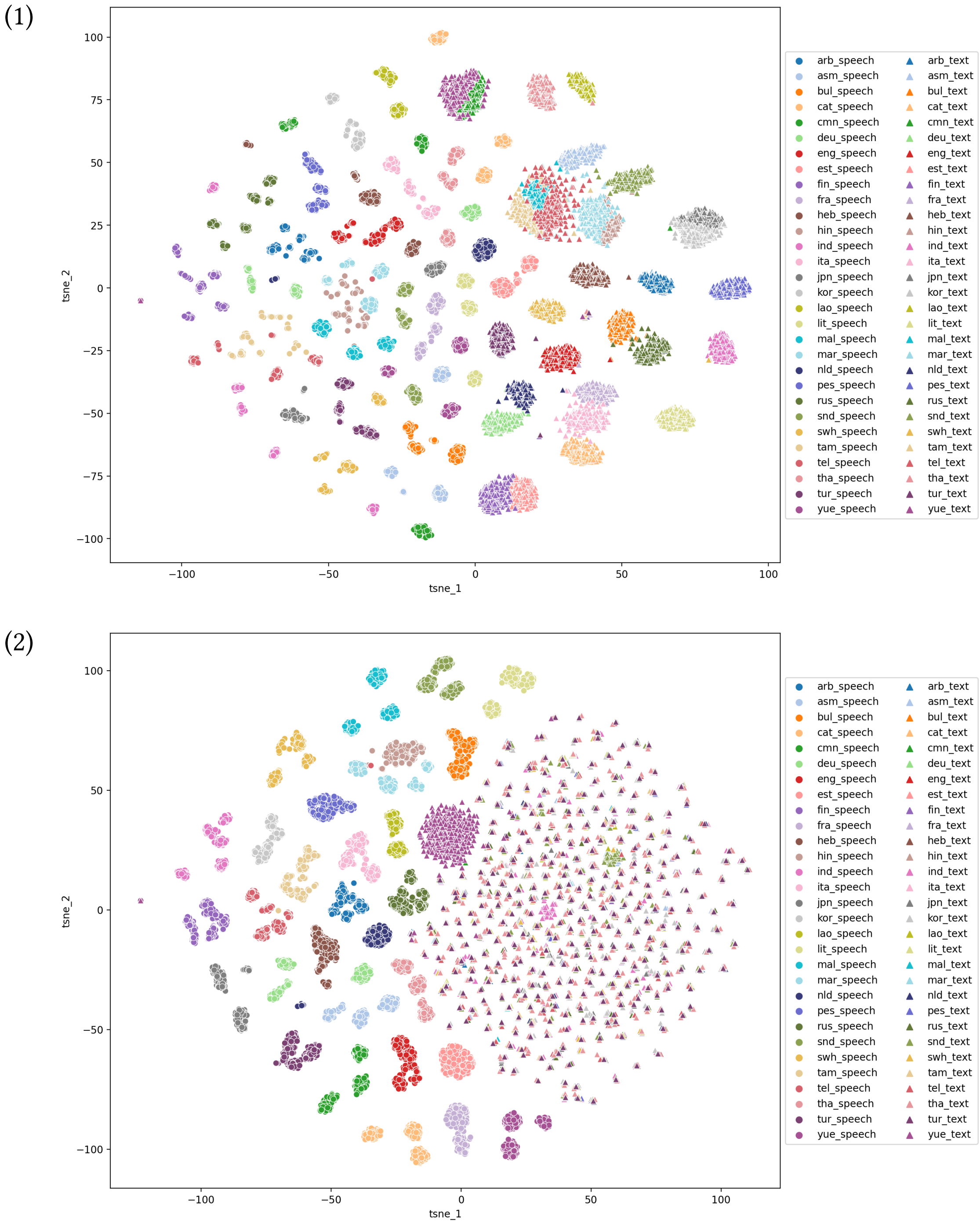


**Figure A.39.: SONAR t-SNE Results 2.** With the results of (1) the sixth and (2) the tenth encoder layer.

