Thesis bachelor Artificial Intelligence

The Impact of Bilingual Exposure on Word Segmentation: Evaluating Dutch-German Trained ASR Models on English Speech

Author

Izabelle Auriaux, 0989274

i.g.v.auriaux@students.uu.nl

First reader

F. W. Adriaans

f.w.adriaans@uu.nl

Second reader

M. Fowlie

m.fowlie@uu.nl

Kunstmatige Intelligentie

7,5 ECTS

Utrecht University

Faculty of Humanities

Universiteit Utrecht

Netherlands

18-04-2025

Table of Content

Abstract

This study explores the impact of bilingual exposure on word segmentation performance in an unseen language using Wav2Vec 2.0, a self-supervised Automatic Speech Recognition (ASR) model. Specifically, it examines how varying Dutch-German exposure ratios affect the model's ability to segment English speech. Five models were trained with different Dutch-German exposure distributions: 100%-0%, 75%-25%, 50%-50%, 25%-75%, and 0%-100%, and their segmentation accuracy was evaluated using the TIMIT dataset.

The Dutch-only model achieved the highest segmentation accuracy, while the 50%-50% model performed the worst. Bilingual models (75%-25% and 25%-75%) slightly outperformed the German-only model, suggesting that Dutch's phonetic similarity to English aided segmentation. All models under-segmented sentences, likely due to a bias toward strong syllables and omission of function words, which are more frequent in English.

These findings challenge assumptions that bilingual exposure always enhances generalization, showing that equal bilingual input, without increased total data exposure, can reduce segmentation accuracy.

Limitations included limited training time per model and a small test set. Future research could explore whether increasing data size mitigates the negative effects of bilingual exposure, and whether similar patterns appear when testing on languages more distant from the training set.

This study highlights how linguistic similarity, bilingual data distribution, and input quantity affect performance in multilingual ASR systems. The findings have implications for speech recognition and bilingual language acquisition, emphasizing the importance of balanced multilingual training strategies.

Introduction

Language learning in bilingual children

It is well known that bilingually raised children tend to pick up on learning a new language more quickly and efficiently than monolingual children. Studies suggest that bilingualism enhances cognitive flexibility, phonological awareness, and general language-learning ability (Kovács & Mehler, 2009).
Even in unbalanced bilingual environments, exposure to a second language still offers significant benefits. Research indicates that children with some familiarity with a second language, even if they are not fully proficient, develop greater phonetic awareness and an enhanced ability to distinguish subtle differences in sounds, thus bringing a lot of similar pro's to a child's language learning skills as a child fully raised bilingual (Deanda, Arias-Trejo, Poulin-Dubois, Zesiger, & Friend, 2015).

One such part of language learning is the skill of word segmentation. This is the process of recognizing where one word ends, and another begins. While speaking words often tend to flow into each other instead of having clear boundaries created by a vacuum of sound. This is an important skill for learning a language, whether as a baby hearing speech for the first time or an adult learning a new language. People use different clues to figure out word boundaries, such as the rhythm of speech, common sound patterns, and how often certain sounds appear together (Saffran, Aslin, & Newport, 1996).

What is mainly used to find clear word boundaries during speaking also depends on the language. Different languages have distinct phonological, morphological, and syntactic structures that affect how people identify word boundaries when listening to that language. For example, in English, stressed syllables often signal the beginning of a word (Cutler, 1989). In *"conduct"*, stress on the first syllable signals a noun ("CONduct"), while stress on the second signals a verb ("conDUCT").
Phonatic constraints also play a role, some sound combinations are more likely to occur within the same word than across word boundaries. For example, in Japanese, the syllable structure makes certain sequences (e.g., "n" followed by a "b" sound) unlikely across words (Otake, Hatano, Cutler, & Mehler, 1993). Infants learning any language rely on patterns to segment words, identifying where syllables frequently transition within words rather than across them (Saffran, Aslin, & Newport, 1996). As mentioned earlier, reliance on this cue varies by language. English learners might rely more on stress, while Japanese learners may prioritize vowel length and overall structure.

Bilinguals, however, face a more complex segmentation task. Since different languages rely on different segmentation strategies, bilingual individuals must navigate multiple, sometimes conflicting, sets of linguistic cues. Some research suggests that bilingual children develop more flexible segmentation strategies, as they are exposed to multiple sets of phonatic constraints and rhythmic structures from an early age.
It also introduces potential interference when the languages differ in segmentation cues.

Several studies have explored segmentation in both human learners and artificial systems. Swingley & Algayres (2024) demonstrated that computational models can

simulate infant word segmentation behavior using statistical patterns in raw speech input. Their results show that unsupervised models, like DP-parsers, can infer word boundaries without labeled data. However, their research focused on monolingual exposure and did not test how models respond when given input in more than one language.

In contrast, Adriaans (2024) discussed the complexity of bilingual phonetic acquisition, showing how competing cues from two languages can interfere with segmentation. His work underlined the challenge of developing consistent strategies in the presence of overlapping but distinct segmentation rules. While this gives insight into human bilingual acquisition, it remains unclear whether computational models experience the same kind of interference or if they develop generalizable strategies from bilingual input.

My study builds directly on both of these works. It extends the exposure-based logic of Swingley & Algayres by testing unsupervised segmentation in multilingual input conditions, and it tests Adriaans' claims about interference within a computational model, using different bilingual exposure ratios. This allows me to examine not only if bilingual exposure helps or hurts segmentation performance, but also how the distribution of that exposure (e.g., 50-50 vs. 75-25) shapes the outcome.

Language learning for large language models

Just as humans rely on phonetical patterns to segment speech into words, artificial intelligence models also face the challenge of breaking down continuous speech into meaningful units. This process is important for Automatic Speech Recognition (ASR) systems and Natural Language Processing applications. While similar, it is important to distinguish between word segmentation, word splicing, and ASR, as these terms describe different aspects of speech processing.

Where word segmentation referred to the process of identifying boundaries between words in continuous speech. Word splicing, on the other hand, is a related but, distinct concept that focuses on extracting and rearranging segments of speech. While segmentation aims to recognize word boundaries, splicing involves manipulating speech fragments, often for speech synthesis, voice cloning, or data augmentation purposes.

Automatic Speech Recognition (ASR) encompasses both segmentation and splicing but serves a broader goal: converting spoken language into written text. ASR models must segment words correctly but, also transcribe them accurately, handling variations in pronunciations, accents, and contextual ambiguities.

Computational models, particularly self-supervised speech models, have become an important tool for studying word segmentation. As they learn patterns in spoken language without relying on labeled data, they serve as an effective means of investigating how different linguistic experiences shape segmentation abilities. They mimic aspects of human language acquisition by identifying recurring structures in speech, much like how infants develop an understanding of word boundaries through statistical learning (Saffran, Aslin, & Newport, 1996).

Baevski et al. (2021) supports this connection between self-supervised models and human language learning. For example, Wav2Vec 2.0, a self-supervised ASR model,

learns to recognize phonemes and word-like units without explicit annotations, similar to how infants rely on exposure-based learning.

Research and hypothesis

Research has shown that ASR models can help us understand how people process and learn languages (Dupoux, 2018). By training these models with different language experiences, we can see how well they recognize word boundaries in a new language. These models, such as Wav2Vec 2.0, learn directly from raw audio without explicit annotations, mirroring how children learn language through exposure rather than instruction. These models identify recurring sound patterns and word-like units in a way that is similar to learning in humans (Saffran, Aslin, & Newport, 1996). By manipulating the input, such as different languages, exposure levels, or linguistic similarity, it's easier to simulate controlled experiments that would be difficult to perform with human learners, specifically children. As Dupoux (2018) suggests, computational models allow us to research the limits of what can be learned from input alone, offering insight into how language experience shapes segmentation ability. In this sense, ASR models serve not only as engineering tools but as experimental platforms for testing theories of language acquisition and transfer learning.

This is the basis of what I will be doing to answer the research question *"Can a computational model with regular exposure to a second language do a better job at word splicing compared to a computational model that did not get this exposure?"*

To better understand the different factors that influence the result of this study I first researched a few sub-questions that were mostly answered in a literature study. *Is there a measurable benefit from testing on a linguistically related language?"* was an important question, since the models were limited in both time and data. If linguistically similar languages could help the models perform better, this would clarify the results and suggest a useful strategy for improving performance. I chose three languages for this research to keep it within reasonable boundaries. The languages: Dutch, German and English, were chosen due to their linguistic similarities and similar approach to tackle word segmentation tasks.

Dutch, German and English all belong to the West Germanic language family (Buccini, 2024) This means they share phonological, morphological, and syntactic structures that could influence segmentation performance (Blevins, 2006). Notably, Dutch is often considered an intermediary between German and English in terms of syntax and phonology (Auwera & Noel, 2011).

Both Dutch and German also use stress-based segmentation strategies similar to English, where stressed syllables tend to mark word boundaries (Cutler, 1989). Prior research has also shown that bilingual individuals often exhibit cross-linguistic transfer in phonological processing (Sebastián-Gallés & Bosch, 2009). By training models on different Dutch-German exposure ratios and evaluating their segmentation performance on English, this study tests whether computational models mirror such transfer effects observed in human learners.

