UPPSALA
UNIVERSITET

Automatic Annotation of Speech: Exploring Boundaries within Forced Alignment for Swedish and Norwegian

Klaudia Biczysko

Uppsala University
Department of Linguistics and Philology
Master Programme in Language Technology
Master's Thesis in Language Technology, 30 ECTS credits

June 21, 2022

Supervisors:
Johan Sjons, Uppsala University
Carl Dehlin, Conversy

Abstract

In Automatic Speech Recognition, there is an extensive need for time-aligned data. Manual speech segmentation has been shown to be more laborious than manual transcription, especially when dealing with tens of hours of speech. Forced alignment is a technique for matching a signal with its orthographic transcription with respect to the duration of linguistic units. Most forced aligners, however, are language-dependent and trained on English data, whereas under-resourced languages lack the resources to develop an acoustic model required for an aligner, as well as manually aligned data. An alternative solution to the training of new models can be cross-language forced alignment, in which an aligner trained on one language is used for aligning data in another language.

This thesis aimed to evaluate state-of-the-art forced alignment algorithms available for Swedish and test whether a Swedish model could be applied for aligning Norwegian. Three approaches for forced aligners were employed: (1) one forced aligner based on Dynamic Time Warping and text-to-speech synthesis Aeneas, (2) two forced aligners based on Hidden Markov Models, namely the Munich AUtomatic Segmentation System (WebMAUS) and the Montreal Forced Aligner (MFA) and (3) Connectionist Temporal Classification (CTC) segmentation algorithm with two pre-trained and fine-tuned Wav2Vec2 Swedish models.

First, small speech test sets for Norwegian and Swedish, covering different types of spontaneousness in the speech, were created and manually aligned to create gold-standard alignments. Second, the performance of the Swedish dataset was evaluated with respect to the gold standard. Finally, it was tested whether Swedish forced aligners could be applied for aligning Norwegian data. The performance of the aligners was assessed by measuring the difference between the boundaries set in the gold standard from that of the comparison alignment. The accuracy was estimated by calculating the proportion of alignments below a particular threshold proposed in the literature.

It was found that the performance of the CTC segmentation algorithm with Wav2Vec2 (VoxRex) was superior to other forced alignment systems. The differences between the alignments of two Wav2Vec2 models suggest that the training data may have a larger influence on the alignments, than the architecture of the algorithm. In lower thresholds, the traditional HMM approach outperformed the deep learning models. Finally, findings from the thesis have demonstrated promising results for cross-language forced alignment using Swedish models to align related languages, such as Norwegian.

Contents

Acknowledgments 5

1. Introduction 6
   1.1. Purpose & Research Questions . . . 6
   1.2. Outline . . . 7

2. Background and Related Work 8
   2.1. Scandinavian Languages & Mutual Intelligibility . . . 8
   2.2. Automatic Speech Recognition . . . 10
   2.3. Toolkits & Algorithms . . . 11
      2.3.1. HTK . . . 11
      2.3.2. Kaldi . . . 11
      2.3.3. Dynamic Time Warping . . . 11
      2.3.4. CTC Segmentation . . . 12
   2.4. Forced Alignment . . . 13
      2.4.1. Description . . . 13
      2.4.2. Munich AUtomatic Segmentation System . . . 14
      2.4.3. The Montreal Forced Aligner . . . 15
      2.4.4. Aeneas . . . 16
      2.4.5. Wav2Vec2+CTC segmentation . . . 17
   2.5. Related work . . . 18

3. Data 20
   3.1. Challenges . . . 20
   3.2. Data Specification . . . 20
   3.3. Corpora . . . 21
   3.4. Data Preprocessing . . . 22

4. Methodology 24
   4.1. Gold Standard . . . 24
   4.2. Testing the-off-shelf forced aligners . . . 25
      4.2.1. WebMAUS . . . 25
      4.2.2. The Montreal Forced Aligner . . . 25
      4.2.3. Aeneas . . . 26
      4.2.4. Wav2Vec2+CTC segmentation . . . 26
   4.3. Evaluation metrics . . . 27

5. Results for the Swedish alignment 29
   5.1. General outline of the results . . . 29
   5.2. Evaluations and comparisons (word-level) . . . 30
      5.2.1. Semi-spontaneous speech data . . . 31
      5.2.2. Spontaneous speech data . . . 35
      5.2.3. Read speech data . . . 39
      5.2.4. Summary . . . 42
   5.3. Evaluations and comparisons (utterance-level) . . . 42
      5.3.1. Semi-spontaneous speech data . . . 44

5.3.2. Spontaneous speech data 48
5.3.3. Read speech data 52
5.3.4. Summary 55
5.4. Qualitative comparison 55

6. Results for the Norwegian alignment 58
6.1. General outline of the results 58
6.2. Evaluations and comparisons 58
6.2.1. Semi-spontaneous speech data 59
6.2.2. Spontaneous speech data 64
6.2.3. Read speech data 68
6.2.4. Summary 71
6.3. Qualitative comparison 71

7. Discussion 73

8. Conclusions 80

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Johan Sjons, and thank him for the dedicated support and guidance, as well as for reviewing my progress constantly and providing invaluable feedback. I have learned a lot. I also wish to thank Carl Dehlin for his practical suggestions and insightful discussions. I am grateful to Conversy for the possibility of writing my thesis with them and the opportunity to be a part of the team. Special thanks to Thea Tollersrud, who helped me with the Norwegian data. Finally, I would like to thank my parents, grandparents, closest friends and my partner for believing in me and for all of their support.

1. Introduction

While watching a video with subtitles, we expect the transcription to be temporally aligned to the audio. The correct temporal alignment of speech and text not only allows us to better experience listening to music with captions, watching movies or recorded lectures but also makes the content more accessible. Moreover, in Automatic Speech Recognition (ASR), a field of growing interest, there is a huge need for training data. Manual speech segmentation has been shown to be more laborious than manual transcription, especially when dealing with tens of hours of speech. To automate the alignment process, tools such as forced aligners have been widely used for more than 10 years.