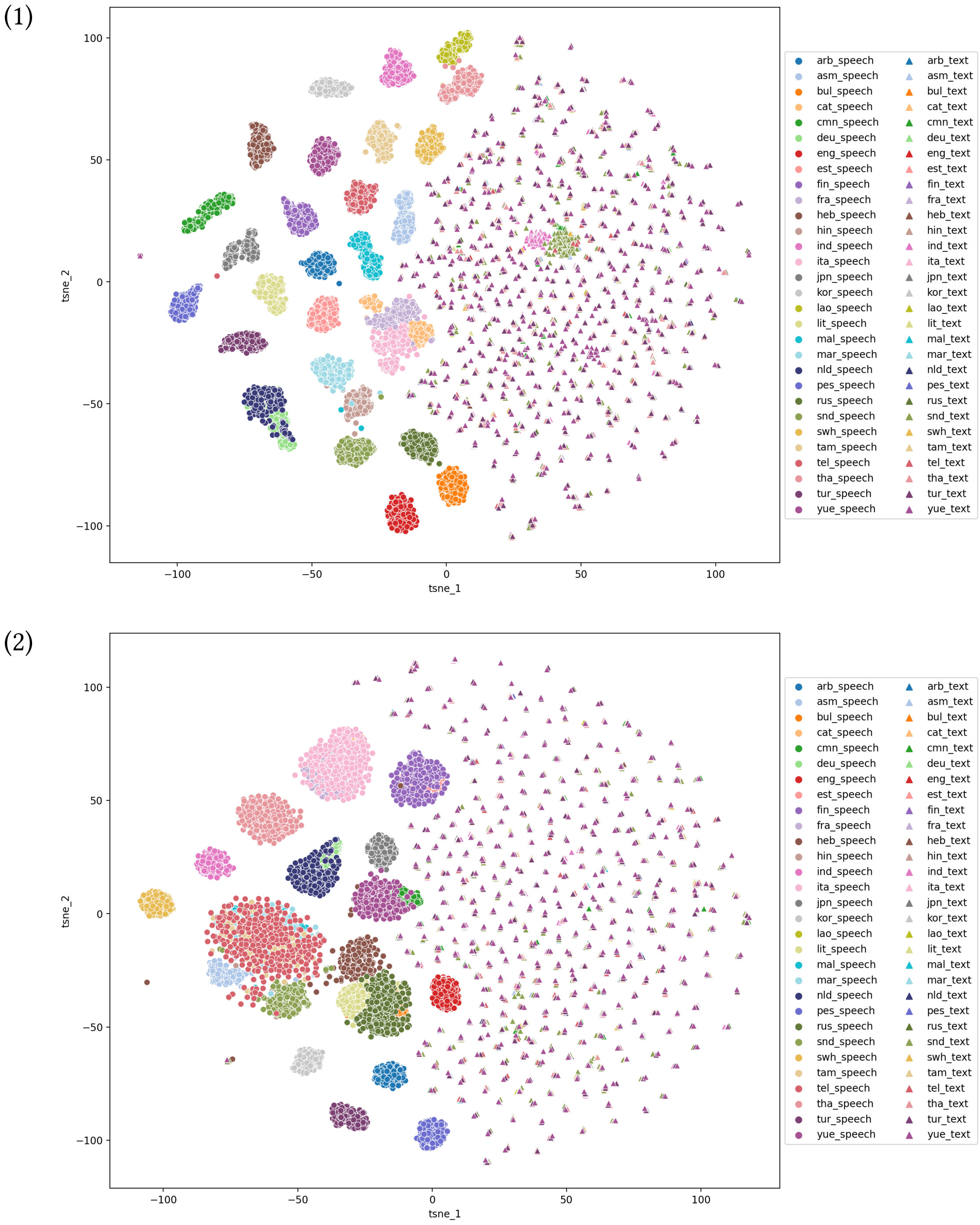


**Figure A.40.: SONAR t-SNE Results 3.** With the results of (1) the 14th and (2) the 22nd encoder layer.

