Universiteit
Leiden
The Netherlands

Data Science & Artificial Intelligence

Phonetic and semantic alignment of spoken language:

Conceptual understanding of Languages without Text

Serdar Bayraktar

Supervisors:
Dr. Rob Saunders & Dr. Zhengjun Yue

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/07/2025

Abstract

Have you ever wondered can languages understand us from the speech. Humans can understand in each other while they speak, understand pauses, continuation, and imagine concepts in different languages in the same objects. Traditional speech processing systems typically rely on a text-based intermediate step, converting speech to text for analysis before any further action. This pipeline, however, is a departure from the way humans naturally process language, as it discards rich paralinguistic information and introduces latency. This thesis investigates the feasibility of achieving a direct, conceptual understanding between languages from spoken audio alone, bypassing the need for text.

The study leverages dubbed audio content in English, Turkish, and Latin American Spanish. Using Wav2Vec for audio representations, sentence-level text embeddings for semantic comparison, and specialized phonetic word embeddings, this work explores the alignment of spoken language on both semantic and phonetic levels. The analysis employs cosine similarity to quantify relationships and t-SNE for visualizing the structure of the embedding space.

While the approach shows potential for developing more natural language processing systems, it also underscores the complexity of spoken language and the need for further refinement in models to disentangle content from speaker-specific features and achieve the robustness of human perception.

Contents

1 Introduction **1**
- 1.1 Research Questions . . . 1

2 Background / Related Work **2**
- 2.1 Audio Embeddings and Wav2Vec . . . 2
- 2.2 Comparison of Vectors . . . 2
  - 2.2.1 t-SNE . . . 3
- 2.3 Traditional Flow and Speech Understanding . . . 3
- 2.4 Dynamic Endpointing . . . 5
- 2.5 Phonetic and Semantic Understanding . . . 5
- 2.6 Open data and lack of Multi-lingual Data . . . 5

3 Methods **6**
- 3.1 Misalignments . . . 6
- 3.2 Targeting Same Content with Dub . . . 6
- 3.3 Embeddings . . . 7
  - 3.3.1 Comparison of Vectors . . . 7
- 3.4 Phonetic Word Embeddings . . . 7
  - 3.4.1 Similar Words . . . 7
  - 3.4.2 Alignment Issues . . . 7
  - 3.4.3 Small Words . . . 8
- 3.5 Multilingual Base . . . 8

4 Experiments and Results **9**
- 4.1 Phonetic and Semantic Embedding Comparisons . . . 9
- 4.2 Multilingual base similarities . . . 9
- 4.3 English phonetic audio and word similarity on shorter segments . . . 9

5 Discussions **17**

6 Conclusions and Further Research **18**

7 Acknowledgments **18**

References **21**

8 Appendix **22**

1 Introduction

Vectorization of data is a critical step in enabling interpretation and processing by Large Language Models (LLMs). Modern vectorization techniques extend beyond textual data to include various modalities such as audio and images, making multi-modal vector representations increasingly important. Recent state-of-the-art models often rely on such multimodal vectorization approaches to enhance their internal architectures and capabilities.

While vectorization plays a foundational role, it is equally important that representations of the same semantic content, even when expressed in different data types, yield similar vectors. This ensures consistency and interpretability across modalities. A well-known example of this concept is the analogy operation in word embeddings, such as "King - Man + Woman" approximates "Queen" [20], which illustrates how semantic relationships can be encoded in vector space.

In recent years, LLMs have seen rapid growth and have already started to expand to Vision Language Models (VLMs), large multi-modal models [7, 17]. These models aim to process data types other than text [16]. One of the data types data seen lots of interest is audio [17]. There are few drawbacks of text compared to audio such as a lack of emotion expression, intuitiveness, prosody etc [31, 30].Although these features enrich communication, they introduce significant challenges. The same sentence can be spoken in multiple ways by different speakers—or even by the same speaker at different times—resulting in variability that must be accounted for during processing. This presents significant challenges for models intended to process audio recordings of speech. Ideally, audio embeddings of semantically identical content should cluster closely, while emotionally or contextually distinct expressions of the same text should diverge appropriately. In this context, vectorization quality and similarity analysis across speech samples become crucial. In this domain, vectorization results becomes an important topic.

Language also plays a key role in representation. In text vectorization, also known as text embeddings, the same words in different languages should result in similar vectorization after multilingual text embedding process [26]. This particularly similar approach can be used for audio. To express it fully, similarity analysis of same content in different languages based on speech data.

The remainder of this thesis is structured as follows: Section 2 includes background information and related work; Section 3 discusses the methods and nature of the data used; Section 4 describes the experiments and discusses their outcome; and Section 6 presents some conclusions and possible future work.

1.1 Research Questions

The main research question addressed in this thesis is:

> **RQ1** Can conceptual understanding between languages be created in multilingual landscape using phonetic embeddings?

This research question can be divided into two sub-questions:

> **RQ1.1** Can phonetic embeddings create similarity between languages?

RQ1.2 Can phonetic word embeddings create similarity between audio embeddings?

2 Background / Related Work

Embeddings is the process of converting input data into vector representations, also referred to as 'tokens' in the literature. Several different approaches have been developed to train/optimize this embedding process [19, 23, 6]. There are different focuses on embeddings such as text, sentence, speaker, etc. Embeddings models have some specific capabilities that change per model basis. Such as their multilingual support. Sentence or text embeddings typically embed semantic meanings. But there is also phonetic word embeddings which leverages phonemes of the words and encoding them [24]. For the experiments `sentence-transformers/LaBSE` [6] and `sentence-transformers/distiluse-base-multilingual-cased-v2` [21] going to be used for vector conversions. The reason why these are selected depends on two factors, availability in the multilingual base and wide usage in other academic papers in the literature.

2.1 Audio Embeddings and Wav2Vec

Audio embeddings are used to vectorize audio data [1]. There are a few more embedding categories in Audio Embeddings, such as speaker Embeddings, speech Embeddings, and joint Embeddings [25, 5]. Wav2Vec was considered state-of-the-art for audio embeddings [23]. Wav2Vec uses a feature encoder with a frame rate of 50 Hz. This makes every vector that Wav2Vec produces equal to 20 ms of audio. This is important because even 2 character words take more than 20 ms of audio speech data. To process these, there are some ways to continue, e.g., by computing mean embeddings across tensors [28, 11].

2.2 Comparison of Vectors

Several commonly used vector comparison methods are possible, such as cosine similarity and Euclidean distance. It is important to choose the correct similarity method because they differ by what they mean. Cosine similarity focuses on direction of the vector while Euclidian distance focuses on distance between vectors. This choice is critical because an inappropriate metric can yield results that are mathematically sound but conceptually meaningless for the given application. Cosine similarity is a metric for measuring angular similarity between vectors. Focuses on the orientation of the vector.

$$\text{cosine\_similarity}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{1}$$

Here, $\mathbf{x}$ is the first vector and $\mathbf{y}$ is the second vector. $\cdot$ represents dot product of the vectors. $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$ represent Euclidean norms (magnitudes).

Euclidean distance is a geometric measurement of a vector distance. It corresponds as a straight line between 2 vectors. Euclidean distance is a higher dimension version of the Pythagorean theorem. The results represent hypotenuse of n-dimensional vectors. Euclidean distance focuses on the

distance, unlike cosine similarity.

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

Here, $\mathbf{x}$ is the first vector and $\mathbf{y}$ is the second vector. x_i and y_i represent the value of the i-th dimension in the vectors $\mathbf{x}$ and $\mathbf{y}$, respectively.

2.2.1 t-SNE

t-distibuted Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction technique used for visualization of high-dimensional data [**?**]. It focuses on pairwise similarities between vectors in high-dimensional space. It aims to create similar similarities in the lower-dimensional space, a two-dimensional "map". It is able to capture non-linear relationships. In the context of our study of homonyms, applying t-SNE is a good visualization. It allows us to render a visual representation of the embedding space, enabling a direct, qualitative assessment of whether the model has learned to place different contextual uses of a homonym into separate, meaningful clusters. Although it is good, since it simplifies multidimensional information, it is not the only technique used in theses.

2.3 Traditional Flow and Speech Understanding

Traditional speech engines use a pipeline to create speech-to-speech engines. This pipeline uses voice activity detection, speech-to-text, large language model, and text-to-speech. Although this flow is complete, it does not reflect how humans communicate. It also has latency issues along with loss of accuracy in the pipeline due to the use of multiple models [22].