Forced alignment is a technique for matching a signal with its orthographic transcription using ASR with respect to the duration of linguistic units. In comparison to end-to-end ASR models, which require only audio, forced aligners need both transcription and audio. Most forced aligners, however, are language-dependent and trained on English data, whereas under-resourced languages lack the resources to develop an acoustic model required for an aligner as well as manually aligned data.

Instead of training new models, which requires a large amount of data, an alternate technique such as cross-language forced alignment, in which an aligner trained with a different language than the speech and transcriptions to be aligned, can be utilized. In this work, three approaches for forced aligners are employed and tested on the small dataset in Swedish and Norwegian: (1) two traditional off-the-shelf forced aligners with a Swedish acoustic model based on Hidden Markov Models, such as Munich AUtomatic Segmentation System (WebMAUS) (Kisler et al., 2017) and the Montreal Forced aligner (MFA) (McAuliffe et al., 2017), (2) a forced aligner Aeneas based on a more classic, signal-processing-based approach, called Dynamic Time Warping (DTW) using text-to-speech (TTS) synthesis (Pettarin, 2017), (3) two pre-trained and fine-tuned deep network Wav2Vec2 Swedish models (Malmsten et al., 2022) with forced alignment based on the CTC algorithm (Kürzinger et al., 2020).

In the first part of the thesis, due to the lack of datasets containing audio, text and time stamps for mentioned languages, small corpora in Norwegian and Swedish are created and manually annotated. The created corpora are then divided into three smaller datasets on the account of the degree of spontaneousness: read, semi-spontaneous and spontaneous speech. The manually labelled dataset is treated as a gold standard and used for comparison purposes.

In the second part of the thesis, the off-shelf-forced alignment systems are evaluated on the Swedish data and their performance is compared.

In the third part of the thesis, it is evaluated whether a cross-language forced alignment can be applied to Scandinavian languages by using the off-the-shelf forced alignment systems trained on Swedish to align Norwegian data.

1.1. Purpose & Research Questions

The purpose of this work is to evaluate the state-of-the-art models for aligning speech and transcription, as well as how effectively the problem of temporal alignment can

be addressed given audio and transcript. Collecting data for training a speech model is a laborious and expensive process, which is often out of reach for under-resourced languages. By employing a model trained solely on one Scandinavian language and testing it on related languages, we want to see if we can achieve valuable temporal alignment for other Scandinavian languages and hence cut the process of creating a separate model for each language.
The following research questions are addressed:

- What is the accuracy of the aligners relative to the alignments of a manually created gold standard?
- How is the performance of a forced alignment system affected when it is used to align a language, on which it was not trained?
- Are there any differences in performance between aligners for 1-2?
- How does the degree of spontaneousness in the speech affect the performance of a forced alignment system?

1.2. Outline

This thesis is structured as follows:
Chapter 2 provides background information about Scandinavian languages, Automatic Speech Recognition, Forced Alignment and the off-the-shelf aligners used in the thesis. Related work is also presented.
Chapter 3 provides information about the challenges of creating speech corpora and the data specification, collection and preprocessing.
Chapter 4 presents the gold standard and the methodology adopted for the evaluation of the off-the-shelf forced aligners. Evaluation metrics are introduced.
Chapter 5 provides a summary of the results on the Swedish speech dataset.
Chapter 6 provides a summary of the results on the Norwegian speech dataset.
Chapter 7 analyzes and discusses the results of the experiments.
Chapter 8 concludes the thesis and discusses future work.

2. Background and Related Work

This chapter provides basic background information about Scandinavian languages, Automatic Speech Recognition, Forced Alignment and the off-the-shelf aligners that were used in the present work, along with a description of the toolkits and algorithms of those aligners.

2.1. Scandinavian Languages & Mutual Intelligibility

Intelligibility does not necessitate that all speakers use the same language. Some genetically related languages are so similar that their users are able to communicate with each other without speaking the same language, nor a *lingua franca*, such as English (Gooskens, 2007; Haugen, 1966).

One group of the languages that are commonly viewed as being mutually intelligible are the Scandinavian languages. The Scandinavian languages belong to the North Germanic group, which is further divided into two branches: *East Scandinavian*, consisting of Danish and Swedish and *West Scandinavian*, containing Faroese, Icelandic and Norwegian. Another categorization is based on mutual intelligibility and divides the North Germanic languages into two groups: *Continental Scandinavian* (Swedish, Norwegian and Danish) and *Insular Scandinavian* (Icelandic and Faroese) (Sahlgren et al., 2021). Moreover, Norwegian has two official forms of the written language, that is *Nynorsk* (New-Norwegian) and *Bokmål* (Dano-Norwegian). According to Sahlgren et al. (2021) Bokmål is said to be an East Scandinavian language due to its similarity with Danish, while Nynorsk is classified as a West Scandinavian language.

Some researchers claim that by Scandinavian languages, we should only understand languages that belong to the Continental Scandinavian group (Delsing, 2005).

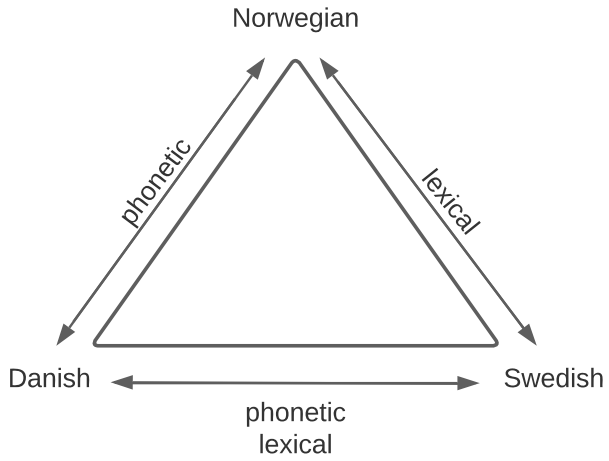


Figure 2.1.: Schematic overview of the linguistic differences that form the largest obstacle for the mutual intelligibility between the Scandinavian languages from Gooskens (2007).