Another important aspect to research before beginning was *"How would the amount and distribution of training data across one or more languages affect the model's ability to segment words in a target language?"*. *Kumar et al*. (2022) found that increasing dataset size generally improves performance across speech models, but how close that language is related plays a distinct role in transfer learning outcomes. This aligns with

Conneau et al. (2020), who emphasize the importance of controlling for training volume when evaluating cross-lingual models to avoid attributing gains solely to exposure quantity. Based on these findings, I decided to keep the total amount of training data constant across all models, monolingual and bilingual alike. This helps focus the research specifically on the effect of language exposure, ensuring that any observed improvements in segmentation performance are due to the presence of a second language, and not because another model has simply had more training data.
It was particularly interesting to see if these models would also perform better on a language that was not included in their training data. Similair to how children that are raised bilingual, also do a better job at language related tasks, in an unfamiliar language.
I hypothesize that computational models trained with different Dutch-German exposure ratios will show varying performance levels on an English word-splicing task.
Specifically, the model trained with equal exposure to both Dutch and German is expected to perform the worst. Since it does not receive sufficient training data in either language to develop strong, language-specific segmentation patterns, it may struggle to establish clear boundaries in English. Prior research suggests that when exposure to multiple languages is not substantial enough, it can hinder linguistic task performance due to weaker statistical representations of either language (Flege, MacKay, & Meador, 1999).
Ultimately, the models trained exclusively on either Dutch or German are predicted to achieve the best performance. Full exposure to a single language is likely to result in more robust statistical and phonetic segmentation patterns, without the potential interference from competing cues in a second language, even if those segmentation patterns do not exactly match the test languages required patterns. While bilingual exposure can help cognitive flexibility in human learners, research has shown that monolingual learners often develop stronger and more consistent segmentation strategies in their dominant language (Weber & Cutler, 2004).
This aligns with the challenges discussed by *Adriaans* (2024), who highlighted the complexity of bilingual phonetical acquisition in computational models, particularly when phonological cues overlap or conflict across languages. Combined with the fact that phonetic constraints in both Dutch and German closely resemble those of English, this suggests that concentrated exposure to a single related language may offer an advantage for English word segmentation.

Resources in language learning for LLM's

These research results are relevant for AI, particularly in understanding how exposure to multiple languages impacts learning in speech models. While it is expected that models trained exclusively on a single language will generally outperform those trained on multiple languages for the same amount of time, this research suggests that models trained on multilingual data can still achieve good performances in tasks like word segmentation. This has important implications for efficiency, as it could rule out the assumption that multilingual models require extensive retraining to perform well. As data is hard and expensive to come by, using these techniques could reduce the need for training on large, diverse datasets, ultimately saving time and resources. The research by *Kürzinger et al*. (2020) on transfer learning and multilingual training, using CTC-segmentation, have shown that it's possible to improve performance across languages without fully retraining models on each dataset. This research continues that

idea, to see if instead of training a model a bit less on each language it would be possible to utilise the same amount of data needed for sufficient comprehension of single language and still have sufficient results in word segmentation tasks.

Methods

Used tools

The following methods were used to train and evaluate five ASR models with varying Dutch-German language exposure ratios, using open-source tools and speech datasets. The initial idea to use the current model & evaluation technique was inspired by the research findings by *Kürzinger et al*. who, in their research on word segmentation tasks on End-to-end Speech Recognition, showed that good results could be found using free, open-source tools without having to use the usually large amounts of training data typical ASR models need to achieve comparable performances. All scripts were run using Google Colab Pro+ with a (A100) GPU.

Wav2Vec 2.0 & CTC

To research the effect of bilingual exposure on word segmentation, I used Wav2Vec 2.0, an open-source Automatic Speech Recognition (ASR) model developed by Meta AI. Wav2Vec 2.0 is a self-supervised model, which means that it learns speech representations directly from audio. Since human infants acquire language primarily by listening rather than through explicit labeling, self-supervised learning offers a similar approach (Baevski, Hsu, Conneau, & Auli, 2022).

Wav2Vec 2.0 learns useful speech representations directly from raw audio data without requiring aligned transcriptions, which mirrors the way humans begin to acquire language through exposure rather than instruction. However, while Wav2Vec alone produces continuous speech representations, a decoding mechanism is required to map these to discrete units such as words or characters.

To handle this, I implemented Connectionist Temporal Classification (CTC) as the decoding layer. CTC is especially well-suited for sequence-to-sequence tasks where the input (audio) and output (character sequences) are not necessarily aligned, and it does not require manual segmentation during training. This is critical for word segmentation tasks, where identifying correct word boundaries is precisely the challenge.

The decision to use CTC was supported by the findings of *Inaguma, Dalmia, Yan, and Sperbe*r (2021), who investigated the effectiveness of CTC in multilingual and cross-lingual Wav2Vec 2.0 setups. Their study demonstrated that CTC, even when used without a language model, can produce competitive results in low-resource conditions, making it particularly relevant to this study, where models were trained on limited amounts of Dutch and German data. Moreover, CTC provides a direct way to evaluate how well a model can segment audio into meaningful units, an essential aspect of this study's focus on bilingual word segmentation performance.

Common Voice

A large amount of speech data was needed to train the models (Synnaeve, et al., 2020). The performance of the model was based on the language, the quality of the data & the variability of the speech (Beech & Swingley, 2023). To train the models, I used Mozilla's

Common Voice dataset, a large, open-source speech corpus that provides speaker-diverse audio samples in multiple languages. Common Voice contains recordings from a wide range of speakers, ensuring that the model learns robust linguistic patterns rather than overfitting to specific accents, voice or speech styles. Since the study required controlled exposure to Dutch and German, Common Voice provided a wide collection of speech samples in both languages, allowing for systematic data allocation. The dataset also includes validated text transcriptions, which are crucial for fine-tuning ASR models and evaluating word segmentation accuracy.

Data preprocessing

Since I used Wav2Vec, which is pretrained on 16kHz audio data, and uses the python CTC segmentation package, which also requires the data to be normalized. Several different tools to preprocess the data were utilized and made sure everything was in the correct format to start training and evaluating the models as efficiently as possible. All preprocessing scripts were executed in Google Colab using Python 3.10

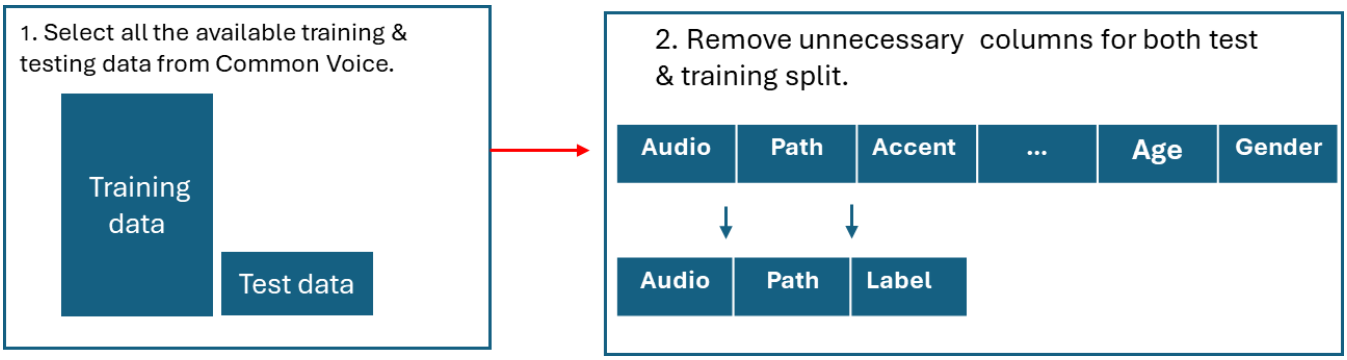

Figure 1: Shows the normalisation of the data for quicker mapping.

Textual data normalization

To make it easier for the CTC tokenizer, it was essential to reduce the amount of unclear textual diversion that the tokenizer could not link to any difference in audio, after all, it is impossible to hear the usage of a comma, capitalization or other special signs. I also changed the data to lists of individual characters, as the CTC tokenizer acquires.

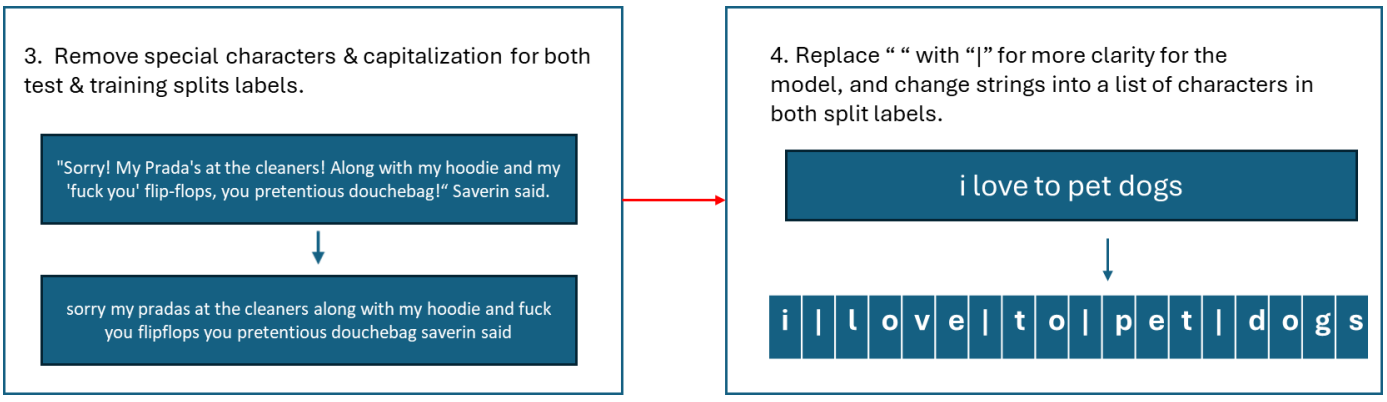

Figure 2: Shows the normalisation of the textual data for the CTC tokenizer.