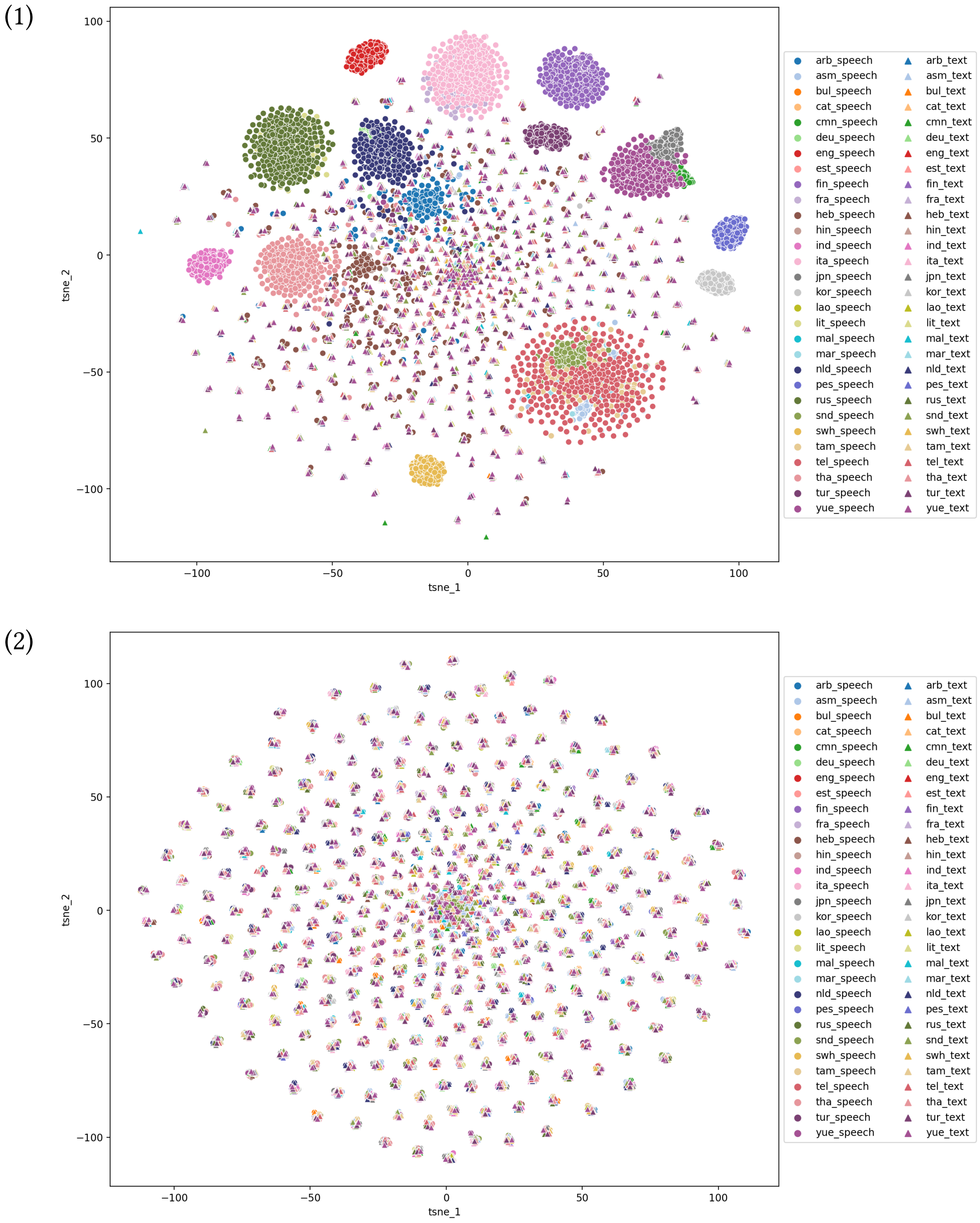


**Figure A.41.: SONAR t-SNE Results 4.** With the results of (1) the 24th and (2) the final language-agnostic embeddings.