Danish, Norwegian and Swedish display many similarities in vocabulary, pronunciation and grammar due to their shared linguistic heritage (all descended from *Old Norse*), as well as the shared history.

As can be seen in Fig. 2.1, Danish and Norwegian differ mostly in pronunciation, but share most of their vocabulary. On the contrary, Norwegian and Swedish share phonetic similarities but face lexical differences. The communication between Danish and Swedish speakers is hindered by both: phonetic and lexical differences (Haugen, 1966). Moreover, Bokmål has strong Danish influences, while the more genuinely Norwegian forms in Nynorsk are more like Swedish (Delsing, 2005).

When it comes to the writing system, Danish and Norwegian share the alphabet, which consists of 29 letters, including three vowels with diacritics, namely *æ*, *ø* and *å*. Nynorsk also has several letters with diacritic signs, such as *é* and *è*, which are not compulsory (Språkrådet, 2017). The Swedish alphabet also includes 29 Latin letters with three vowels with diacritics, that is, *å*, *ä* and *ö*. The letters *ä* and *ö* correspond to *æ* and *ø*, respectively (Haugen, 1966).

To exemplify the differences and similarities between the languages, some word examples from the written languages are presented in Table 2.1.

English	**Danish**	**Norwegian**	**Swedish**
not	ikke	ikke/ikje	inte
and	og	og	och
difference	forskel	forskjell/skilnad	skillnad
begin	begynne	begynne/byrja	börja

Table 2.1.: The examples of differences between Danish, Norwegian and Swedish in the written language. For the Norwegian, the Bokmål variant is presented to the left, while the Nynorsk variant is to the right of the column. The English translation is provided in the first column. The examples presented by Delsing (2005).

Sometimes Norwegian vocabulary stands between Danish and Swedish, which can be noticed in the two last rows of the table. Nynorsk and Swedish often resemble one another, for instance, *skiland – skillnad* (see Tab. 2.1).

A number of studies have been conducted to determine the actual level of understanding between speakers of the Scandinavian languages (Delsing, 2005; Gooskens, 2007; Haugen, 1966). The obtained results were consistent among the mentioned studies. Norwegians tend to comprehend other Scandinavian languages the best, while Norwegian is marked as the most easily understood language for both Swedes and Danes. When it comes to the writing language, Bokmål is easier to understand for Danes, while Nynorsk is easier for Swedes (Delsing, 2005). Danish has a large deviation between speech and writing, which results in incomprehension of the language by its Scandinavian neighbours. Hence, Danish is the hardest language to understand for both Swedes and Norwegians. Moreover, most of the time, the intelligibility is not symmetrical. For instance, Danes understand Swedish better than Swedes understand Danish (Gooskens, 2007).

The study conducted by Gooskens (2007) confirmed that the phonetic distances between languages have a bigger influence on the intelligibility than the lexical distances. Moreover, as Gooskens (2007) highlighted the differences in intelligibility are dependent on the region where the speakers live. Scandinavian communities speak a wide variety of local dialects, especially in rural areas. Sahlgren et al. (2021) argued that from a typological perspective the difference between dialects within the Scandinavian languages is in some cases probably greater than the difference between the standard languages. Since boundaries between dialect regions are gradual, a

dialect continuum is created, which is not always consistent with country borders (Gooskens, 2020). For instance, the dialects used in southern Sweden were treated as Danish dialects until 1658, when present southern Sweden was part of Denmark. After the region was ceded to Sweden, the local dialects started to be classified as Swedish dialects despite no linguistic changes (Chambers and Trudgill, 1998). Thus, a Swedish speaker from Malmö can comprehend Danish better than a Swedish speaker from Stockholm. In addition, the level of semi-communication can be improved by being exposed to another language (Haugen, 1966).

Even though there are debates about the meaning of *Scandinavia* and the language categorization within the group, due to the time limits, the thesis focuses on two languages from the Continental Scandinavian group, that is, Swedish and Norwegian.

2.2. Automatic Speech Recognition

Automatic Speech Recognition (ASR) has been a subject of study for more than 60 years (Morgan, 2010). Nowadays ASR technology has been used for tasks such as personal assistants, chatbots and generating captions for audio or video. The main goal of ASR is to find the optimal string of words, given a waveform (see Fig. 2.2).

Figure 2.2.: Simplified visualization of the Automatic Speech Recognition process. A waveform of the Swedish word *Ingress* ("Preamble") is transformed to a string.

The basic components of ASR systems are signal processing and feature extraction, acoustic modelling, language modelling and hypothesis search (Benesty et al., 2008), which will be explained in turn.

Since the raw speech signal has high dimensionality, speech recognition systems require feature extraction to decrease variability and filter out undesirable sounds, such as silence, noise or channel distortions (Kamath et al., 2019). Features are extracted based on the short-term amplitude spectrum from speech (i.e., phonemes) and amplify signals that imitate the human ear. The analysis of sound by the human ear is performed on a nonlinear frequency scale, which is also known as *Mel scale* (Shrawankar and Thakare, 2013). One of the most popular audio feature extraction methods is the *Mel-frequency cepstral coefficients* (MFCC) (Furui, 1986). After the acoustic signal is converted from time-domain to frequency-domain and features are extracted, the feature vectors are fed into acoustic models. The main goal of an acoustic model is to map the audio signal to the fundamental units of speech such as graphemes or phonemes. Implementations of these models include HMMs and neural networks (Benesty et al., 2008). A language model computes the probability of a sequence of words. The acoustic model is guided by the language model, which eliminates the predictions, which are not likely to occur due to, for example, proper grammar (Benesty et al., 2008). Finally, the hypothesis search is the combination of acoustic and language modelling. It returns the word sequence with the highest probability score as the output (Benesty et al., 2008).

2.3. Toolkits & Algorithms

In this section, the toolkits and algorithms of forced aligners are introduced and described.