XLSR Feature Extractor

The extractor normalized and down sampled audio to 16 kHz, as this is the only sample size the model will accept. Padding was added, so all audio samples were the same size, further helping data normalization. The extractor converted waveforms into feature vectors that capture important acoustic patterns, volume and tones.

Because I used a standardized feature extraction process, the model generalized across different speakers and recording conditions. This makes sure that consistent input representation helps minimize variability, allowing for a more controlled and consistent evaluation.

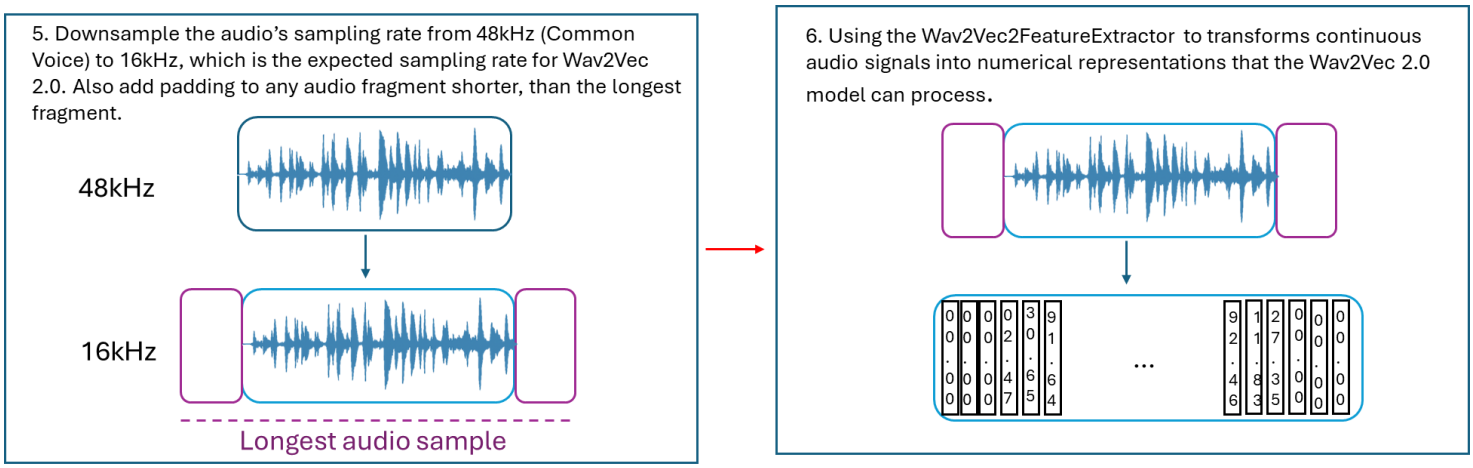

Figure 3: Shows the normalisation of the audio data for the Wav2Vec 2.0 model.

CTC Tokenizer

To process and decode the speech data, I used the Wav2Vec2CTCTokenizer, which is designed for Connectionist Temporal Classification (CTC)-based models. Since CTC models do not specifically predict word boundaries, the tokenizer played an important role in reconstructing text from raw model predictions.

The tokenizer converts raw speech model outputs into character-esque sequences, making sure alignment between spoken and written text were made. It was specifically useful in this study, because it uses a blank token to account for uncertain predictions, allowing for more flexible segmentation in the models that did not get sufficient data. By using a character-based tokenizer, the model learns segmentation patterns without

explicit labelled data, or word level-supervision. This helps us give a clearer view of what the model is predicting, without having to give it exclusively labelled data.

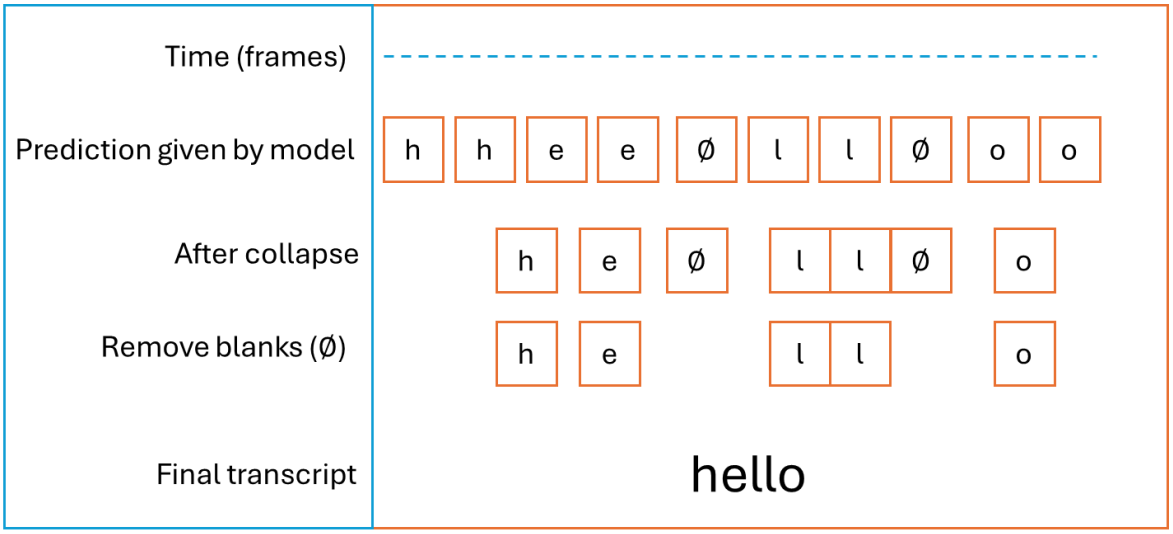

Figure 4: Shows process of the CTC tokenizer, from output of the Wav2Vec model towards its final transcript.

Mixed data models

Initial run

Before training the multilingual models, I conducted an initial model run using only Dutch data. This step was crucial in determining the ideal training parameters and assessing baseline segmentation performance. Loss convergence, which is the stabilization of training loss over time, was monitored closely to determine when further training would no longer give improvements. Training loss measures how well the model learns from the training data, while validation loss assesses its performance on unseen data, which is where I used the test slice. A decreasing training loss with stable validation loss suggests successful learning, whereas an increasing validation loss could indicate overfitting.

To evaluate model accuracy, I also tracked the Word Error Rate (WER), which quantifies the difference between the predicted and actual transcriptions. The WER is a standard metric in speech recognition research and serves as an important benchmark. For reference, the Wav2Vec 2.0 large model, pre-trained on 60,000 hours of LibriVox data and fine-tuned with only 10 minutes of labelled data, achieved a WER of 5.2% on the LibriSpeech clean test set and 8.6% on the other test set (Baevski, Zhou, Mohamed, & Auli, 2020). This table provides context for evaluating the segmentation accuracy of my model.

During the initial model run, I found that training loss stabilized around 10,000 to 12,000 steps, confirming that this range would be appropriate for subsequent bilingual models. Validation loss showed diminishing returns beyond 12,000 steps, suggesting that further training would not significantly improve performance. This informed my decision to standardize training across all models at 12,000 steps to ensure a fair comparison.

This experiment also provided a baseline WER, serving as a control against which I could measure the performance of bilingual models. By making sure that the model processed Dutch correctly and confirming optimal training parameters, I ensured that later experiments could isolate the effect of language exposure. The insights gained from this initial run helped refine hyperparameter selection and reinforced the importance of structured, comparative evaluation in multilingual speech modelling.

Table 1: Progress of Training Loss, Validation Loss and Word Error Rate per 500 steps during the initial run.

Step	Training Loss	Validation Loss	Wer
500	0.785300	0.633843	0.658550
1000	0.299500	0.305755	0.350489
1500	0.244300	0.255516	0.306746
2000	0.206400	0.239935	0.295399
2500	0.134700	0.221388	0.283107
3000	0.151600	0.223319	0.282405
3500	0.127200	0.228786	0.275587
4000	0.120000	0.215997	0.270990
4500	0.122200	0.211115	0.266432
5000	0.126100	0.216379	0.268613
5500	0.118300	0.224526	0.267240
6000	0.109700	0.225219	0.262185
6500	0.130400	0.210010	0.256818
7000	0.119300	0.208931	0.259292
7500	0.083400	0.213268	0.257383
8000	0.087600	0.209142	0.253331
8500	0.075900	0.204434	0.250662
9000	0.043600	0.210561	0.251071
9500	0.074700	0.202508	0.248880
10000	0.053700	0.210251	0.251705
10500	0.064900	0.203347	0.247497
11000	0.054900	0.201678	0.244156
11500	0.053700	0.201886	0.242432
12000	0.087900	0.206302	0.241886

Mixed data models

Catastrophic forgetting is a phenomenon where neural networks rapidly lose previously acquired knowledge when introduced to new data (French, 1999). This issue arises when a model trained on one language is fine-tuned on another without mechanisms to retain the prior knowledge, often leading to a significant decline in performance on the original language.