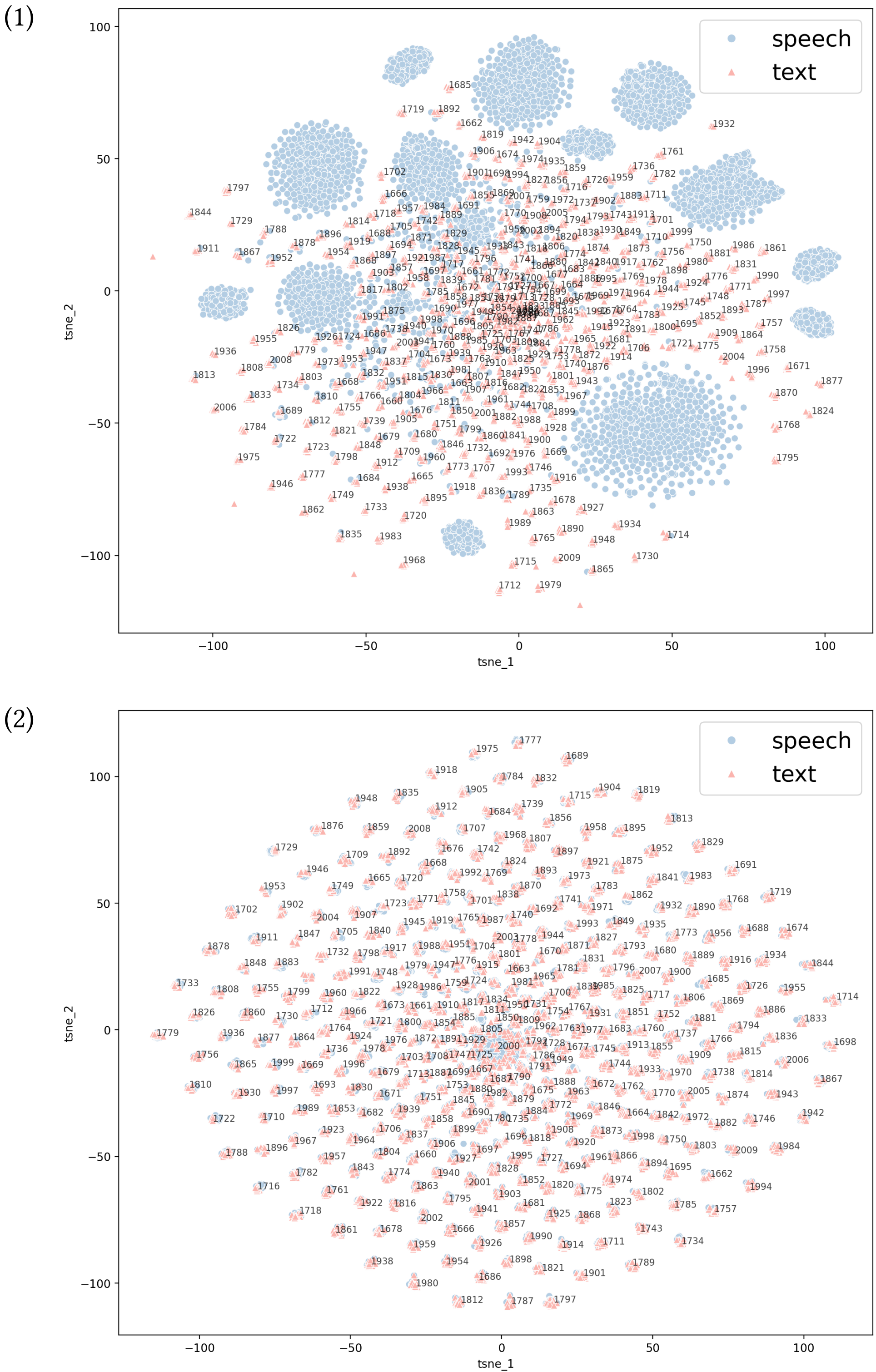


**Figure A.42.: SONAR t-SNE Results 5.** With the results of (1) the 14th encoder layer and (2) the final language-agnostic embeddings with only the ids of the input sentences visible. Sentences with the same semantic meaning have the same id in FLEURS across all languages, meaning that the representations of each bundle is based on the same semantic meaning.