2.3.1. HTK

Hidden Markov Model Toolkit (HTK) is a portable speech recognition toolkit for building and manipulating Hidden Markov Models (HMMs), as well as Gaussian Mixture Models modelling that was developed at the University of Cambridge by Young et al. (2002). HTK is one of the most popular toolkits for speech recognition and is primarily aimed at developing HMM-based speech processing tools, mainly speech recognizers, however, it also can be adapted to compute forced alignments. The toolkit has a restrictive license, that is, the toolkit is not free for commercial purposes, although it can be purchased from the University of Cambridge.

2.3.2. Kaldi

Kaldi is an open-source speech recognition toolkit developed by Povey et al. (2011). The toolkit is written in *C++* and is dependent on two external libraries, both of which are publicly available: one for the finite-state framework and the second for numerical algebra libraries.

The standard approach for feature extraction is Mel-Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) features, which can be adjusted with, for instance, the number of mel bins. Other common feature extraction methods, such as Linear Discriminant Analysis (LDA) or cepstral mean and variance normalization, are also supported. When it comes to acoustic modelling, Kaldi supports traditional models, such as diagonal Gaussian Mixture Models (GMMs) and Subspace Gaussian Mixture Models (SGMMs), but new types of models are possible to employ.

Users should be able to use any language model, which can be described as a Finite State Transducer (FST). Finally, training and decoding algorithms employ Weighted Finite State Transducers (WFSTs) (Povey et al., 2011).

The authors highlight that using Kaldi has benefits over the HTK toolkit (described in section 2.3.1), which is employed by the majority of current aligners. First, the distribution is much simpler, due to Kaldi's more permissive license, moreover, the code is modern, flexible, cleanly structured and has better Weighted Finite-State Transducers and math support (Povey et al., 2011).

2.3.3. Dynamic Time Warping

Traditionally, to determine the similarity of two sequences, sequences are converted into vectors to calculate *Euclidean distance*, that is the absolute distance between the values of the two sequences. However, in speech recognition *Euclidean matching* – the traditional time series matching – is too restrictive in the case when durations for two sequences are different (Furtună, 2008). For instance, words in the same sentence, recorded by various speakers, can have different durations, hence they cannot be correctly matched by Euclidean matching, since even small-time shifts can result in incorrect identification. To solve this time shift and align words correctly the Dynamic Time Warping algorithm is used (Amin and Mahmood, 2008). In speech recognition, the Dynamic Time Warping algorithm (DTW) is a dynamic programming algorithm that is used to determine the similarity between two time series (audio signals). The algorithm calculates the warping path values and the

minimum distance between the two audio signals (Furtună, 2008). The smaller the warping path is, the more similar two time series are (Permanasari et al., 2019).

In the DTW algorithm given time series are divided into smaller, equal parts, dividing the entire problem into sub-problems. For each sub-problem, the distance measure is calculated. The overall solution is developed after calculating the optimum paths (i.e., the shortest path) for all sub-problems (Amin and Mahmood, 2008).

The algorithm starts with measuring local distances between the elements of the two given sequences, which results in a matrix based on the length of the time series, where distances are stored. After calculating the distances between elements in the matrix, the algorithm looks for the minimal distances between elements of two time series in the matrix in result creating a warping path. To calculate the similarity between the two time series, the minimum distances are summed up (Amin and Mahmood, 2008).

Due to the high space complexity of the algorithm, the forced aligner Aeneas uses Sakoe-Chiba Band Dynamic Time Warping (Sakoe and Chiba, 1978). While the DTW algorithm computes distances over the full matrix, which requires a huge amount of space, Sakoe-Chiba Band DTW computes only a band around the main diagonal and the algorithm path is limited to staying within this band. The approach is an approximation of the exact DTW since the calculated path is an approximation of the minimum cost path (Pettarin, 2017). However, as Pettarin (2017) highlighted, the Sakoe-Chiba approach often returns the optimal solution.

2.3.4. CTC Segmentation

Connectionist Temporal Classification (CTC) is an algorithm often used for end-to-end ASR systems (Graves et al., 2006). For every frame of an audio recording, the CTC algorithm outputs a single character, in a manner that the input has the same length as the output. To encode duplicate characters, a special symbol called *blank* is used. Then a collapsing function is applied combining sequences of identical letters (Graves et al., 2006).

Kürzinger et al. (2020) proposed a segmentation algorithm based on a CTC network for extracting temporal alignments, also in case of additional unidentified speech segments at the beginning or end of the audio. To generate an alignment, first, a CTC-based end-to-end network (e.g., implemented in an ASR system) that was trained on already aligned data, is used to estimate the frame-wise label probability of each audio frame.

In the next step, dynamic programming is used to calculate all possible maximum joint probabilities $k_{(t,j)}$ for aligning the transcript until character index j to the audio up to frame t. The probabilities are mapped into a trellis – a 2-dimensional matrix with a time axis and label axis – that symbolizes the probability of transcript labels aligned at each time frame.

As Kürzinger et al. (2020) claims, to calculate the maximum joint probability at a point, the most probable of the two possible transitions is taken: a) only a ***blank*** symbol, or b) the next character is consumed. To align the transcription start to an arbitrary point of the recording, the transition cost is set to zero for staying at the first character.

In the last step, the character-wise alignment is computed by using backtracking. We start with the most probable time step of the last character in the transcript. Based on the highest post-transition probability, we traverse back in time (choosing stay or transition) and determine the alignment of the audio frame to its character from the transcription. To sort out utterances with deviations between speech and corresponding text, a confidence score for each segment is derived. For instance, if a

word within an utterance is missing, the value of the confidence score is low. Finally, the characters are merged. For the final probability of each segment, the frame-wise probability from the emission matrix is taken (Kürzinger et al., 2020).

CTC segmentation speed is influenced by a few factors such as window size, length of audio, length of text, how well the text fits audio and the preprocessing functions. The memory consumption of the algorithm strongly depends on the model architecture and the used toolkit (Kürzinger et al., 2020).