To mitigate this, I used a rehearsal-based approach, which is where both Dutch and German data were continuously presented to the model in different proportions throughout training. Instead of sequentially training on one language and then another, I trained five separate models with fixed ratios of Dutch and German exposure. Despite the varying ratios, each model was trained on the same amount of time of total audio, ensuring that all models were exposed to an equal amount of linguistic data while varying the degree of the languages.

By interleaving Dutch and German throughout training, the model continuously reinforced prior knowledge while incorporating new linguistic patterns. This approach aligns with established research on mitigating catastrophic forgetting in neural networks, which suggests that repeated exposure to prior data helps maintain previously learned representations. The effectiveness of this method has been demonstrated in multilingual ASR research by Baevski et al. (2022), where exposure to multiple languages improves the model's ability to generalize across languages while preserving performance in each individual language.

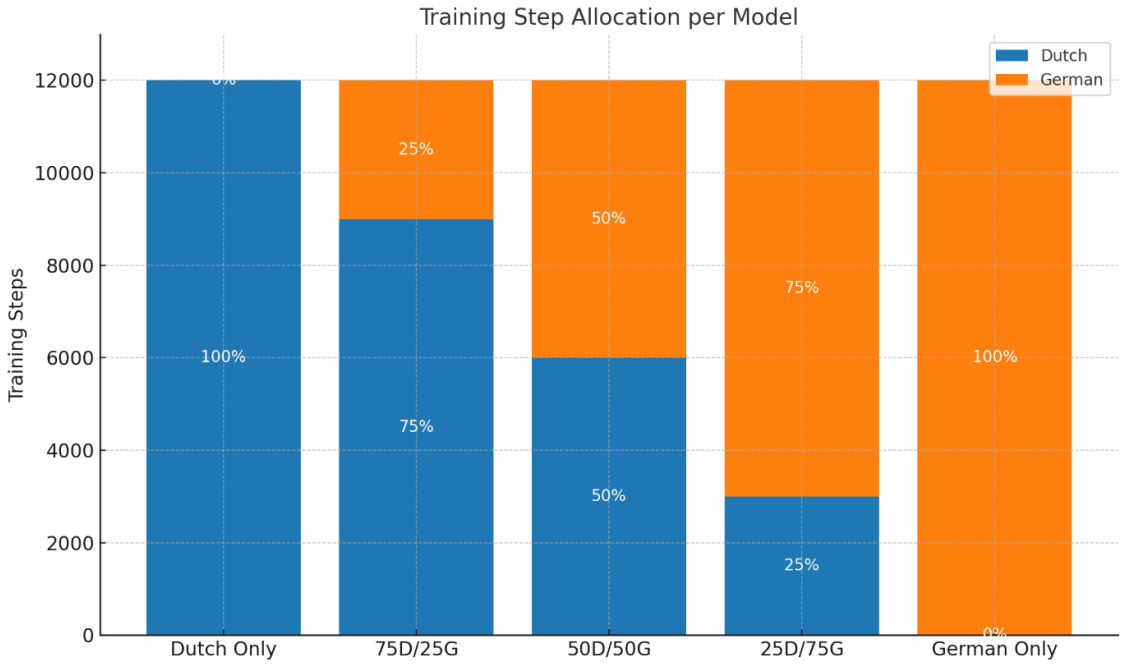

Figure 5: Data distribution per model.

To get a clear view on how the distribution of the languages used in the training data influenced the model, I made 5 models with different data distributions as seen in the graph above.

Model training

To implement this research, I used existing open-source code from publicly available repositories, including the CTC segmentation toolkit by Kürzinger et al (2022) and the Hugging Face implementation of Wav2Vec 2.0. These resources provided a great start for setting up ASR training and evaluation, which I then adapted to fit the specific requirements of this study.

The provided scripts were updated to ensure compatibility with the latest versions of Hugging Face's Transformers library and the Common Voice dataset format.

Custom scripts were written to balance bilingual data exposure while preventing catastrophic forgetting in the multilingual training models.

The models were trained using Google Colab Pro+, leveraging their NVIDIA A100 GPU for faster processing times. The A100 GPU allowed for larger batch sizes and longer training runs, reducing convergence time while keeping a high computational efficiency. Training

scripts were optimized with mixed precision training (fp16) to increase the speed and reduce memory consumption, as this is limited to 200GB in Google Colab.

Evaluation

After training, the models were evaluated to determine their word segmentation performance using CTC-segmentation, an open-source tool developed for aligning ASR outputs with reference transcriptions.

TIMIT

The segmentation accuracy was assessed using TIMIT, a widely used dataset containing phonetically diverse sentences with manually annotated word boundaries, serving as a "golden rule" for comparison (Garofolo, et al., 1993). TIMIT is similar to Common Voice in terms of diversity of speakers, accents and quality of audio, making it a good match for the models. However, I was limited by the free sample, which consisted of only 50 test sentences.

By applying CTC-segmentation, I was able to align the predicted word boundaries with the ground truth annotations in TIMIT, enabling a direct and objective performance assessment.

Mean Absolute Error

To evaluate the accuracy of the model's predicted word boundaries, a custom Mean Absolute Error (MAE) metric was computed in python based on the alignment between predicted and true word boundaries. The method is illustrated in the accompanying figure.

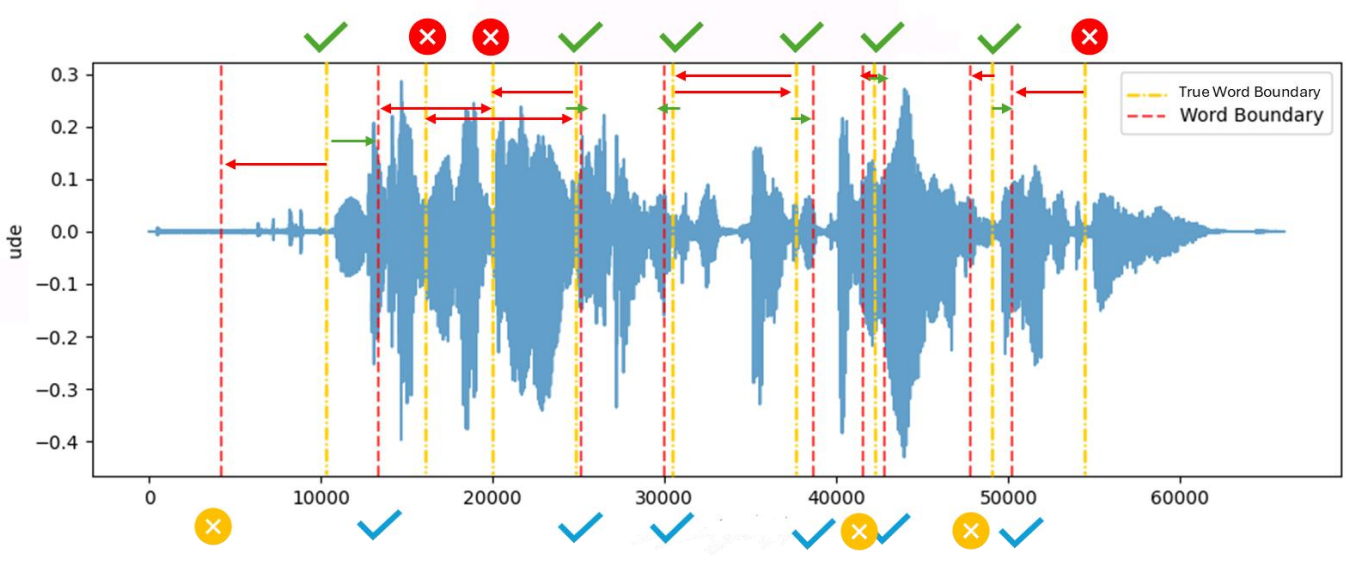

Figure 6: Explanation of the Mean Absolute Error calculations. Where each true boundary is seen to "claim" a closest predicted word boundary (green arrow) and reject any boundaries that are further away or that are not predicted word boundaries.

Each true word boundary (indicated by yellow dashed lines) was assigned the closest predicted boundary (red dashed lines) on either its left or right. A predicted boundary could only be assigned to one true boundary; in cases of conflict, it remained assigned to the true boundary to which it was closest. This ensured a one-to-one matching process, which prevented false positives. True boundaries that found a match were marked as successful, while unmatched true boundaries were treated as errors.

Similarly, predicted boundaries were evaluated based on whether they were matched to any true boundary. Both unmatched true and predicted boundaries contributed to the final error calculation by incurring a penalty. This approach allowed the metric to account for both missed word boundaries (false negatives) and spurious predictions (false positives).

The MAE was then computed as the average of the absolute distances in milliseconds between each matched pair of boundaries, along with additional penalty values for unmatched items. All distances were normalized, resulting in final MAE values between 0 and 1. A value of 0 indicates perfect prediction alignment, while values closer to 1 reflect greater deviation and/or more unmatched boundaries.