### A.4.3. SALMONN

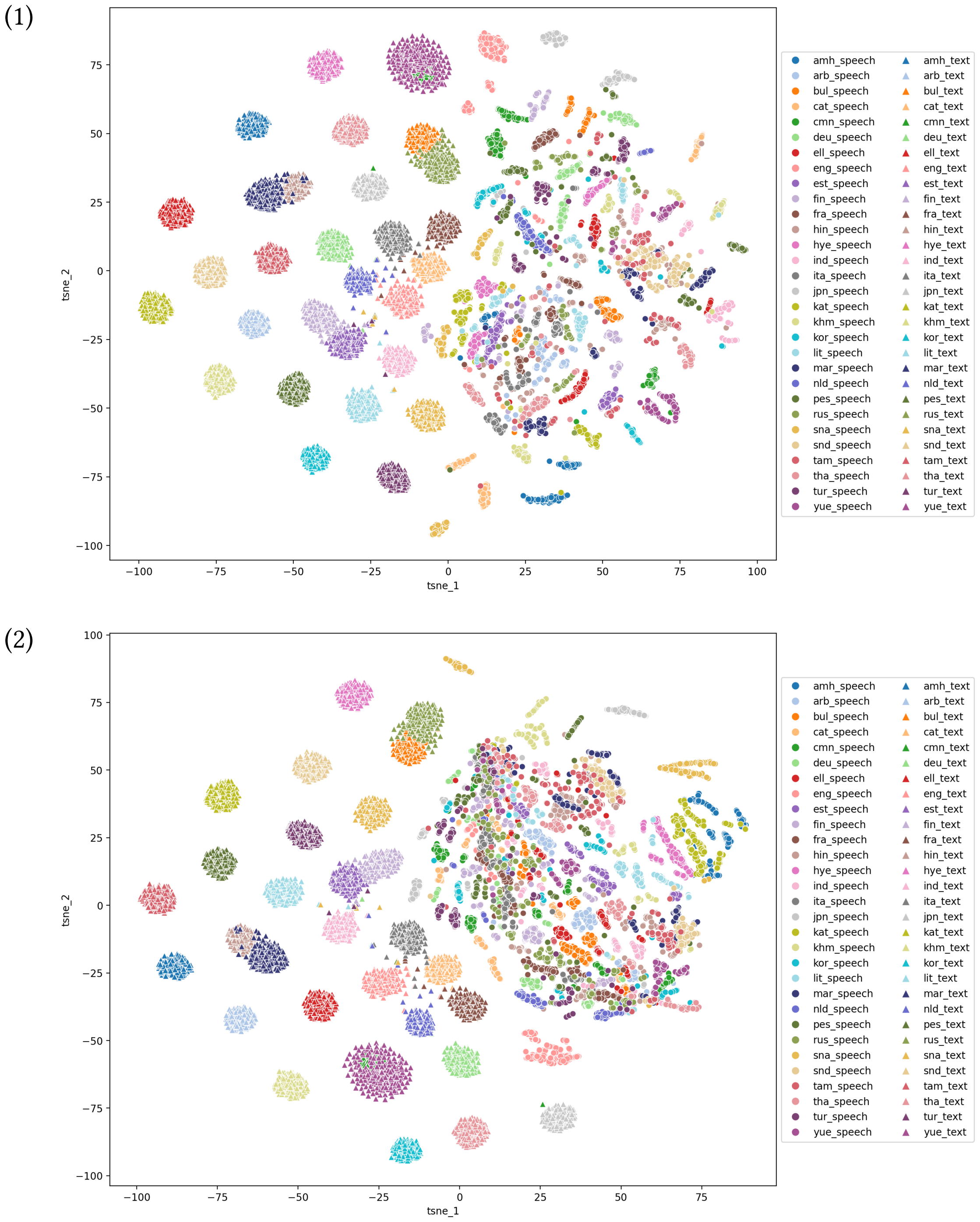


**Figure A.43.: SALMONN t-SNE Results 1.** With the results of (1) the encoder outputs before and (2) after the Q-Former. The text embeddings are the same for both t-SNE maps.