2.4. Forced Alignment

In this section, I briefly discuss forced alignment and then the state-of-the-art force aligners used in this study are described.

2.4.1. Description

The method of matching speech to a corresponding transcription, such that each written word from the text is linked to its respective sound, is known as *Forced Alignment (FA)*. Given a method of mapping graphemes to phonemes (e.g., pronunciation dictionary consisting of orthographic transcription of words and their phonetic pronunciation) and a statistical model of phone realization, the audio and text files are time-aligned at various granularity levels (e.g., chapter, sentence, word or phoneme). Forced aligners are often included in a speech recognition toolset, however, in comparison to ASR tools, which require only audio as input, forced aligners require both audio and the transcript (see Fig. 2.3). Some forced aligners require also a pronunciation dictionary of words used in the transcripts. The output of these systems is (most of the time) a time-aligned Praat TextGrid (Boersma, 2001) (see Fig. 2.4).

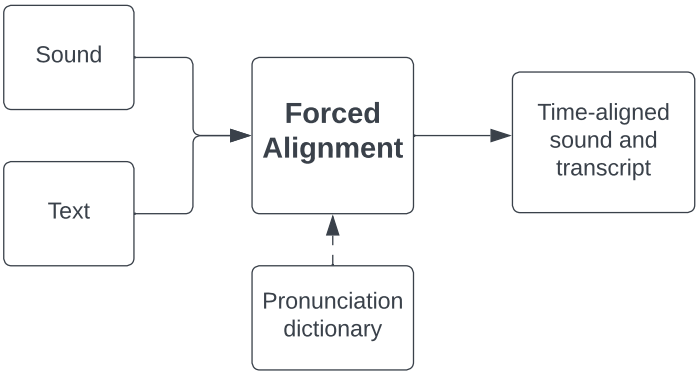


Figure 2.3.: Visualization of the simplified forced-alignment process.

Most forced aligners differ in two aspects: *trainability* and *architecture*. For *trainability*, some aligners can be retrained on new data, while others allow usage of only pre-trained models. While concerning *architecture*, there are differences in the acoustic model, which is employed to form the realization of phones (McAuliffe et al., 2017).

Overall, there are two approaches for forced alignment (Leinonen et al., 2021):

- **Cross-language forced alignment (CLFA)** - also known as cross-linguistic; an aligner trained on a different language is used for aligning data in another language.

- **Language-specific forced alignment (LSFA)** - training a model on the target language by using existing transcriptions.

For under-documented languages, training a language-specific forced alignment system is not always possible. Typically, training a well-performing forced aligner requires at least several hours of collected and transcribed data (Tang and Bennett, 2019).

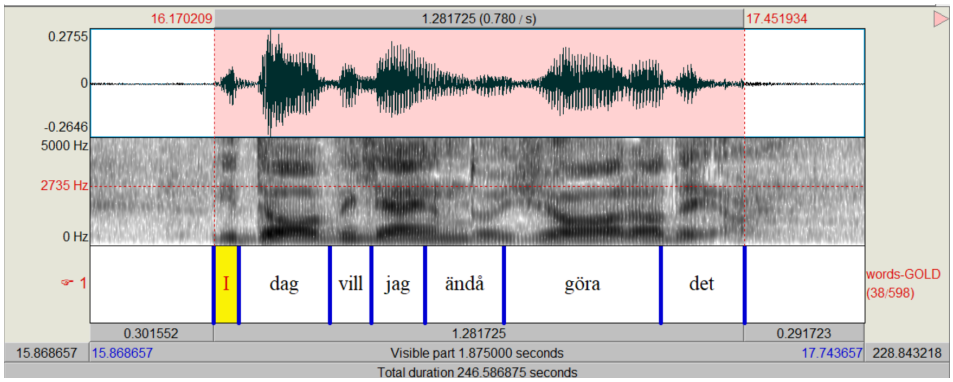


Figure 2.4.: Sample word alignment for a Swedish sentence *I dag vill jag ändå göra det* ("Today, I want to do it anyway") from Praat TextGrid (Boersma, 2001). From top to bottom: audio wave, spectrogram and the gold standard.

Over the last ten years, forced alignment has become widely employed in the scientific study of language, including sociolinguistics and phonetics, due to the accessibility of efficient, pre-built and simple-to-use aligners (McAuliffe et al., 2017). While forced aligners help to reduce time and expenses, the output is still affected by a variety of aspects such as audio quality, speech rate, data type and the algorithm and the training data behind the aligner.

2.4.2. Munich AUtomatic Segmentation System

The Munich AUtomatic Segmentation system (MAUS) is a forced aligner developed at the Department of Phonetics, University of Munich (Schiel, 1999). MAUS employs rules that are statistically weighted for the prediction of possible pronunciation variants, as well as HTK toolkit, more precisely its HMM search algorithm for finding the most likely segmentation and labelling.

According to Kisler et al. (2017), MAUS is distinguished from other forced aligners by its two-step modelling approach: pronunciation prediction and signal alignment. MAUS pipeline follows 3 steps (Schiel, 1999):

1. First, the given transcript is tokenized and stripped of punctuation. After that, it is fed to the grapheme-to-phoneme algorithm. The G2P algorithm is implemented in one of two ways: a) rule-based, or b) a mix of lexicon lookup and rule-based fallback. The output of the first step is a string of phonological symbols.

2. In the second phase, a probabilistic model of all possible variants of pronunciation is calculated. The string of phonetic symbols is fed to the production system with language-specific knowledge of pronunciation and statistically weighted re-write rules are applied to the string. After that, a Markov model, with all the possible pronunciation variants of the given utterance in addition to the conditional probabilities, is computed. Each path through the graph

depicts a unique potential pronunciation, whereas the product of all the probabilities on the arcs provides the overall predicted probability of that variant of pronunciation.

3. In the final stage, the produced Markov model, along with the speech signal are fed into a Viterbi alignment algorithm (Young et al., 2002), that identifies the most probable path through the model by computing the combined probability of the overall predicted probability and acoustical score. Finally, MAUS outputs the segmentation into words and single phonemes.