Humans understand speech conceptually. We do not convert it to text in our brain and think about our response. We just understand how it is spoken. This can be imagined as when the words "chair" have been said, we understand chair as an object, not as a word. This process can also be performed in multilingual landscape. Our understanding of the word does not change when the same word is heard in different languages. In all languages, we understand chair as object no matter which language it is. In this point, semantic meaning is also important where meaning of the word is understood from context. Similar semantic needs will also be mentioned in Section 2.4.

These issues have been recognized by the literature and some solutions have been proposed. One of them is Step-Audio [8]. Step-Audio, focuses on this issue by training a unified speech-text model. The power of this model comes from its architecture with both semantic and linguistic tokenizer in LLM. The linguistic tokenizer leverages phonemic and linguistic features.

Another example of new speech systems is Baichuan-Audio by Tianpeng et al. [13]. Baichuan-Audio proposes a unified framework for audio understanding for both Chinese and English. The model can accept both audio and text input as human. For audio understanding, it does not convert text to speech like a traditional flow. Interprets speech directly from the audio itself with its audio embedding structure. Baichuan-Audio have been trained with the two-stage strategy as in Chen et al. [3]. First, while the LLM parameters remain the same, the audio embedding layers are updated. Then, all parameters except the LM embedding and the head-trained. This shows us that when it comes to audio interpretation as humans, multistage training with audio and LLM is needed. This

is also one of the reasons why we try to answer research questions in multiple steps for phonetic and semantic embedding similarities [13].

Another issue with traditional flow is the use of words with similar pronunciation, such as "affect" and "effect". This challenge has been discussed by Li et al. in Chinese [14], also Ma et al. proposed a method using a LLM along with Automatic Speech Recognition (ASR) [18]. The N-best (with values of 1,3,5, and 10) predictions of the ASR and LLM rerank and correct simultaneously. To elaborate more, ASR sends the best N output to the transformer then the transformer converts the best N prediction to one output as shown in Figure 1. This is called Generative Error Correction (GEC). When human speech is investigated, we also understand if the word is "affect" or "effect" from the meaning of the full sentence by thinking of full meaning. The method proposed by Li et al. approximates how humans understand spoken language, study is not a cognitive model but result of the mechanism is close like humans do. The results of the proposed method show a 3 to 5 percent improvement in character error rate and entity recall for end performance. In terms of cosine similarity, the method offers a significant improvement for text and pinyin vectors compared to LLaMa-3-8B-Chinese. This study suggests that understanding conceptual meaning with semantics with mixing audio and semantic embeddings has the potential to improve models of non-English languages [14].

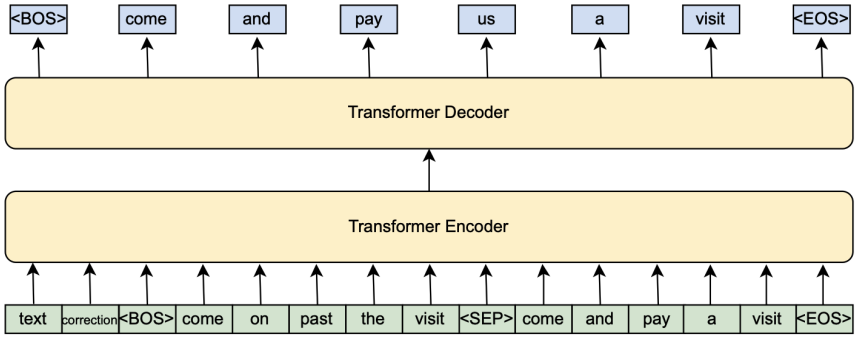

Figure 1: N-best T5 error correction model structure [18]

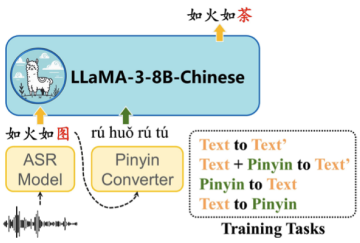

Figure 2: The flowchart for PY-GEC [14]

2.4 Dynamic Endpointing

Another issue with traditional flow is that it does not react directly to what has been spoken. It was created to give some answers and aims to do this. But for some of the sentences we use in daily life, we stop talking and pause with sound [15]. In the meantime, some process goes on. To elaborate on this more, one of the examples is checking weather. Here is the following scenario.

```
Speaker 1 : How is the weather?
Speaker 2 : Let me check emmmm (checks window)
Speaker 2 : It is sunny
```

Here, traditional flow tries to answer between Speaker 2's speech. But by the meaning, its clear that Speaker 1 should wait. So in more general terms, dynamic endpointing understands when to stop and when to continue while recognizing a meaningful action in speech.

2.5 Phonetic and Semantic Understanding

Phonetic and semantic embeddings do not align with each other due to their focus. This creates a challenge of focusing on spoken content. Chen et al. focus on solving this with spoken content retrieval with both phonetic and semantic embeddings [3]. The idea here is not only to retrieve phonetically similar structures, but also to retrieve semantically similar. To elaborate on this more, when "Netherlands cities" is said; it also needs to be able to retrieve the word "Amsterdam". Chen et al. create their architecture with 2 stages, first runs phonetic embeddings with dis-entangling speaker characteristics [3]. This approach emphasizes that what is spoken is more important than who is speaking. In the second stage of the model, semantic embeddings are applied to the phonetic embeddings obtained from the first stage. These resulting embeddings are subsequently aligned or parallelized for evaluation purposes. In this context, parallelization refers to processing audio and text embeddings in a manner that presumes an underlying relationship between the two modalities, even though they may not occupy the same representational space. The evaluation is conducted using the LibriSpeech dataset, and although the results are not flawless, both the retrieval and ranking performances are reported to be reasonable and indicative of meaningful cross-modal alignment.

2.6 Open data and lack of Multi-lingual Data

In the literature most of the models are trained using a few public datasets. These datasets typically come from high-resource languages such as English and Chinese. For low resource languages, limitations increase due to low data [2, 22] and high differences between different language families. To prevent this there are few techniques proposed such as cross-lingual transfer learning [10]. Cross-lingual transfer learning leverages high-resource languages to learn low-resource languages. Another proposition by Banerjee et al. suggests significant improvement with the "each language for itself" and "each language for others" approaches [4]. These are focusing on using high-resource languages to learn understanding speech for low-resource languages.

Libri speech [1], VCTK [32], LJ speech [2], and Common Voice [3] are among the most popular

[1] http://www.openslr.org/12

[2] https://keithito.com/LJ-Speech-Dataset/

[3] https://commonvoice.mozilla.org/en/datasets

open-source datasets. The size of the English portion of datasets varies, but often the datasets are biased and skewed for high-resource languages such as Spanish, French, German, as per Sarim et al. [22, 9, 29]. Another limitation is linguistic bias. Typically, there are more data on some languages, which creates an imbalance. Some of the datasets are synthetic dialog rather than natural, which directly affects the performance of natural conversation [22].

Another point in this direction is phonetic recognition in multilingual basis. In a paper by Joachim, it has been tested to identify acoustic and phonetic similarities across languages [12]. The proposed method aims to exploit phonetic similarities between languages. This can be thought of as understanding which languages have been spoken by others without knowing the language. Although the study was promising, there are still many questions left to be answered in paper.

3 Methods

In this study, data are provided by a production company to be studied. One commercial hour corresponds to 45-48 minutes of data. There are 9 commercial hours of data provided. All of these are in different languages. For 3 of the languages, there was also transcribed text. But for the other 4 languages, no transcribed text is found. The Speech-to-Text method had been run on all of them. Elevenlabs Scribe is used for this purpose. All of the audio files have been run over Elevenlabs to get high precision data for both timestamp in milliseconds and word detection.

Additionally, FFMPEG [4], a widely used open-source tool for audio and video processing, was utilized at various stages of the workflow to handle audio format conversion, segmentation, and preprocessing tasks required for subsequent analysis.

3.1 Misalignments

Due to the varying temporal characteristics of speech across different languages, the duration required to express the same content can differ significantly, resulting in second or minute-level misalignments between corresponding segments. Moreover, the dubbing process prioritizes natural and contextually appropriate translations over literal or word-for-word equivalence. As a result, the semantic content of the texts across languages may not align precisely, leading to further discrepancies in cross-lingual textual alignment.