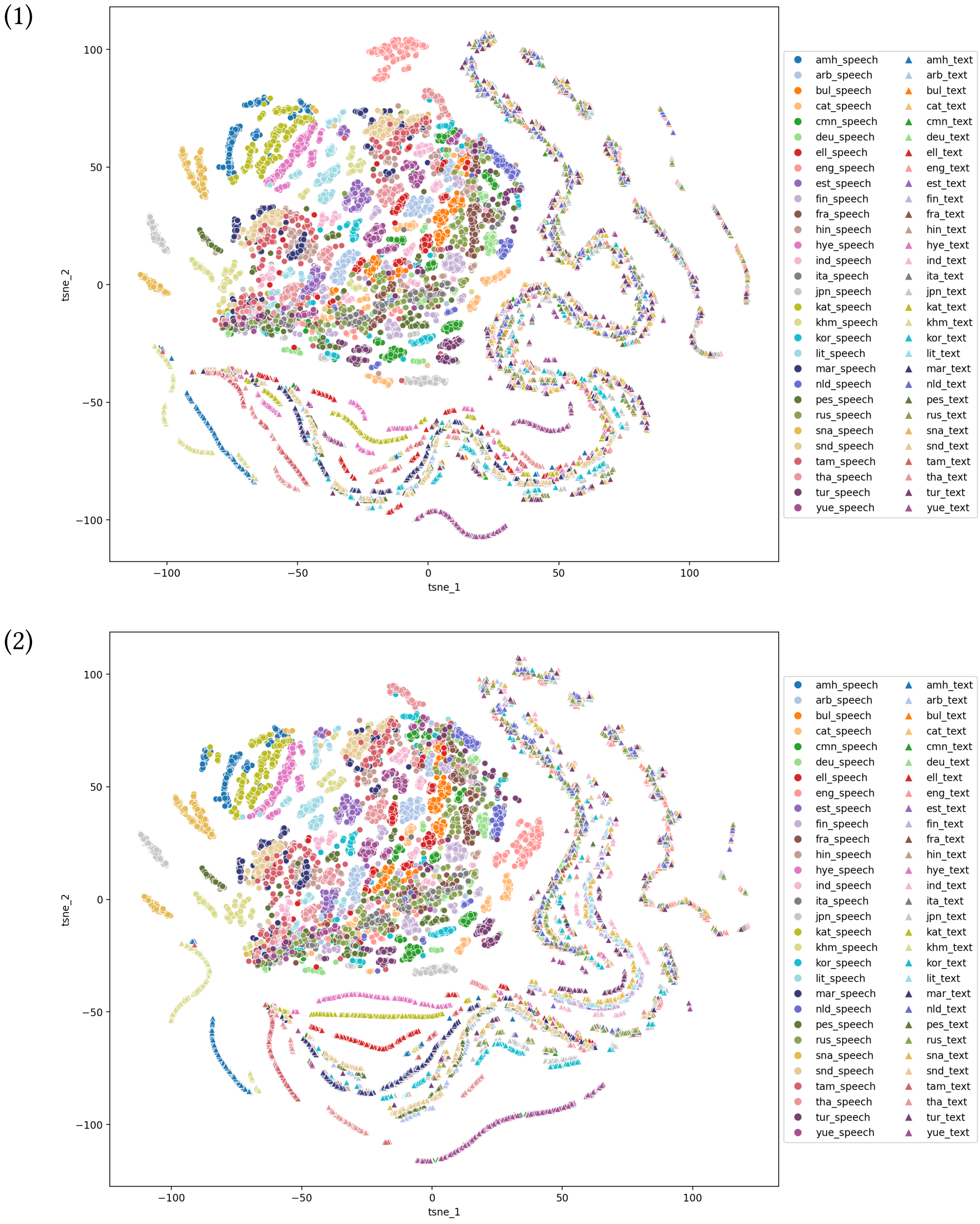


**Figure A.44.: SALMONN t-SNE Results 2.** With the results of (1) the second and (2) fourth decoder layer.