In addition to numerical evaluation, I generated visual graphs using matplot to illustrate the segmentation differences between the model's predicted word boundaries and TIMIT's confirmed word boundaries. These graphs plotted the predicted and actual segmentation points, providing a good representation of how closely each model aligned with the expected boundaries.

Results

The results of the evaluation show an interesting contrast between the monolingual and bilingual models in their ability to segment English speech. The Dutch-only model exhibited the lowest Mean Absolute Error (MAE), indicating that it aligned most closely with the golden rule segmentation provided by TIMIT.

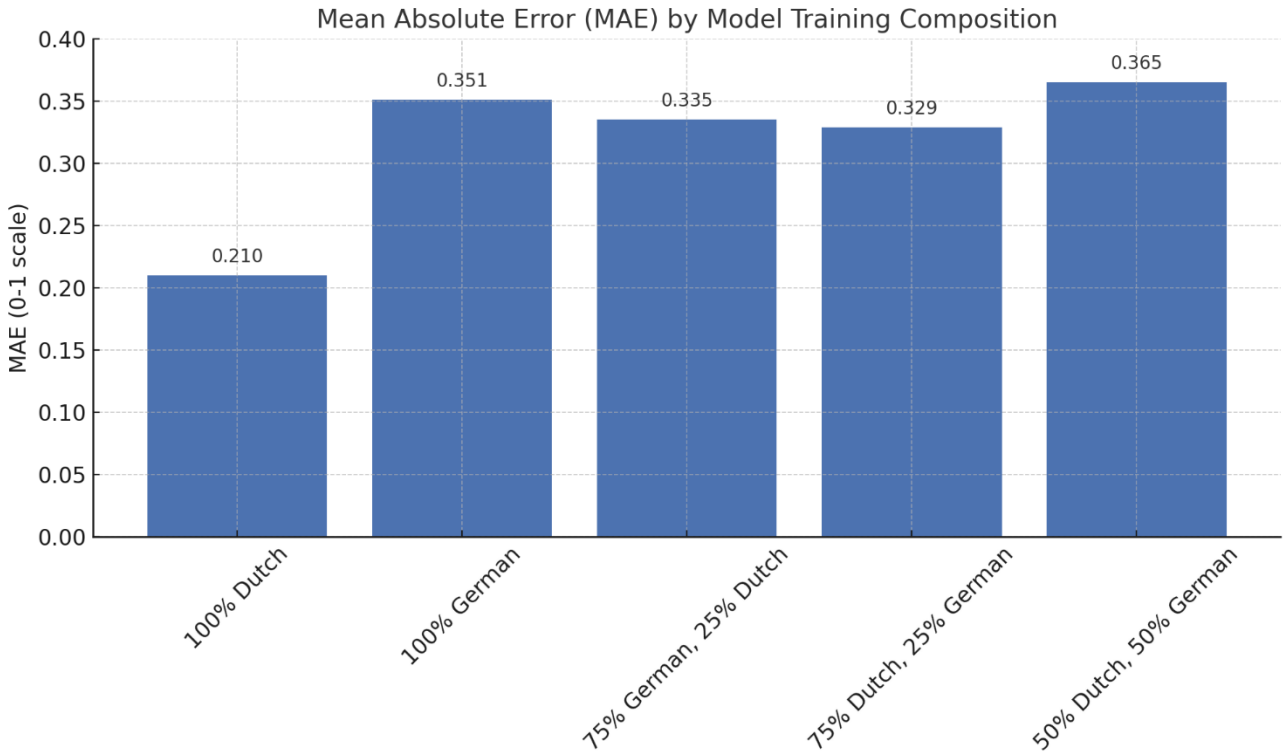

Figure 7: Mean Absolute Error, based on 50 test sentences in English, on our 5 different ratio models.

In contrast, the 50%-50% Dutch-German model performed the worst, displaying the highest MAE. The models with 75%-25% and 25%-75% Dutch-German distributions fell in the middle, with increasing German exposure correlating with slightly higher segmentation errors. Although this did not significantly impact the performance, making it a near symmetrical result.

A closer examination of the visual graphs further reinforced these findings. The Dutch-only model consistently placed segmentation boundaries closest to the reference points, whereas the bilingual models, particularly the 50%-50% model, demonstrated misalignment and inconsistent boundary placement. We can also see that the better performing models, guessed a closer number of words than the confirmed amount, whereas the worst performing model encountered on average, 5 less words than the best performing models. This finding suggests that while exposure to multiple languages may enhance general linguistic flexibility in some cases, it does not necessarily translate to improved segmentation performance in an unseen language

Model trained exclusively on Dutch data

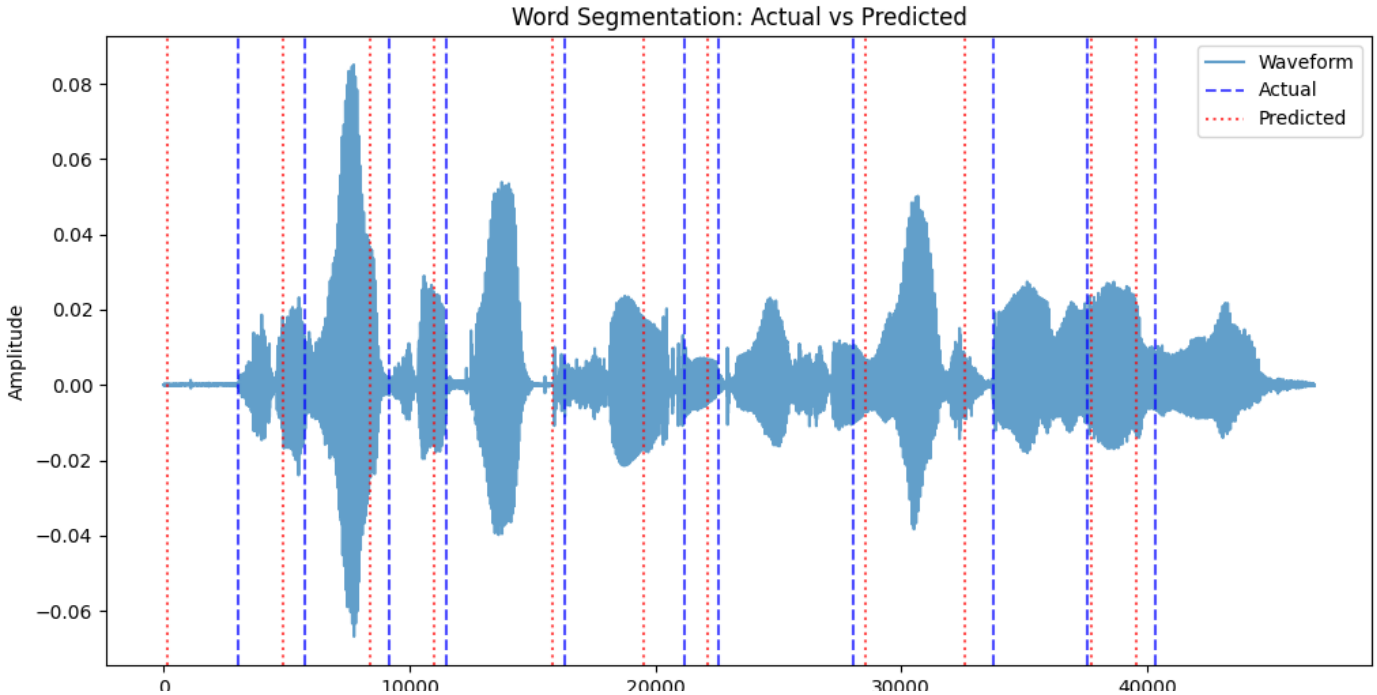

Figure 8: Word segmentation result on a test sentence from the Dutch only model.

The Dutch-only model produced a transcript that most closely matched the original sentences, both in content and segmentation boundaries. While it did not transcribe the sentences perfectly, it predicted a structure that was rhythmically and semantically aligned with the true words, and it was the only model to consistently predict close to the correct number of word boundaries in the test set. This model achieved the lowest MAE (0.21), suggesting high temporal precision in identifying word transitions. In the example sentence above the Dutch model predicted the transcript: "*zie hed je dak zi in grisie as worden al jeer*". Which compared to the original transcript, "She had your dark suit in greasy wash water all year", is eerily close. Another thing the Dutch-only model excelled at, was correctly predicting words, with quite a few correctly predicted words in the English language, even if it did not get any English training data.