3.2 Targeting Same Content with Dub

This study focuses on speech data in Latin American Spanish, Turkish, and English speech. Given the nature of text embeddings, the goal was to identify segments that convey exactly the same meaning between languages, along with the precise start and end times of each script segment. So, exact same meaning retrieved along with the start and end of the script. To detect exact same content OpenAI and Gemini have been used to automate process. There are approximately 400 lines of script per recording. However, only 5 of them hit exact match, common expressions in languages are excluded. All of this process is run over Python and the final results have been saved as pandas dataframe. When dataframes have all text and timing data, FFMPEG is used to cut and retrieve the required audio data.

[4] https://www.ffmpeg.org

After these process was completed due to computational extensiveness of embeddings, Google Colab was used as a compute resource for this.

3.3 Embeddings

Embeddings have been run on 2 different versions of Wav2Vec. These are facebook/Wav2Vec-base [5] and facebook/Wav2Vec2-base-960h [6]. Due to similar results, only facebook/Wav2Vec-base is presented. On the semantic text embedding side, setu4993/LaBSE [7] have been used.

3.3.1 Comparison of Vectors

The audio files was in sampling rate 48 KHz but Wav2Vec requires 16 KHz sampling rate due to its original settings. After resampling, the specific interval needed for embeddings to be cut. Due to the fact that 20 ms corresponds to a vector, this embedding process gave a tensor with multiple vectors inside it. To run a similarity analysis on these embeddings, the mean of these embeddings had been run.

This process compares the similarity between audio embeddings, more specifically Wav2Vec, and semantic text embeddings. Due to their nature, audio embeddings tend to create more similarity across homonym words, while synonyms create more similarity across semantic embeddings.

3.4 Phonetic Word Embeddings

For phonetic word embeddings, `rahulsrma26/phonetic-word-embedding` pretrained models from GitHub [8] have been used. It is provided in English and Hindi. The aim of this repository and the pre-trained model is to get a phonetic embedding word basis [24]. So in other words, having more similarity across homonyms. Figure 3 shows the similarity of the homonyms with the t-SNE parameters: flow perplexity = 15, component number = 2, iteration = 3500, and random state = 32 with PCA init. With these parameters, embeddings are visualized in 2 dimensions.

3.4.1 Similar Words

In the experimental setup, phonetically similar words are selected for analysis. However, given that the same word can be articulated differently across speakers or contexts, multiple recordings of the same word are included to capture this variation. Consequently, the analysis includes multiple instances per word, which enables better overview of phonetic consistency and variability in the embeddings.

3.4.2 Alignment Issues

Some alignment discrepancies were observed in the timestamps provided by ElevenLabs, primarily due to limitations in temporal precision. Typically, words examined have a duration of 50-300 ms.

[5] https://huggingface.co/facebook/wav2vec2-base
[6] https://huggingface.co/facebook/wav2vec2-base-960h
[7] https://huggingface.co/setu4993/LaBSE
[8] https://github.com/rahulsrma26/phonetic-word-embedding.

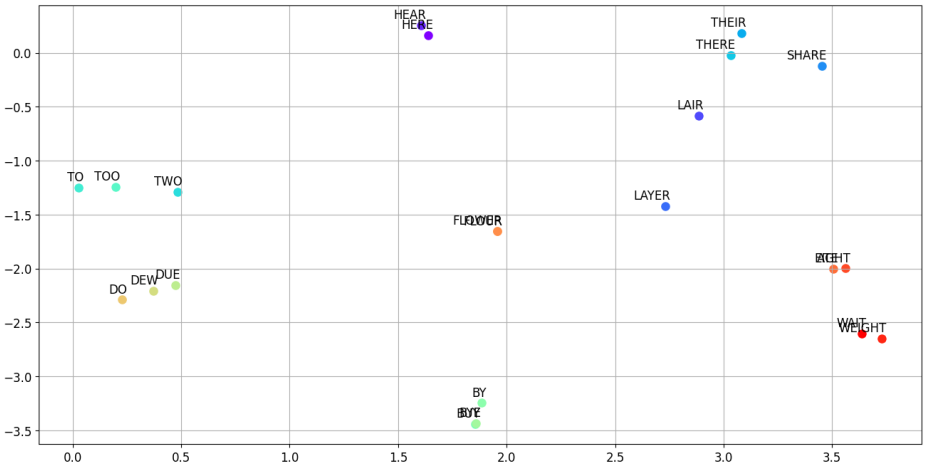

Figure 3: Phonetic word alignments of Homonyms

Since this is a small interval, there are misaligned timestamps from Elevenlabs. These alignments are manually checked and corrected to find the ground truth of the timestamp.

3.4.3 Small Words

Some of the small words are ignored. Especially words with 2 and 3 characters show poor results. Because the words are not clear from the recording. Even listening to the audio, it is hard to understand which words have been spoken.

3.5 Multilingual Base

The main motivation behind these experiments is to have a spoken semantic understanding of the languages without relying on intermediate speech-to-text layers. To support this motivation more, Turkish language also examined with the recordings. But because limitations around phonetic word embeddings for Turkish could not be found, the reference phonetic word embedding figure is missing.

4 Experiments and Results

4.1 Phonetic and Semantic Embedding Comparisons

Figure 4 shows a comparison of embedding similarity with cosine similarity. LAS represents Latin American Spanish, TR represents Turkish, and EN represents English. As explained in Section 3, phonetic Wav2Vec embeddings and semantic text embeddings do not have a direct relation. This structural difference contributes to the observed fluctuations in similarity scores across both modalities and languages.

Notably, at Index 23 in Figure 4, TR-LAS and TR-EN similarity scores are 0.00. This is because of a technical issue while running embeddings on the TR audio. While these values do not reflect meaningful similarity, they are reported here to maintain transparency in the presentation of results.

In Table 1 the audio scripts have been provided below. The reason for providing those is to be able to understand differences in translation and corresponding similarity rates. In most of the scripts the words used have been changed while transcribing languages to protect the meaning of the sentence and make sentences in the same feelings.

Language	Text
English	As if there is a chance of your son getting that money, Father. You're always so eager. He'll put himself in the ground to make that money while we're celebrating.
Latin American Spanish	Lo dices como si ese dinero le llegara al bolsillo de tu hijo, papá. Pero solo son números. Esa idea de trabajar bajo tierra, no sé cuándo generará dinero.
Turkish	Sanki o kadar parayı oğlunun cebine koyacaklar. Baba ya adamdaki hevese bak. İki kuruş kazanmak için yerin dibine girecek. Biz burada uçuyoruz.

Table 1: Scripts to corresponding languages in index 1 of figure 4

4.2 Multilingual base similarities

To capture similarities between multilingual audio recordings, cosine similarity was applied over the audio file. Figure 5 represents cosine similarities in 35 minutes for 3 language pairs with each 500 ms audio segment. Figure 6 represents cosine similarities in 35 minutes for 3 language pairs with each audio segment of 200 ms. As the duration of the segment decreases, an increase in similarity scores is observed. Additionally, the sharp drops in similarity seen in the 500 ms analysis become smoother in the 200 ms analysis, indicating improved temporal resolution.

Notably, substantial drops in similarity (e.g., values below 0.8) in the original recordings are primarily associated with non-speech content, such as background music or silent intervals. Furthermore, the use of shorter segments enhances the detection of high-similarity regions, further supporting the importance of temporal granularity in cross-lingual audio similarity analysis.