The authors recommend processing recordings that last less than 10 minutes since the processing power grows quadratically as the complexity of the Markov model increases (Kisler et al., 2017). In case of recordings longer than 10 minutes, the recordings should be chunked into smaller files.

Nowadays, MAUS supports around 23 languages or language variants (Kisler et al., 2017). Even though MAUS offers a Swedish model, it is far from optimal: there is no real Swedish acoustic model since no phonetically labelled and segmented Swedish data was used for training. Acoustic models were cloned from the manually segmented part of the Norwegian corpus *NB Tale* (Schiel, 2022).[1] Hence there is no pronunciation model and there is no phonetic recognizer provided. The Norwegian model was enriched with the Swedish NST dataset (Birkenes, 2020).

MAUS offers 3 ways to run the software: a) locally on the user's computer, b) accessed through a web service and c) WebMAUS basic, which can be accessed online as a web interface and through a web service. WebMAUS basic offers automatic segmentation based on orthography, which does not require the canonical pronunciation transcript as the input, in comparison to the MAUS. The web service combines G2P, which transfer an orthographic transcript into the most-likely standard pronunciation and the segmentation tool. In this research, WebMAUS is used due to its simplicity (Kisler et al., 2012).

2.4.3. The Montreal Forced Aligner

The Montreal Forced Aligner (MFA) is a forced alignment system developed by McAuliffe et al. (2017) as a successor to the Prosodylab-Aligner (Gorman et al., 2011). MFA is built upon an open-source speech recognition toolkit - Kaldi (Povey et al., 2011). In opposition to Prosodylab-Aligner, where monophone acoustic models are employed, MFA utilizes triphone acoustic models to capture contextual variability in phone realization (McAuliffe et al., 2017). A standard GMM/HMM architecture adapted from existing Kaldi recipes is used.

MFA offers both train/align and pre-trained models providing acoustic models[2] and G2P models[3] for more than 20 languages, as well as pronunciation dictionaries[4] for 7 languages.

Most of the acoustic and G2P models, which can be used for MFA, were developed by Schultz et al. (2013) as a part of *GlobalPhone*, a multilingual data corpus created in cooperation with the Karlsruhe Institute of Technology (KIT). Acoustic models use fully-continuous 3-state left-to-right HMMs, while Gaussian Mixtures with diagonal covariances model the emission probabilities. Pronunciation lexicons were constructed

[1] NB Tale is a speech database for Norwegian produced for the National Library of Norway. The database contains recordings of 380 speakers from 24 different dialect areas (Lingit, 2015).
[2] https://montreal-forced-aligner.readthedocs.io/en/latest/user_guide/models/acoustic.html
[3] https://montreal-forced-aligner.readthedocs.io/en/latest/user_guide/models/g2p.html
[4] https://montreal-forced-aligner.readthedocs.io/en/latest/user_guide/models/dictionary.html

by using phone-based pronunciation (IPA-based naming of phones) by using language-specific phone sets and manual pre-processing by native speakers, who corrected errors. To create a dataset, each language had roughly 100 adult native speakers reading articles from national papers reporting political news and economic news, e.g. *Göteborgs-Posten* for Swedish. The Swedish pronunciation dictionary consists of 48 phones, 25k words and 25k dict entries. The Swedish data was recorded in Stockholm and Värnamo.

Users may align data by using either pre-trained acoustic models or by training them from scratch on their data (train/align). However, according to the authors of MFA, when employing acoustic models developed from scratch, alignment may be greatly improved, especially when the dataset to be aligned is sufficiently vast and diverse. They also recommend experimenting with pre-trained models and re-training them.

To train a model, MFA implements the following pipeline:

1. First, monophone GMMs are iteratively trained and used to produce a basic alignment.

2. Next, triphone GMMs are trained to take into consideration surrounding phonetic context, as well as triphone clustering to overcome sparsity.

3. For acoustic features MFA employs MFCCs. Thirteen MFCCs are computed using a window size of 25 ms and a frame shift of 10 ms.

4. Feature transforms for each speaker are computed in the last round of training using feature space Maximum Likelihood Linear Regression (fMLLR). When aligning with pre-trained models, speaker adaptation is also performed, although it can be deactivated for faster alignment.

MFA supports both short recordings with an associated transcription file and long recordings which contain time-aligned periods of transcribed speech. The alignment is possible on two-level granularity: at the word and phone levels. In comparison to the Prosodylab-Aligner, MFA – rather than requiring that every word from the transcripts in the pronunciation lexicon – has a model for unknown words including a unique phone, allowing them to be modelled while keeping surrounding words aligned.

2.4.4. Aeneas

In contrast to other forced aligners, which are based on the automatic speech recognition methods, Aeneas (Pettarin, 2017) utilizes Text-To-Speech (TTS) + Dynamic Time Warping (DTW) – a more traditional, signal-processing-based method. First, a TTS engine is employed to convert the given text into audio. Second, the DTW algorithm is used for aligning the actual audio (MFCCs representation of the audio wave) with the one produced by a TTS engine, warping the time axis. Thanks to that, the software can predict the beginning and end of each speech passage and the synchronization map can be computed.

To explain it in more detail, Aeneas workflow consists of 5 steps:

1. The given audio file is converted to a mono wave file.

2. A text-to-speech system (TTS) is used for synthesizing the given text file, which generates the audio and time mapping. As the author highlights, the audio produced by the TTS does not need to sound natural, but it needs to be understandable. Moreover, in the TTS+DTW approach, this is the only step,

which is language-dependent, due to the audio output being produced from text using rules that vary from language to language.

3. MFCCs are computed from the given and synthesized audio. As a result, two new objects are created: a) a matrix containing the MFCC coefficients for the given audio file and b) a matrix containing the MFCC coefficients for the audio created by a TTS engine.

4. The DTW algorithm is applied and computed between the two matrices. First, the dot product of the two matrices is computed for the cost function. Subsequently, to find the minimum cost path of transforming the synthesized audio into the given one, DTW is used for the cost matrix.