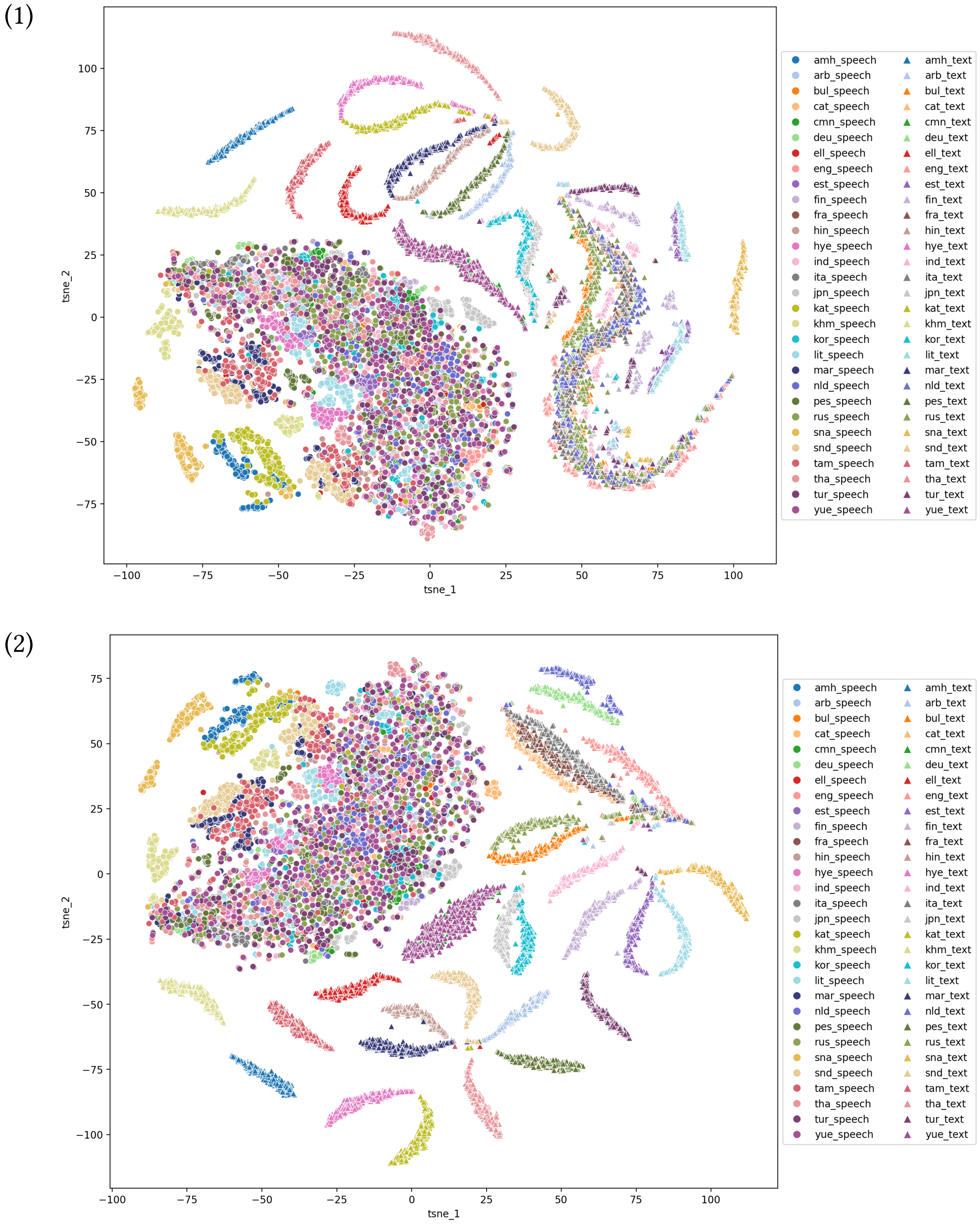


**Figure A.45.: SALMONNN t-SNE Results 3.** With the results of (1) the 12th and (2) 18th decoder layer.