4.3 English phonetic audio and word similarity on shorter segments

To preserve more focused study on phonetic audio and phonetic word similarity, an experiment with similar or the same words was conducted. To visualize results, t-SNE applied over the embedding results in Wav2Vec. Figure 7 shows the phonetic similarity over the words. It can be seen that

Figure 4: Cosine similarity of speech and text of same sentences across 4 different samples in 3 language pairs using LaBse and Wav2Vec

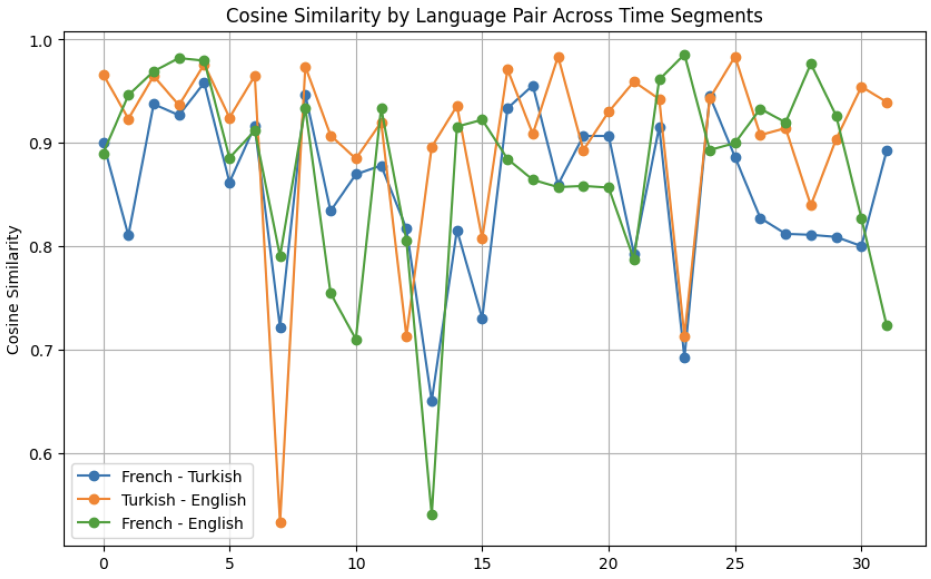

Figure 5: Cosine similarity of speech across language pairs with 500 ms segment size for 35 minutes

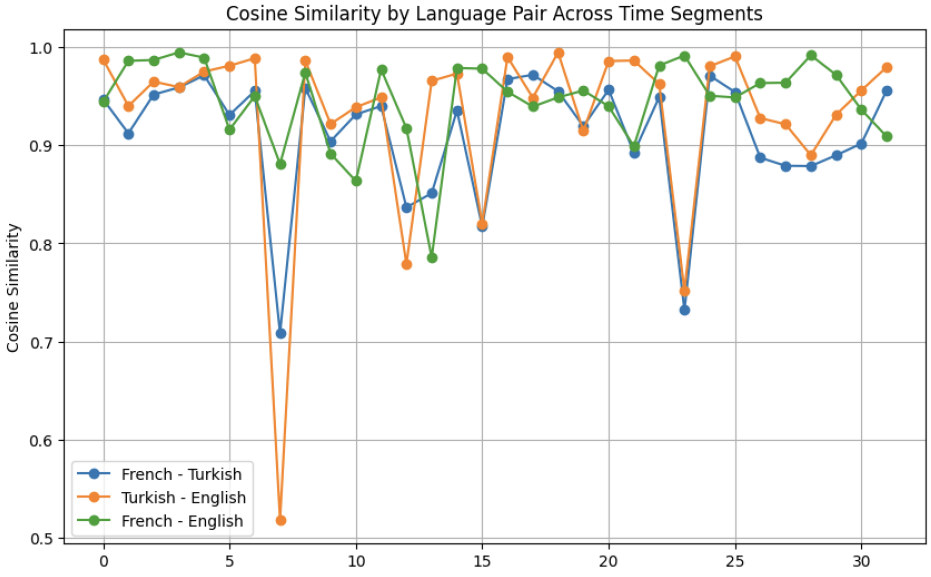

Figure 6: Cosine similarity of speech across language pairs with 200 ms segment size for 35 minutes

homonyms are closer in each other. Figure 8 shows the phonetic similarity with the audio of the words. Each dot is labeled in the format `word (speaker_id) timestamp`. Due to misalignment of the speech to text, words' timestamps are realigned manually. Regarding t-SNE settings, the perplexity of 12, PCA as function, 3500 iterations, and 2 components used.

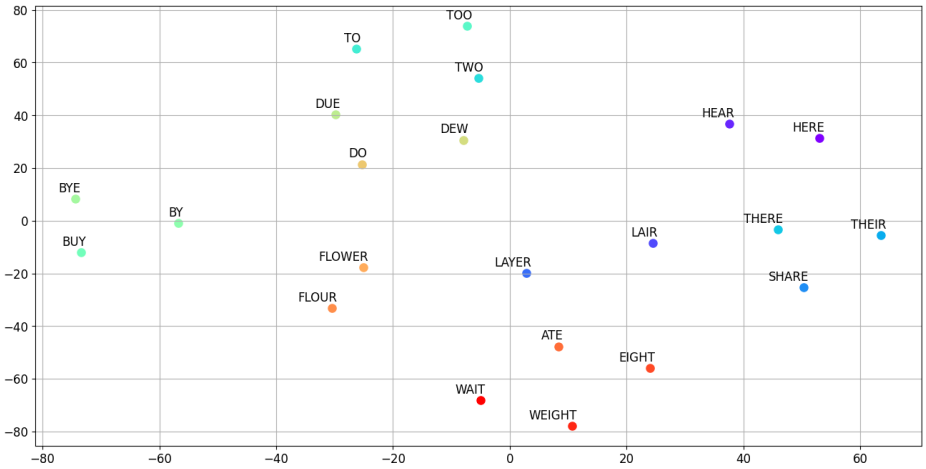

Figure 7: Phonetic word embeddings base phonetic similarity visualization with t-SNE

After reviewing the words left in Figure 8, it was found that some words are closer to each other without being homonyms. One of the reasons for this was that they are from the same speaker. So to examine this more, another episode of the series had been compared with the speaker id's. In this episode, all speakers are labeled with Elevenlabs ids. In addition, all the dots in the figure 8 had been aligned manually. In most of the instances, it is easily seen that either there is a same speaker close to the point, or there is a homonym.

To actually prove that the same speaker is an important parameter, the points in Figure 8 are filtered for speaker 21 only and shown in Figure 9. Although not all, some of the words clearly represent close embedding results, e.g. "know" and "No" on the middle left. Although this visualization looks nice, a similarity calculation is also run for these points. Because we measure cosine similarity across embeddings, normalization of the embeddings was run to get better results.

Also cosine similarity of the same data in Figure has been run and shown in Table . This data is focused on interword and intraword cosine similarity values across samples. For the values with just one sample label itself, have do not have a similarity value since its not possible to calculate across one sample. The values in the table have been interpreted in 2 ways. First, we compare similarity of 2 homonyms from a nonhomonym word and look for similar cosine similarity levels. Second, we compare each word similar to its phonetically closest word.

In the first comparison, the strongest alignment is observed between the words "know" and "no", which consistently exhibit high similarity scores across nearly all samples. This observation is further supported by the visualization of the t-SNE in Figure 9, where these words appear close together, reinforcing the phonetic similarity captured by the embeddings. Conversely, certain word pairs such as "buy" and "by", despite being homophones, do not consistently yield high similarity scores as might be expected. Similar inconsistencies are observed for other homophonous sets such as "too" and "two", as well as "hear", "here", and "there". While some similarity values