5. Based on the DTW path, time mapping for the given audio file is computed.

Even though Aeneas provides word-level alignment, it was designed for phrase/sentence level granularity. Hence, the accuracy of the synchronization map for word-level alignment is worse in comparison with ASR–based forced aligners for languages with trained ASR models (Pettarin, 2017). Aeneas is reported to be robust against background noise, local rearrangements of words, as well as misspelt or mispronounced words (Pettarin, 2017).

A language is supported by Aeneas only if it is supported by at least one of the built-in TTS engine wrappers. The default engine *eSpeak* does not require additional licenses as other built-in TTS engines such as *AWS Polly TTS* or *Nuance TTS* API wrappers. A custom TTS engine Python wrapper, provided by a user, may be also employed.

According to Pettarin (2017), Aeneas has mainly two limitations: a) a wrong synchronization map may be computed due to the errors within the audio or transcription and b) the amount of RAM locally needs to be sufficient for the maximum duration of a single audio recording, that is, to align 10 hours audio, 16 GB RAM is required. The intent behind Aeneas was an application for many languages that is easy to install and run, without heavy language models. This is a distinction from other forced aligners, which most of the time are developed in academia specifically for sociolinguistics studies, or are commercial products being heavily engineered and most of the time are side products of ASR models (Pettarin, 2017).

2.4.5. Wav2Vec2+CTC segmentation

The Wav2Vec2 model is a Transformers framework, proposed by Baevski et al. (2020), for self-supervised learning of speech representations, which masks latent representations of the raw waveform and solves a contrastive task over quantized speech representations.

For the present work, Wav2Vec2 is used solely as a CTC-based end-to-end network for estimating the frame-wise label probability of each audio frame and generating labels. Hence, it is not a forced aligner on its own, but an important component of the CTC-segmentation algorithm. Two pre-trained Wav2Vec2 models, which were fine-tuned by the researchers from the National Library of Sweden (KBLab), were employed.

1. *Wav2Vec2 Large XLSR 53 Swedish* - a fine-tuned version of *facebook/wav2vec2-large-XLSR-53* in Swedish using the NST Swedish Dictation. The XLSR model learns cross-lingual speech representations by pretraining a single model from the raw waveform of speech in multiple languages (Conneau et al., 2020). XLSR used Common Voice, BABEL and MultiLingual LibriSpeech as training data.

The model was further pre-trained with a corpus of 1000 hours of spoken Swedish from various radio stations. Secondly, NST Swedish Dictation and Common Voice datasets were used for fine-tuning. Two evaluation scores on the Common Voice test set were reported: Word Error Rate (WER) of 14.30% and Character Error Rate (CER) of 4.93% (Malmsten, 2021).

2. *Wav2vec 2.0 large VoxRex Swedish* - a fine-tuned version of KBs VoxRex large model using Swedish radio broadcasts, NST and Common Voice data. The model was pretrained on the P4-10k corpus, containing 10,000 hours of Swedish local public service radio along with 1500 hours of audiobooks and other speech from KBs collections. Without a language model, WER for NST + Common Voice test set (2% of total sentences) achieves 2.5%. While for Common Voice test set the reported WER is 8.49% and 7.37% with a 4-gram language model (Malmsten et al., 2022).

2.5. Related work

The existing literature on the forced alignment task focuses particularly on three issues, such as forced alignment for low-resource languages, alignment of long recordings and alignment of imperfect transcriptions. Due to the topic of the thesis, chosen articles concerning the first issue will be introduced.

A surprisingly high number of languages and dialects is affected by a lack of resources, such as training data or models, for speech recognition tasks. Due to that, there is a need for multilingual models that would take advantage of similarities between dialects and languages from the same language group (Schultz et al., 2013). The majority of forced alignment tools are language-dependent and under-resourced languages hardly ever have adequate resources to build an acoustic model for an aligner (Leinonen et al., 2021). Moreover, according to Schultz et al. (2013), to develop a stand-alone good-quality speech processing system a developer needs to have not only the technical background but also the native expertise of a target language.

As MacKenzie and Turton (2020) noted, there is a need among sociolinguists for off-the-shelf forced alignment systems performing well on a variety of different language dialects without training or significant manipulation. In their study, MacKenzie and Turton (2020) assessed the accuracy of two off-the-shelf forced aligners - *FAVE* (Rosenfelder et al., 2011) and *DARLA* (Reddy and Stanford, 2015) - on varieties of British English, focusing on a phoneme level of granularity. The placement of phone boundaries by each aligner was compared with that of a human annotator. Both aligners were trained on data in Mainstream American English and they were intended to use in that language. *FAVE* was based on the Penn Phonetics Lab Forced Aligner and employs the HTK speech recognition toolkit. On the other hand, *DARLA* was built on the Montreal Forced Aligner and - hence - employs Kaldi. The researchers found out that both aligners perform extremely well on a range of British dialects, including both spontaneous and read speech. However, the aligners performed worse on the Westray variety of Scots, which was radically different from Mainstream American English. Moreover, their performance dropped in the case of faster and reduced speech. Since their study was centred around sociolinguistics, MacKenzie and Turton (2020) pointed out that these forced aligners should be treated as a tool, not a replacement for a researcher. Furthermore, it was highlighted that in case of working with dialects far from the training language, or with data consisting of rapid speech rates and massively reduced speech, the alignment should be double-checked (MacKenzie and Turton, 2020).

Leinonen et al. (2021) presented a new Finnish grapheme-based forced aligner and validated its performance by aligning Uralic languages (Estonian, Northern Sami) and one unrelated language - English. As they noted, *cross-language forced alignment* (CLFA), in which an aligner trained on a different language is used for aligning data in another language, may be an alternative solution to the training of new models, especially in the case of low-resource languages, where a target language does not have enough transcribed data. Leinonen et al. (2021) introduced a word-level forced aligner based on Kaldi. Researchers used their own G2P mapping for target languages. As they reported, a rough mapping between Estonian graphemes and Finnish phonemes was developed due to that the languages are genetically related. For Northern Sami, the grapheme-to-phoneme mapping introduced by Leinonen et al. (2015) was used. Their research has established that even a simple grapheme-to-phoneme mapping developed by non-experts can generate valuable word alignments (Leinonen et al., 2021).