Model trained exclusively on German data

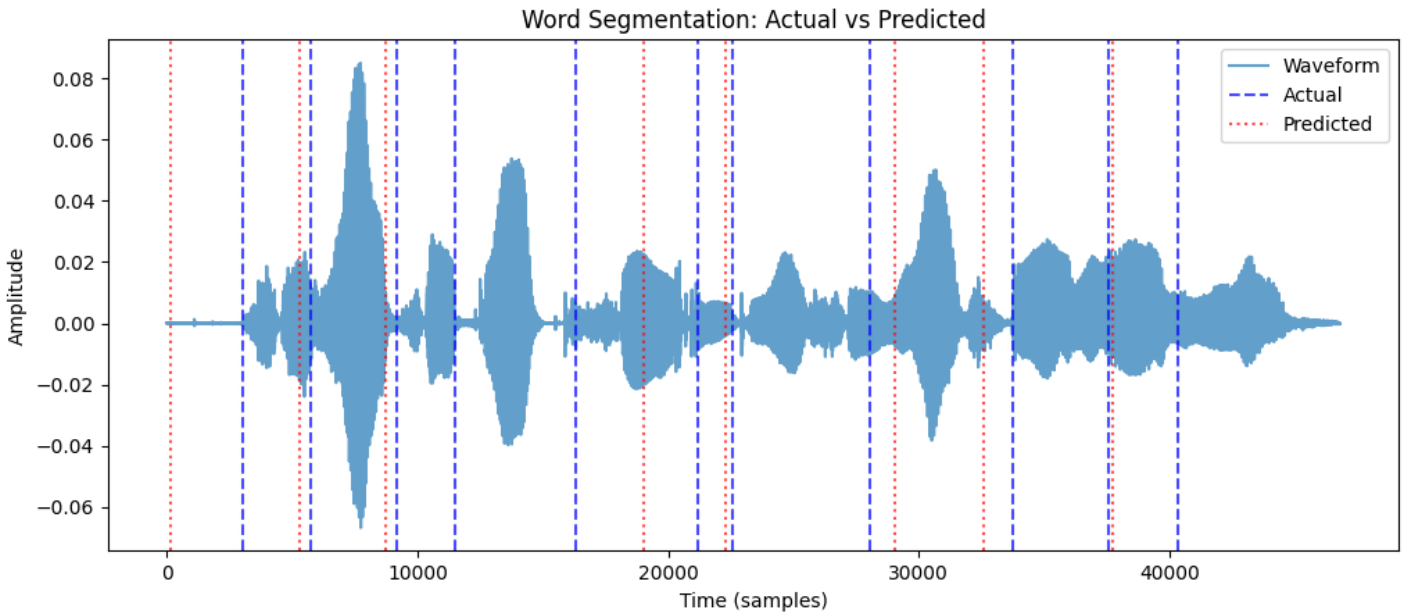

Figure 9: Word segmentation result on a test sentence from the German only model

The German-only model performed notably worse, with a higher average MAE of 0.351. Its predicted transcripts often contained fewer words than the original and exhibited greater deviations in both pronunciation and segmentation. In the example sentence, it produced: "*chie het jedox se in grecey whwade oiere.*" While some individual sounds are recognizable, the word boundaries were less precise, often merging adjacent words or skipping function words. The segmentation graph shows multiple true boundaries left unmatched or paired with distant predictions, contributing to the elevated error score.

Model trained on a majority of German, and some Dutch data

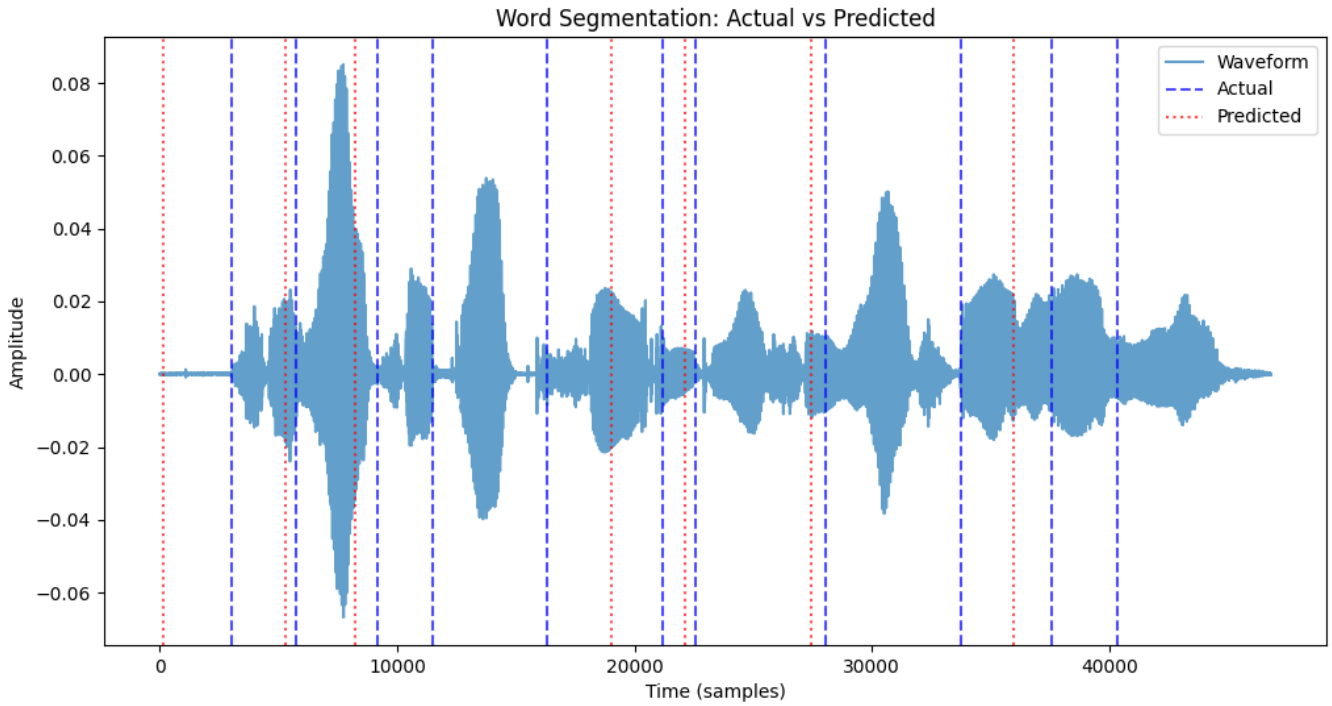

Figure 10: Word segmentation result on a test sentence from the 75% German and 25% Dutch exposure model

This model showed modest improvement over the fully German model, with an average MAE of 0.335. Its predicted transcripts still demonstrated signs of under-segmentation, but occasionally captured content words more clearly. For the test sentence, its output was: "*shie ha jadocu in gresy wacwal roer*." Though still imperfect, it shows some influence from Dutch training data in how it structures the phrase. The segmentation graph supports this, with a few closely predicted boundary matches present, but many misalignments and unmatched predictions remaining.

Model trained on a majority of Dutch, and some German data

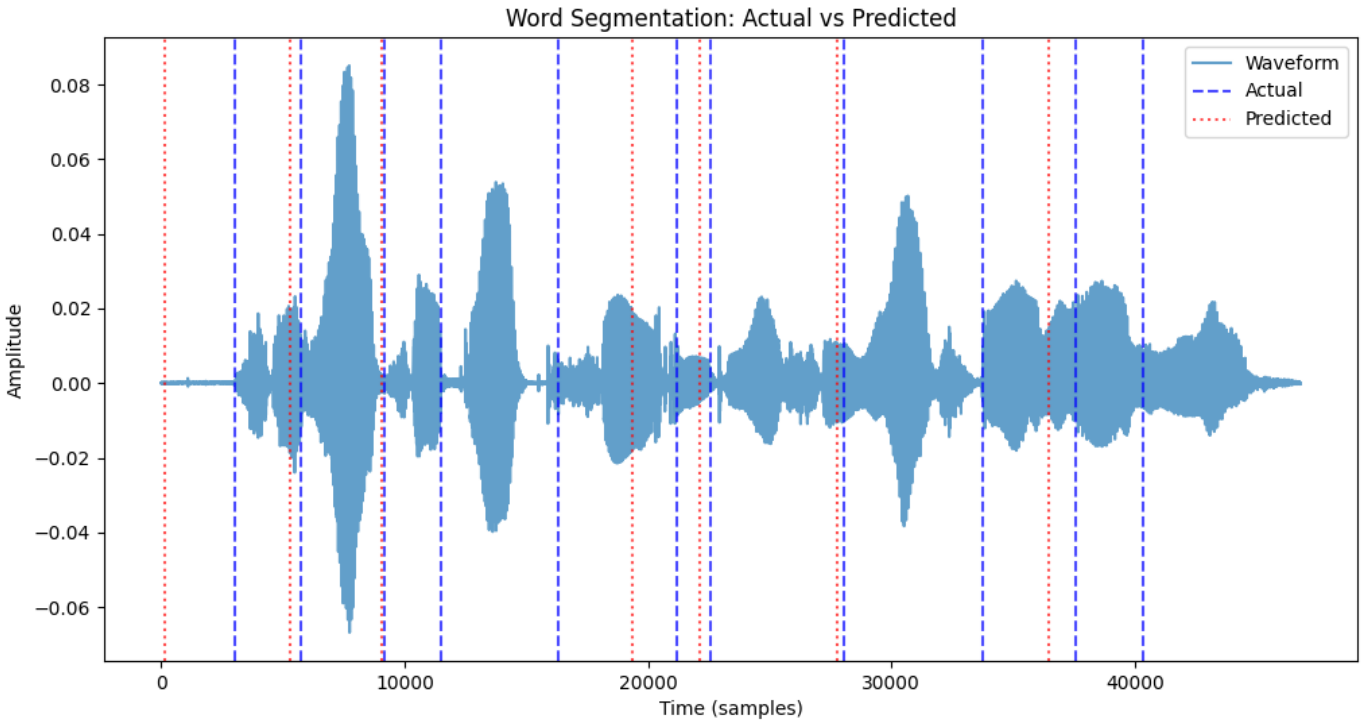

Figure 11: Word segmentation result on a test sentence from the 75% Dutch and 25% German model.