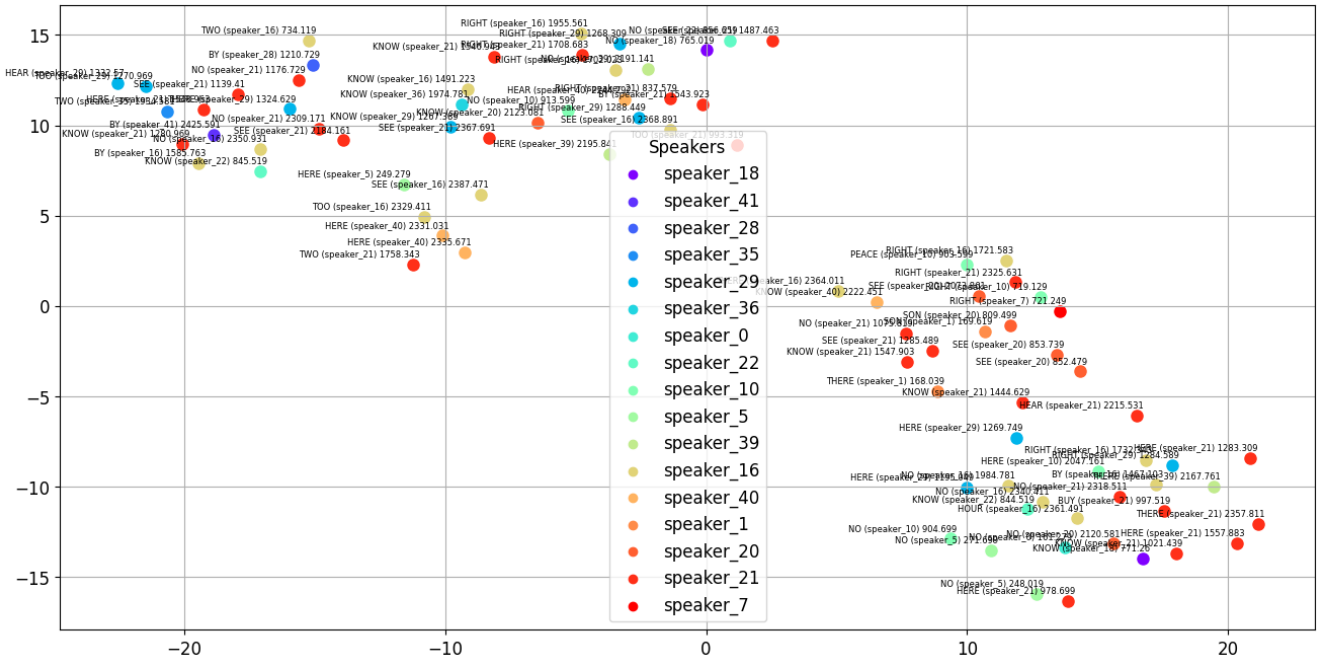

Figure 8: Audio embeddings visualization with t-SNE over speakers

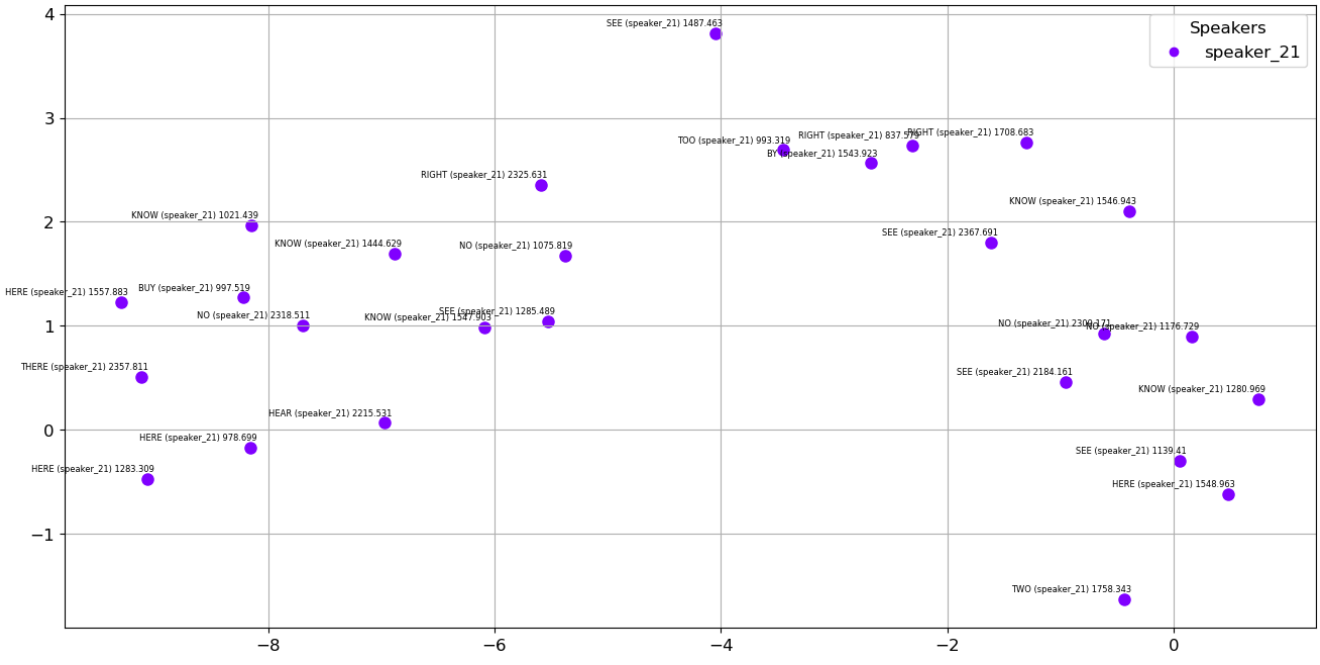

Figure 9: Audio embeddings visualization with t-SNE over speaker 21

align with homonym expectations, others deviate notably. These anomalies are likely attributable to the inherent variability and complexity of spoken language data, which introduces noise and subtle phonetic differences that the embeddings may not fully capture. As such, these discrepancies highlight a key limitation of the current approach and suggest that further refinement is necessary to achieve more consistent performance in capturing homophonic relationships.

In the second comparison, similarity values generally indicate clear phonetic alignment across several pairs of words. However, notable anomalies are also present. For example, the word "here" exhibits its highest similarity with "buy" (0.55), followed closely by "there" (0.53). The high similarity score between "here" and "buy" is unexpected and does not align with intuitive phonetic or semantic expectations. Such inconsistencies highlight limitations in the current methodology and suggest that the embeddings may not fully capture phonetic structure in all cases. These anomalies underscore the need for further refinement of the embedding models and more rigorous evaluation to ensure consistent phonetic interpretability.

Table 2: Pairwise Average Cosine Similarities Between Normalized Target Word Embeddings of speaker 21

Word	buy	by	hear	here	know	no	right	see	there	too	two
buy	*N/A*	0.3679	0.7226	0.5505	0.4653	0.3857	0.2472	0.2549	0.6669	0.3513	0.4186
by	0.3679	*N/A*	0.4512	0.1876	0.5036	0.5148	0.7363	0.5157	0.1645	0.7895	0.3960
hear	0.7226	0.4512	*N/A*	0.4946	0.4674	0.4391	0.3174	0.3367	0.6378	0.4542	0.5381
here	0.5505	0.1876	0.4946	0.3838	0.3391	0.2983	0.1340	0.2230	0.5391	0.1881	0.4377
know	0.4653	0.5036	0.4674	0.3391	0.3827	0.4545	0.4251	0.3953	0.3225	0.4863	0.4307
no	0.3857	0.5148	0.4391	0.2983	0.4545	0.3421	0.4274	0.4350	0.2621	0.4763	0.4075
right	0.2472	0.7363	0.3174	0.1340	0.4251	0.4274	0.5337	0.4340	0.1127	0.7100	0.3029
see	0.2549	0.5157	0.3367	0.2230	0.3953	0.4350	0.4340	0.3764	0.1482	0.4757	0.3989
there	0.6669	0.1645	0.6378	0.5391	0.3225	0.2621	0.1127	0.1482	*N/A*	0.1877	0.3480
too	0.3513	0.7895	0.4542	0.1881	0.4863	0.4763	0.7100	0.4757	0.1877	*N/A*	0.3126
two	0.4186	0.3960	0.5381	0.4377	0.4307	0.4075	0.3029	0.3989	0.3480	0.3126	*N/A*

Note: Diagonal elements represent average intra-word cosine similarity. "N/A" indicates insufficient samples (1) for intra-word similarity calculation.

The same values from Table 2 were used below for the selected target words, sorted from most to least similar. The corresponding homonym for each target word is highlighted in red. This format helps to quickly assess which words are semantically or contextually closest in the embedding space for this speaker.

The nine tables in Figure 10 show the average cosine similarity scores between a target word and a list of other words. The target word for each table is labeled underneath it. The words in the "Word" column are ranked from highest to lowest similarity, so the word at the top of the list is the most similar to the target word. The words highlighted in red are homonyms (or close phonetic matches) of the target word. Each similarity score is the average of all pairwise cosine distances between instances of that word pair.

From these sorted lists, several interesting patterns emerge for speaker 21:

- **Homonyms are not always closest:** In most cases, the homonym (e.g., "here' for "hear",

Figure 10: Pairwise average cosine similarities for selected words, sorted from most to least similar for each target word. The corresponding homonym is highlighted in red.