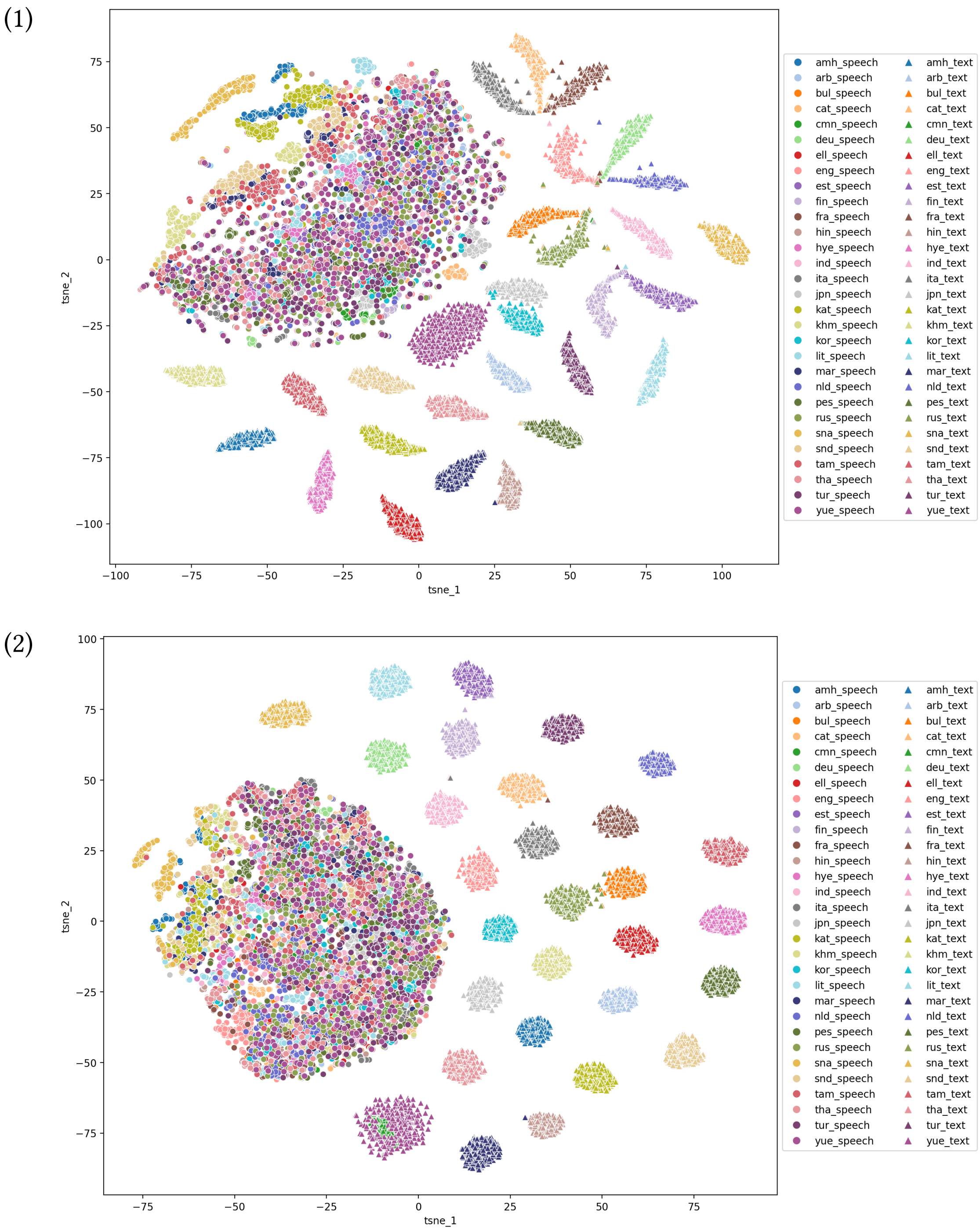


**Figure A.46.: SALMONN t-SNE Results 4.** With the results of (1) the 24th and (2) 32nd decoder layer.