The 75%-25% Dutch-German model outperformed the German-heavy and balanced models, achieving an average MAE of 0.329. The predicted transcript for the example sentence was: "*cie hert jedacsu in grisi wacwor aljeer*." This result retains a certain English-like structure. In the segmentation graph, a few true boundaries are matched. These results suggest that dominant exposure to Dutch, even with some German interference, still yields segmentation strategies somewhat closely aligned with English. Interestingly, this graph shows that, compared to the German-majority model, the predicted boundaries are almost perfectly aligned, with the Dutch-majority model only offering a slight improvement toward the true boundaries.

Model trained on both German and Dutch equally

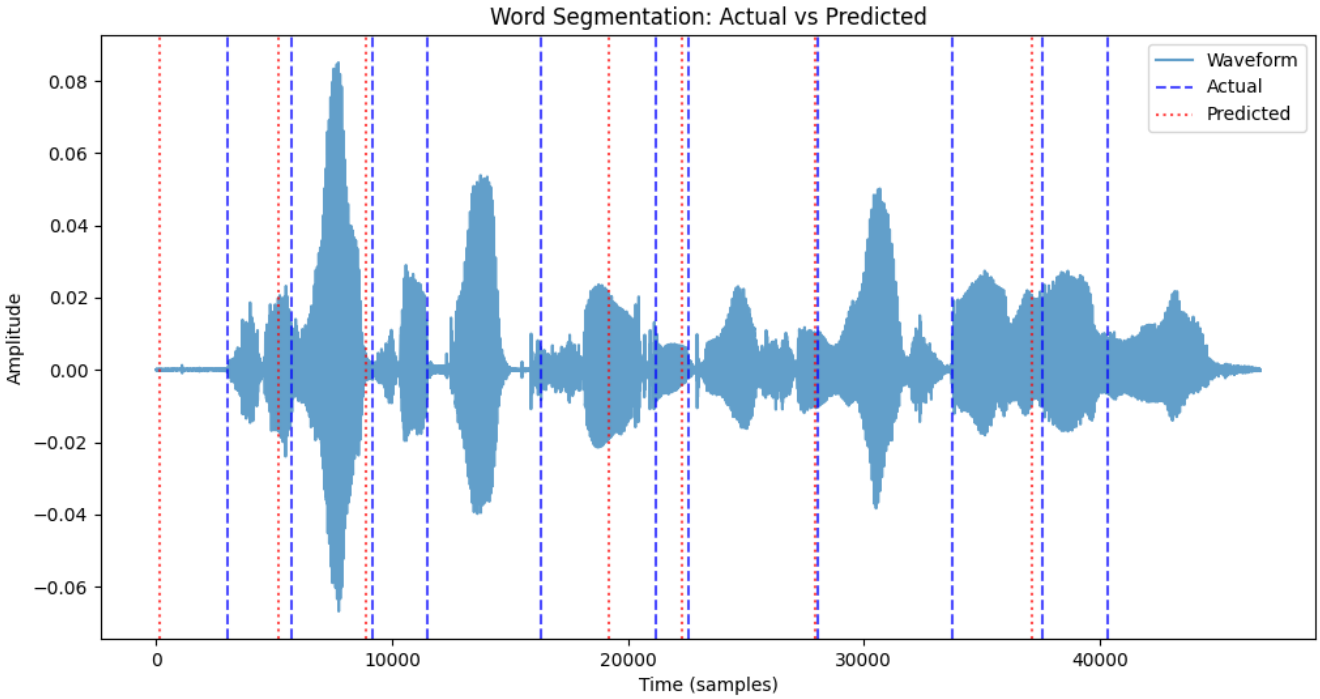

Figure 12: Word segmentation result on a test sentence from the 50% German and 50% Dutch model

Despite being trained on the same total amount of data, the 50%-50% model performed the worst across the test set, with the highest average MAE of 0.365. Its predicted transcripts frequently under-segmented the sentences and distorted the phonetic structure. In the example sentence, the output was: "*zihed je dakte ingrissie was voorden aljeer.*" While partially recognizable, the segmentation was inconsistent, and the number of predicted boundaries was often too low. The segmentation graph confirms this, with sparse boundary predictions and numerous unmatched lines.

Final results

The Dutch-only model achieved the lowest MAE (0.21), while the 50%-50% model performed the worst (MAE: 0.365). Bilingual models with partial Dutch exposure (75%-25% and 25%-75%) performed better than the German-only model but worse than the Dutch-only model. Across all models, the number of predicted words was consistently lower than the number of actual words in each English sentence. The worst-performing model (50%-50%) predicted five fewer words, on average, than were included in the official transcript.

Discussion

The results of this study demonstrate a clear pattern in how different language exposure ratios influence word segmentation in an unseen language. The fully Dutch model emerged as the best-performing model, while the 50%-50% Dutch-German model performed the worst. The intermediate models, 75%-25% and 25%-75%, performed slightly better than the fully German model, but still fell short of the Dutch-only model's segmentation accuracy. These results somewhat support the original hypothesis that monolingual training in a language would result in better segmentation performance than bilingual training under limited data conditions. The Dutch-only model outperformed all others, while the 50%-50% model performed the worst. This directly addresses the research question and suggests that bilingual exposure does not necessarily enhance segmentation performance in unseen languages when exposure time is limited.

However, the fact that both mixed models with a majority in one language performed (slightly) better than the fully German model suggests that having (some) training in a language phonetically closer to English, does improve performance in this task. These findings also directly answer the research question by showing that bilingual exposure, without additional training time, can both improve or deteriorate segmentation accuracy, depending on the secondary language it is exposed to.

Why did the fully Dutch model perform the best?

The Dutch-only model achieved the lowest Mean Absolute Error (MAE), meaning its predicted word boundaries aligned most closely with the correct English segmentation points. This is likely due to the phonetic and rhythmic similarities between Dutch and English. Both languages share similar stress-timed prosody, meaning they tend to group syllables in a rhythmically similar way (Blevins, 2006). Additionally, Dutch and English have a high degree of lexical overlap, meaning that Dutch-trained models may have been able to recognize words and word boundaries in English more effectively than models trained on German.

Furthermore, because this model received 100% of its training data in one language, it developed a deep understanding of Dutch segmentation cues without interference from a second language. This clarity in linguistic pattern recognition likely contributed to its superior performance in English segmentation.

Why did the 50%-50% Dutch-German model perform the worst?

The poorest performance came from the 50%-50% model, which exhibited the highest MAE. This finding aligns with previous research suggesting that training a model with limited data per language can hinder its ability to generalize effectively (Liu, 2023). Since the total amount of training data was the same across all models, dividing the data equally between Dutch and German meant that neither language received sufficient exposure for the model to fully grasp its segmentation rules.

Essentially, the 50%-50% model had half the training material per language compared to the monolingual models, forcing it to learn two linguistic systems without mastering either one. This could explain why it struggled the most with English segmentation, it lacked a dominant phonetic framework to rely on.

Another possible reason is interference between Dutch and German linguistic patterns. Although these languages are both West Germanic, their phonetic constraints, syllable structures, and rhythm patterns are not identical. The model likely had difficulty reconciling conflicting segmentation cues from both languages, leading to greater errors in word boundary predictions. These results offer computational evidence for Adriaans' (2024) theory of segmentation interference caused by overlapping phonetic systems in bilingual learners.

Why did the 75%-25% and 25%-75% models perform better than the fully German model?

The fully German model performed worse than the Dutch model, but slightly better than the 50%-50% model. Interestingly, the 75%-25% (Dutch-German) and 25%-75% (Dutch-German) models both outperformed the fully German model, despite having exposure to two languages. This suggests that even a small percentage of Dutch training data provided segmentation advantages over training in German alone.

There are a few possible explanations for this, Dutch is phonetically closer to English. To the point that even a 25% exposure to Dutch may have given these models better segmentation intuition compared to a purely German-trained model. This is particularly relevant given that English shares more vocabulary and phonetic structures with Dutch than German.

It's also suggested that partial bilingual training can be beneficial. Some research suggests that partial bilingual exposure (where one language dominates) can help models generalize better than either monolingual training or perfectly balanced bilingual training (Toshniwal, et al., 2017). This is because the dominant language provides a strong foundation, while the secondary language introduces variability that enhances pattern recognition.

These findings suggest that bilingual exposure can be helpful for segmentation tasks, but only when there is a dominant language to provide stability, and the secondary language better reflects the test language.

Why did models predict fewer words than the English sentence contained?

Across all models, a consistent trend emerged: they predicted fewer words per sentence than the actual number of words in English. For example, when an English sentence contained 11 words, models typically predicted around 8 words. This suggests that the models tended to over-merge words rather than insert unnecessary boundaries.

One possible explanation for this is phonological contraction, both Dutch and German have different word boundary cues than English, particularly in terms of function words

and weak syllables. English frequently reduces unstressed syllables and function words (e.g., "to," "the," "of"), while Dutch and German tend to preserve clearer phonetic boundaries of such words. The models may have learned to segment speech in a way that favoured stronger, more clearly articulated syllables, skipping over reduced function words, as seen in the English language.