Word	Similarity
buy	0.7226
there	0.6378
two	0.5381
here	0.4946
know	0.4674
too	0.4542
by	0.4512
no	0.4391
see	0.3367
right	0.3174

Similarity to **"hear"**

Word	Similarity
buy	0.5505
there	0.5391
hear	0.4946
two	0.4377
know	0.3391
no	0.2983
see	0.2230
too	0.1881
by	0.1876
right	0.1340

Similarity to **"here"**

Word	Similarity
hear	0.7226
there	0.6669
here	0.5505
know	0.4653
two	0.4186
no	0.3857
by	0.3679
too	0.3513
see	0.2549
right	0.2472

Similarity to **"buy"**

Word	Similarity
too	0.7895
right	0.7363
see	0.5157
no	0.5148
know	0.5036
hear	0.4512
two	0.3960
buy	0.3679
here	0.1876
there	0.1645

Similarity to **"by"**

Word	Similarity
by	0.7895
right	0.7100
know	0.4863
no	0.4763
see	0.4757
hear	0.4542
buy	0.3513
two	0.3126
here	0.1881
there	0.1877

Similarity to **"too"**

Word	Similarity
hear	0.5381
here	0.4377
know	0.4307
buy	0.4186
no	0.4075
see	0.3989
by	0.3960
there	0.3480
too	0.3126
right	0.3029

Similarity to **"two"**

Word	Similarity
by	0.5036
too	0.4863
hear	0.4674
buy	0.4653
no	0.4545
two	0.4307
right	0.4251
see	0.3953
here	0.3391
there	0.3225

Similarity to **"know"**

Word	Similarity
by	0.5148
too	0.4763
know	0.4545
hear	0.4391
see	0.4350
right	0.4274
two	0.4075
buy	0.3857
here	0.2983
there	0.2621

Similarity to **"no"**

Word	Similarity
by	0.7363
too	0.7100
see	0.4340
no	0.4274
know	0.4251
hear	0.3174
two	0.3029
buy	0.2472
here	0.1340
there	0.1127

Similarity to **"right"**

"by‘ for "buy") is not the most similar word in the embedding space. This suggests the model is successfully capturing contextual or acoustic differences rather than just phonetic similarity.

- **"buy" vs "by":** The similarity between "buy" and its homonym "by" is quite low (0.3679). For this speaker, the embedding for "buy" is much closer to verbs like "hear" (0.7226) and location words like ‘there‘ (0.6669).
- **"too" vs "two":** The similarity between "too" and ‘two" is also surprisingly low (0.3126). The word "too" is extremely similar to the functional words "by‘ (0.7895) and "right" (0.7100).
- **"hear" vs "here" vs "there":** The homonyms is ranked close in these instances, falling behind "buy", and "two". This indicates a dominant similarity except "buy".
- **Strong Cluster of "by", "too", ‘right":** These three words show very high pairwise similarity. This likely reflects their use as short, functional adverbs or prepositions in similar sentence structures.

5 Discussions

In Section 4.1, the phonetic and semantic embeddings will be compared in 3 languages with the same sentences. The results open the door to an experiment with phonetic word embeddings which is done in Section 4.3. In addition, results have led to the paper that assesses similarity in a long recording instead of small recordings in Section 4.2. The aim is to understand whether we can detect a language with just phonetics of it. This can be thought of as understanding which language has been spoken by others without knowing the language as human. The results in Section 4.2 showed that as the length of the recording decreased, the similarity score increased due to non-speech events [33]. Another reason of this is that language-agnostic human speech is more similar in different languages than some other sound such as nature sounds, door closing sound, etc. [27]. Although Section 4.2 provided an understanding of multilingual basis, it does not provide any significant proof. The findings in both Section 4.1 and Section 4.2 have led to a final experiment of Section 4.3. Section 4.3 shows a similarity in phonetic audio embeddings and phonetic word embeddings. If these phonetic word embeddings can be aligned with semantic word embeddings, it can result in a conceptual understanding of a language.

In general, the results in t-SNE and similarity values show potential. There are values that show a clear proof of the research question, but there are also some data that show otherwise. The main reason for this is the limitation of the research. As the nature of audio is rich, there are many more features that audio embeddings capture, but not all of this can be quantized and shown in similarity analysis and/or t-SNE. To provide a clearer response to **RQ1.1**, Sections 4.1 and 4.2 offer evidence suggesting a degree of similarity. However, these findings are not robust enough to conclusively demonstrate that the embeddings capture similarity in a consistent or meaningful way. In contrast, regarding **RQ1.2**, there is stronger evidence indicating that the embeddings capture aspects of similarity. While some cases exhibit weaker results, these can largely be attributed to methodological limitations or constraints in the dataset.

In general, in addressing **RQ1**, it can be concluded that a conceptual understanding of similarity between languages is achievable through the use of phonetic embeddings within a multilingual context. However, achieving this understanding requires extensive data preprocessing. Even with thorough preprocessing, mismatches may still occur due to inherent limitations in the approach, and further investigation into accuracy and reliability is necessary.

Importantly, while humans are capable of performing such cross-linguistic similarity recognition in real time, the computational processes required to replicate this behavior remain resource-intensive and challenging to implement in real-time systems. This gap underscores the need for continued technical advancements in embedding models and processing pipelines to approximate the efficiency and accuracy of human perception.

6 Conclusions and Further Research

The papers have shown a potential for using the joint method for audio embeddings and phonetic word embeddings to align both semantic and phonetic structures of the languages. Multilingual usage of this structure remains an important rule for speech-to-speech engines.

Research can be extended to both multilingual and embedding perspectives. Joint architecture for embeddings can be a significant add-on for the study. In addition, using this architecture on multilingual landscape could be a good improvement in the work.

7 Acknowledgments

A big thanks to my supervisors Dr. Rob Saunders and Dr. Zhenghun yue for their contributions to my bachelor thesis. I also thank my family for their contributions for all of my bachelors. Finally, I thank the Kadraj group for providing data needed to study on this particular topic.

References

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[2] Somnath Banerjee, Avik Halder, Rajarshi Mandal, Sayan Layek, Ian Soboroff, Rima Hazra, and Animesh Mukherjee. Breaking boundaries: Investigating the effects of model editing on cross-linguistic performance, 2025.

[3] Yi-Chen Chen, Sung-Feng Huang, Chia-Hao Shen, Hung yi Lee, and Lin shan Lee. Phonetic-and-semantic embedding of spoken words with applications in spoken content retrieval, 2019.

[4] Mithun Das, Somnath Banerjee, and Animesh Mukherjee. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 32–42, New York, NY, USA, 2022. Association for Computing Machinery.

[5] Benjamin Elizalde, Shuayb Zarar, and Bhiksha Raj. Cross modal audio search and retrieval with joint embeddings based on text and audio. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4095–4099, 2019.

[6] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852, 2020.

[7] Roberto Gallotta, Graham Todd, Marvin Zammit, Sam Earle, Antonios Liapis, Julian Togelius, and Georgios N. Yannakakis. Large language models and games: A survey and roadmap. *IEEE Transactions on Games*, pages 1–18, 2024.

[8] Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, Peng Liu, Ruihang Miao, Wang You, Xi Chen, Xuerui Yang, Yechang Huang, Yuxiang Zhang, Zheng Gong, Zixin Zhang, Hongyu Zhou, Jianjian Sun, Brian

Li, Chengting Feng, Changyi Wan, Hanpeng Hu, Jianchang Wu, Jiangjie Zhen, Ranchen Ming, Song Yuan, Xuelin Zhang, Yu Zhou, Bingxin Li, Buyun Ma, Hongyuan Wang, Kang An, Wei Ji, Wen Li, Xuan Wen, Xiangwen Kong, Yuankai Ma, Yuanwei Liang, Yun Mou, Bahtiyar Ahmidi, Bin Wang, Bo Li, Changxin Miao, Chen Xu, Chenrun Wang, Dapeng Shi, Deshan Sun, Dingyuan Hu, Dula Sai, Enle Liu, Guanzhe Huang, Gulin Yan, Heng Wang, Haonan Jia, Haoyang Zhang, Jiahao Gong, Junjing Guo, Jiashuai Liu, Jiahong Liu, Jie Feng, Jie Wu, Jiaoren Wu, Jie Yang, Jinguo Wang, Jingyang Zhang, Junzhe Lin, Kaixiang Li, Lei Xia, Li Zhou, Liang Zhao, Longlong Gu, Mei Chen, Menglin Wu, Ming Li, Mingxiao Li, Mingliang Li, Mingyao Liang, Na Wang, Nie Hao, Qiling Wu, Qinyuan Tan, Ran Sun, Shuai Shuai, Shaoliang Pang, Shiliang Yang, Shuli Gao, Shanshan Yuan, Siqi Liu, Shihong Deng, Shilei Jiang, Sitong Liu, Tiancheng Cao, Tianyu Wang, Wenjin Deng, Wuxun Xie, Weipeng Ming, Wenqing He, Wen Sun, Xin Han, Xin Huang, Xiaomin Deng, Xiaojia Liu, Xin Wu, Xu Zhao, Yanan Wei, Yanbo Yu, Yang Cao, Yangguang Li, Yangzhen Ma, Yanming Xu, Yaoyu Wang, Yaqiang Shi, Yilei Wang, Yizhuang Zhou, Yinmin Zhong, Yang Zhang, Yaoben Wei, Yu Luo, Yuanwei Lu, Yuhe Yin, Yuchu Luo, Yuanhao Ding, Yuting Yan, Yaqi Dai, Yuxiang Yang, Zhe Xie, Zheng Ge, Zheng Sun, Zhewei Huang, Zhichao Chang, Zhisheng Guan, Zidong Yang, Zili Zhang, Binxing Jiao, Daxin Jiang, Heung-Yeung Shum, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Xinhao Zhang, and Yibo Zhu. Step-audio: Unified understanding and generation in intelligent speech interaction, 2025.

[9] Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. Europarl-st: A multilingual corpus for speech translation of parliamentary debates, 2020.

[10] Sameer Khurana, Nauman Dawalatabad, Antoine Laurent, Luis Vicente, Pablo Gimeno, Victoria Mingote, and James Glass. Improved cross-lingual transfer learning for automatic speech translation, 2024.

[11] Ondrej Klempir, Adela Skryjova, Ales Tichopad, and Radim Krupicka. Ranking pretrained speech embeddings in parkinson's disease detection: Does wav2vec 2.0 outperform its 1.0 version across speech modes and languages? *medRxiv*, 2025.

[12] J. Kohler. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 4, pages 2195–2198 vol.4, 1996.

[13] Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, Jianhua Xu, Haoze Sun, Zenan Zhou, and Weipeng Chen. Baichuan-audio: A unified framework for end-to-end speech interaction, 2025.

[14] Yuang Li, Xiaosong Qiao, Xiaofeng Zhao, Huan Zhao, Wei Tang, Min Zhang, and Hao Yang. Large language model should understand pinyin for chinese asr error correction, 2024.

[15] Dawei Liang, Hang Su, Tarun Singh, Jay Mahadeokar, Shanil Puri, Jiedan Zhu, Edison Thomaz, and Mike Seltzer. Dynamic speech endpoint detection with regression targets. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[16] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, CAICE '24, page 405–409, New York, NY, USA, 2024. Association for Computing Machinery.

[17] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration, 2023.

[18] Rao Ma, Mengjie Qian, Potsawee Manakul, Mark Gales, and Kate Knill. Can generative large language models perform asr error correction?, 2023.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[20] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[21] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019.

[22] Mohammad Sarim, Saim Shakeel, Laeeba Javed, Jamaluddin, and Mohammad Nadeem. Direct speech to speech translation: A review, 2025.

[23] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *CoRR*, abs/1904.05862, 2019.

[24] Rahul Sharma, Kunal Dhawan, and Balakrishna Pailla. Phonetic word embeddings. *CoRR*, abs/2109.14796, 2021.

[25] Adriana Stan and Johannah O'Mahony. An analysis on the effects of speaker embedding choice in non auto-regressive TTS. In *12th Speech Synthesis Workshop (SSW) 2023*, 2023.

[26] Chen-Tse Tsai and Dan Roth. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, 2016.

[27] Léo Varnet, Maria Clemencia Ortiz-Barajas, Ramón Guevara Erra, Judit Gervain, and Christian Lorenzi. A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, 142(4):1976–1989, 10 2017.

[28] Mercedes Vetráb and Gábor Gosztolya. Aggregation strategies of wav2vec 2.0 embeddings for computational paralinguistic tasks. In Alexey Karpov, K. Samudravijaya, K. T. Deepak, Rajesh M. Hegde, Shyam S. Agrawal, and S. R. Mahadeva Prasanna, editors, *Speech and Computer*, pages 79–93, Cham, 2023. Springer Nature Switzerland.

[29] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online, August 2021. Association for Computational Linguistics.

[30] Liyan Wang, Jun Yang, Yongshan Wang, Yong Qi, Shuai Wang, and Jian Li. Integrating large language models (llms) and deep representations of emotional features for the recognition and evaluation of emotions in spoken english. *Applied Sciences*, 14(9), 2024.

[31] Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances, 2024.

[32] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92), 2019. University of Edinburgh. The Centre for Speech Technology Research (CSTR).

[33] Goksenin Yuksel, Marcel van Gerven, and Kiki van der Heijden. General-purpose audio representation learning for real-world sound scenes, 2025.

8 Appendix

Language	Text
English	Oh, give me time to dry. It's my boy. He got 90%. He passed another class. At this rate, he's going to graduate.
Latin American Spanish	Ah. Permítame. Ese es mi hijo. Obtuvo noventa puntos. Ya pasó el otro examen. Muy pronto terminará la escuela.
Turkish	Bir dakika sen kullan. Aslan oğlum doksan puan almış. Bir sınav daha vermiş. Bu gidişle bitirecek okulu.

Table 3: Scripts to corresponding languages in index 2 of figure 4

Language	Text
English	Listen, please call your mother. Tell her to cook something extra special tonight.
Latin American Spanish	Por cierto, ¿puedes llamar a tu madre y pedirle que prepare algo rico de cenar?
Turkish	Bana bak, annene telefon et. Akşama şöyle güzel bir sofra hazırlasın.

Table 4: Scripts to corresponding languages in index 7 of figure 4

Language	Text
English	I will help you.
Latin American Spanish	Déjame ayudarte.
Turkish	Ben yardımcı olayım.

Table 5: Scripts to corresponding languages in index 13 of figure 4

Language	Text
English	No way.
Latin American Spanish	No, ¿cómo?
Turkish	Yok oğlum, sorma.

Table 6: Scripts to corresponding languages in index 23 of figure 4

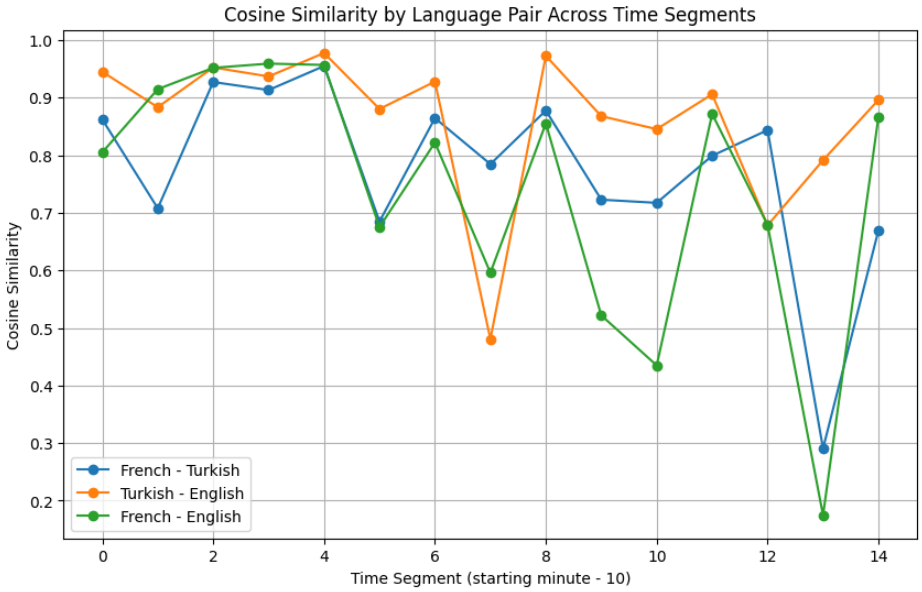

Figure 11: Consine similarity of speech across language pairs with 1 second segment size for 15 minutes