Expected results

The Dutch-only model performed the best. Given its phonetic similarity to English, this result was anticipated. The 50%-50% model performed worst, as dividing training data between two languages could lead to weaker language modelling capabilities.

Unexpected results:

The 75%-25% and 25%-75% models performed better than the fully German model. Their relatively strong performance suggests that even limited bilingual exposure can significantly aids foreign language segmentation if it is phonetically and rhythmically similar to the test language.

Models predicted fewer words than actually present in English, indicating over-merging tendencies rather than random segmentation errors.

Limitations and open questions

While this study gave somewhat expected results, some limitations should be acknowledged.

Each model was trained on the same amount of data, but this amount may not have been enough to fully develop bilingual segmentation skills, particularly for the 50%-50% model. If, let's say, the model got 10 times the amount of data it currently got, it might have had time to properly generalise segmentation rules for both languages.

The test set for this research was quite limited, with only 50 sentences. Evaluating segmentation accuracy on a larger dataset might have provided a clearer picture of generalization performance.

The drop-off point between 75%-25% and 50%-50% remains unclear. Would a 60%-40% model perform more like the Dutch-dominant models or more like the 50%-50% model?

The results gave a favourable result for Dutch, linking these results to the fact that Dutch is phonetically & rhythmically closer to English. Would research tested entirely on an unrelated language (e.g., Japanese or Arabic) show more favourable results for models with broader training data?

Conclusion

This study investigated how different Dutch-German language exposure ratios influence word segmentation performance in an unseen language (English) using Wav2Vec 2.0. The results revealed that linguistic similarity and the proportion of training data dedicated to a single language significantly affect segmentation accuracy. The fully Dutch model achieved the best performance, demonstrating that training on a phonetically closer language to English provides a strong foundation for word segmentation tasks in said language. Conversely, the 50%-50% bilingual model performed the worst, indicating that an equal split between two languages can dilute a model's ability to generalize segmentation rules. The 75%-25% and 25%-75% models performed slightly better than the fully German model, suggesting that even a small percentage of Dutch exposure enhanced performance.

Key findings and their implications

The most important outcome of this study is that balanced bilingual training (50%-50%) led to the poorest performance, while monolingual training in a language structurally closer to English (Dutch) produced the most accurate segmentation. This finding supports prior research suggesting that splitting training data between two languages without significantly increasing the total amount of exposure weakens a model's linguistic competence in either language. Essentially, by dividing resources equally between Dutch and German, the 50%-50% model never acquired a strong foundation in either language, leading to reduced segmentation accuracy when tested on English.

The success of the fully Dutch model reinforces the importance of linguistic similarity in word segmentation tasks. Dutch shares more phonetic rules, rhythmic structures, and syllable patterns with English than German does. This suggests that, in speech recognition models, exposure to a linguistically adjacent language enhances the ability to transfer segmentation strategies to a third language.

An unexpected, but important finding was that partial bilingual exposure (75%-25% or 25%-75%) resulted in better segmentation accuracy than the fully German model. While this result was minimal it still suggests that exposure to a dominant language with a smaller influence from a secondary language can provide certain benefits. The presence of some Dutch exposure appears to have given these models an advantage over the German-only model, though not enough to surpass the fully Dutch model. This aligns with theories in bilingual language acquisition that suggest some degree of bilingualism enhances flexibility.

Another notable trend was that all models under-segmented sentences, meaning they predicted fewer words than actually present in the English sentences. Rather than randomly misplacing boundaries, they consistently merged words together, suggesting that their segmentation strategies favoured strong syllables while overlooking function words and weak syllables. This behaviour reflects cross-linguistic differences in prosodic structure, where languages like Dutch and German emphasize stronger syllables more than English does.

Finals thoughts & future research directions

This thesis contributes to the understanding of multilingual ASR by showing that linguistic similarity and training data distribution critically shape segmentation performance in unseen languages. These results challenge assumptions about the benefits of bilingual exposure and highlight the importance of data balance in multilingual training strategies. Future work should explore these dynamics with larger datasets, analyse intermediate ratios to pinpoint the threshold at which bilingual training begins to harm segmentation performance and measure the performance of these models against on unrelated languages. I further made a custom code to record, and manually segment sentences in any language, to further test the model's performance against.

References

Adriaans, F. (2024). Computational Approaches to Bilingual Phonetics and Phonology. *The Cambridge Handbook of Bilingual Phonetics and Phonology*, 126-144.

Auwera, J., & Noel, D. (2011). Raising: Dutch Between English and German. *Journal of Germanic Linguistics*, 1-36.

Baevski, A., Hsu, W. N., Conneau, A., & Auli, M. (2022, May 2). *Unsupervised Speech Recognition*. From Facebook AI: https://arxiv.org/pdf/2105.11084

Baevski, A., Scheider, S., Hsu, W. N., Conneau, A., Zhou, H., Collobert, C., . . . Auli, M. (2021, October 25). *Wav2vec.* From Meta AI: https://ai.meta.com/research/publications/unsupervised-speech-recognition/

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: a framework for self-supervised learning of speech representations. *34th International Conference on Neural Information Processing Sustems* (pp. 12449 - 12460). Red Hook: Curran Associates Inc.

Beech, C. (2023, June). *Consequences of phonological variation for algorithmic word segmentation.* From ScienceDirect: https://linkinghub.elsevier.com/retrieve/pii/S0010027723000355

Beech, C., & Swingley, D. (2023, February). *Consequences of phonological variation for algorithmic word segmentation*. From Science Direct: https://www.sciencedirect.com/science/article/abs/pii/S0010027723000355

Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics*, 531-573.

Buccini, A. H. (2024, july 19). *West Germanic languages*. From Encyclopedia Britannica: https://www.britannica.com/topic/West-Germanic-languages/German

Cutler, A. &. (1989). *The role of strong syllables in segmentation for lexical access.* From APA PsycNet: https://psycnet.apa.org/record/1988-19602-001

Deanda, S., Arias-Trejo, N., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2015). Minimal second language exposure, SES, and early word comprehension: New evidence from a direct assessment. *Cambridge University Press*, 162–180.

Dupoux, E. (2018, February 14). *Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner.* From Arxiv, Cornell University: https://arxiv.org/abs/1607.08723

Flege, J. E., MacKay, I. R., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *Journal of the Acoustical Society of America*, pp. 2973-2987.

French, R. (1999, May). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, pp. 128-135.

Gabriel Synnaeve, Q. X. (2020, July 15). *End-to-end ASR: from Supervised to Semi-Supervised Learning with Modern Architectures.* From Arxiv, Cornell University: https://arxiv.org/abs/1911.08460

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, N. L. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus.* From Linguistic Data Consortium: https://catalog.ldc.upenn.edu/LDC93S1

Hirofumi Inaguma, S. D. (2021, september 27). *FAST-MD: FAST MULTI-DECODER END-TO-END SPEECH TRANSLATION.* From paperswithcode: https://arxiv.org/pdf/2109.12804v1

Inaguma, H., Dalmia, S., Yan, B., & S., W. (2021, september 27). *Fast-MD: Fast Multi-Decoder End-to-End Speech Translation with Non-Autoregressive Hidden Intermediates*. From ARVIX: https://arxiv.org/abs/2109.12804

Kovács, A. M., & Mehler, M. (2009, April 13). *Cognitive gains in 7-month-old bilingual infants*. From PubMed: https://pubmed.ncbi.nlm.nih.gov/19365071/

Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T., & Rigoll, G. (2020, september 29). *CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition*. From Springer Nature Link: https://link.springer.com/chapter/10.1007/978-3-030-60276-5_27

Liu, Z. S. (2023). Investigationo datapartitioning strategies for crosslinguistic low-resource ASR evaluation. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (p. 44). Dubrovnik: Association for Computational Linguistics.

Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T., Kamo, N. & Rigoll, G., (2022, february 2). *CTC - segmentation*. Github: https://github.com/lumaku/ctc-segmentation

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or Syllable? Speech Segmentation in Japanese,. *Journal of Memory and Language,*, Pages 258-278,.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science.*, 1926-1928.

Sebastián-Gallés, N., & Bosch, L. (2009). Developmental shift in the discrimination of vowel contrasts in bilingual infants: is the distributional account all there is to it? *Devloped Science*, 874-887.

Singh, L., & Foong, J. (2022). ilingual infants' speech segmentation strategies: A comparative look at monolingual and bilingual learning environments. *Developmental Psychology*, 209-225.

Swingley, D., & Algayres, R. (2024, March 25). *Computational Modeling of the Segmentation of Sentence Stimuli From an Infant Word-Finding Study.* From Wiley, Online Library: https://onlinelibrary.wiley.com/doi/full/10.1111/cogs.13427

Synnaeve, G., Qiantong, X., Kahn, J., Likhomanenko, T., Grave, E., Pratap, V., . . . Collobert, R. (2020, July 15). *END-TO-END ASR: FROM SUPERVISED TO SEMI-SUPERVISED*. From ARVIX: https://arxiv.org/pdf/1911.08460

Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., & Rao, K. (2017). *Multilingual Speech Recognition With A Single End-To-End Model*. Chicago: Toyota Technological Institute at Chicago.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, pp. 1-25.