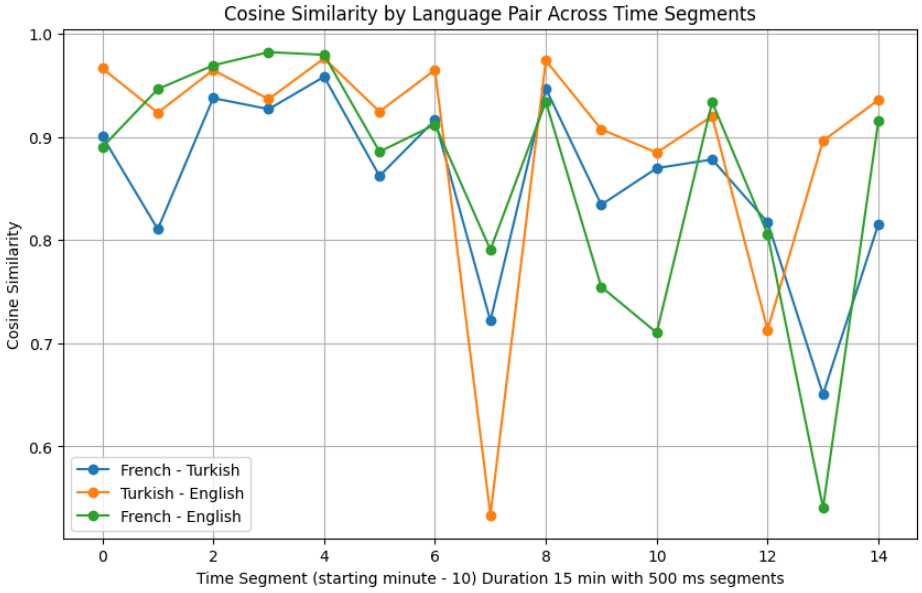

Figure 12: Consine similarity of speech across language pairs with 500 ms segment size for 15 minutes

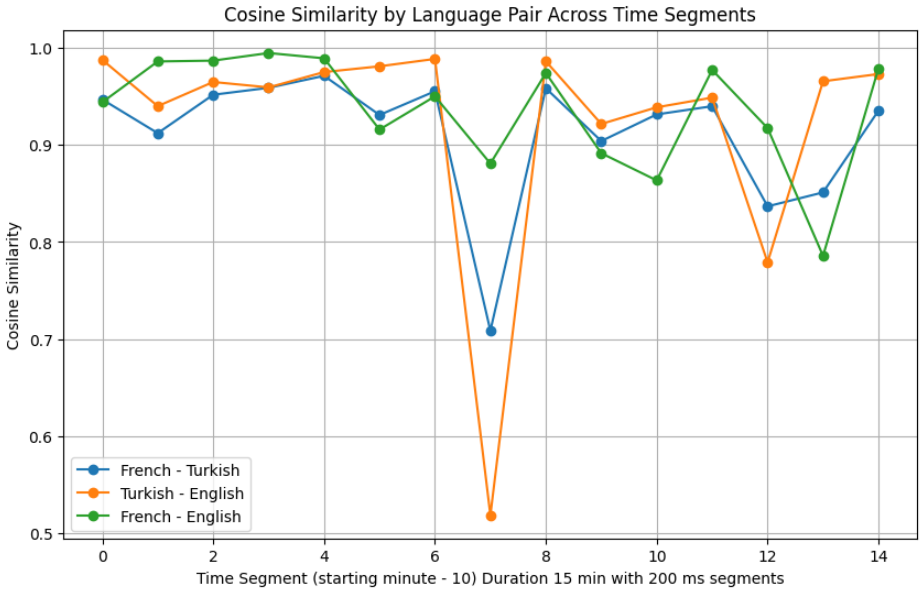

Figure 13: Consine similarity of speech across language pairs with 200 ms segment size for 15 minutes

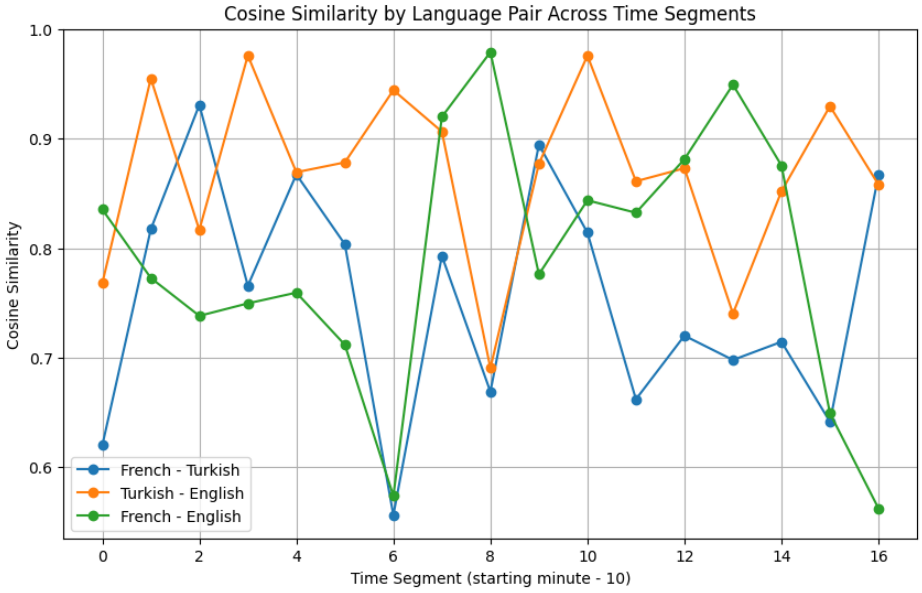

Figure 14: Consine similarity of speech across language pairs with 1 second segment size for 20 minutes

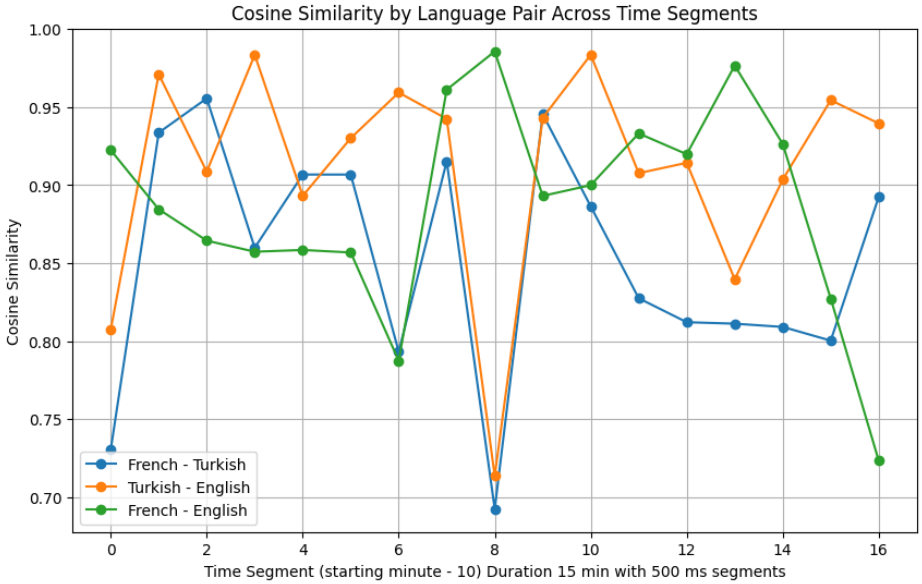

Figure 15: Consine similarity of speech across language pairs with 500 ms segment size for 20 minutes

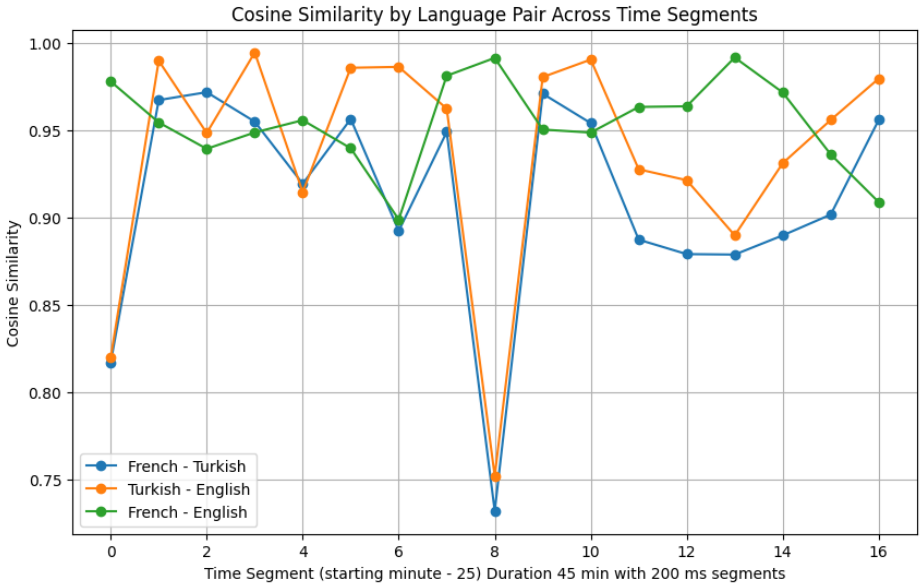

Figure 16: Consine similarity of speech across language pairs with 200 ms segment size for 20 minutes