Hoffmann and Pfister (2013) indicate that the alignment between transcripts and speech may also help in building data for further speech processing. Traditionally, this is accomplished by using ASR on the speech data, however, an ASR system for the target language must be available. A different approach to an HMM-based forced alignment was presented. According to their method, HMMs do not need to be trained in the target language. A set of HMMs might be utilized as a universal model to align a large corpus in an arbitrary language with sentence granularity. The paper focuses on the sentence segmentation of large speech recordings using Viterbi forced alignment (Hoffmann and Pfister, 2013).

Tang and Bennett (2019) highlighted two issues with the cross-language forced alignment: a) Which of the existing language models should be utilized for alignment? and b) How should the phones in the target language be mapped to the phones in the language that is used for training the alignment model? According to the authors, CLFA's overall performance varies significantly among studies. In comparison to that, the language-specific alignment models result in better accuracy of alignment. However, a good LSFA model requires enough data for the training process, which is not possible in the case of low-resource languages. In their article, Tang and Bennett (2019) decided to take a different approach and combine CLFA and LSFA by training a forced aligner by pooling smaller quantities of data from genetically-related languages (two Mayan languages). They found that the method creates an efficient forced alignment system despite the small amounts of data used for the training.

The academic literature on forced alignment has revealed the urge for cross-language forced aligners without a need for additional training or other resources.

3. Data

To answer the research questions, I needed two small datasets for Scandinavian languages such as Norwegian (both Bokmål and Nynorsk) and Swedish, which would be used for testing, as well as creating a gold standard. Since no model was trained, there was no need for huge corpora and data could be web-scraped from the internet.

In this chapter, the dataset used for the thesis is discussed. I describe the challenges of creating speech corpora, how the data was created, in order to address the mentioned challenges, the categorization of corpora and its specifics.

3.1. Challenges

The process of creating speech datasets is time-consuming and costly, especially when it comes to under-resourced languages, due to the lack of resources. Forced alignment requires not only recorded audio but also accurate orthographic transcriptions of the audio. Publicly available speech datasets that meet these demands are hard to find.

Moreover, even if researchers decide to web-scrape data from the internet and audio happens to be transcribed, most of the time the transcriptions resemble subtitles in that i) the transcriptions lack, so-called *filled pauses*, that is, filler words (marking pauses or hesitations in speech, such as *um*, *uh* in English) and slips of the tongue (saying a wrong word by mistake), ii) repetitions are removed, iii) words are substituted for their synonyms and iv) the syntax is changed. Hence, most of the time, transcriptions need to be rewritten, which is problematic for researchers who do not know the transcribed language.

Many speech corpora that include other languages than English, such as *Common Voice* (Ardila et al., 2019), consist of unrelated short sentences spoken in isolation (Braunschweiler et al., 2010), which restricts testing since there is an interest in using forced aligners also for audio recorded in different environments, for instance, with background noise.

Since there was no dataset available that would be useful for answering the research questions, a new dataset had to be created, while the mentioned challenges had to be met.

3.2. Data Specification

The data can be divided into different categories with respect to the spontaneousness of the recorded speech. In this case, the data has three categories: *read*, *spontaneous* and *semi-spontaneous*.

Read Audiobooks are recorded in a controlled environment with a good quality microphone and without background noise. The transcript is read without any filler words or changes to the source material. However, even though audiobooks seem like perfect material for forced aligners, they may be a challenge due to their length. In MAUS the processing time increases quadratically with the number of words.[1]

[1] https://clarin.phonetik.uni-muenchen.de/BASWebServicesTest/help#WhatShouldIDoIfMyInputFileContainsTooManyWordsAndIDonTHaveAChunkSegmentation

	Semi-spontaneous			Spontaneous			Read		
Language	Time (min)	Utterances	Words	Time (min)	Utterances	Words	Time (min)	Utterances	Words
Norwegian	13:05	100	1734	8:27	147	1190	4:70	32	503
Swedish	13:35	100	1758	8:13	216	1253	3:54	22	456

Table 3.1.: Specification of the data by time, amount of utterances and words. Data has been categorized by languages and labels.

In addition, the Wav2Vec2 models do not accept audio recordings longer than 40 seconds.

Spontaneous The spontaneous speech in the conversational dataset consists of many filler words and slips of the tongue between words or sentences, as compared to the read speech. In the present work, the spontaneous speech data is a collection of street interviews and the quality of the audio is worse in comparison to the datasets of read and semi-spontaneous speech. Since the spontaneous speech data was not recorded in a controlled environment, but on streets – hence having poorer quality and lot of background noise – it may affect the aligner.

Semi-spontaneous Parliamentary speeches can be classified as semi-spontaneous as thus lies somewhere between read and spontaneous speech, since even though most of the speeches were scripted before any public appearance, trained speakers such as politicians do not need to follow scripts to the letter. However, there are moments, when speakers lose track of the written word (i.e., there is a larger amount of word fillers, repetitions and mistakes). The recordings of the parliamentary speeches also have a good quality, without any noticeable background noise.

3.3. Corpora

Semi-spontaneous Semi-spontaneous data consist of parliamentary speeches. For parliamentary speeches in Swedish, the official website for parliament was used, *Riksdagen*[2]. The website provides videos with transcriptions. First, the videos were downloaded and then audios were extracted. For parliamentary speeches in Norwegian, the Norwegian Parliamentary Speech Corpus (NSPC),[3] developed by the Norwegian Language Bank at the National Library of Norway, was used. The NSPC corpus consists of audio recordings of meetings in *Stortinget* (the Norwegian parliament) and corresponding orthographic transcriptions. The recordings total 140 hours of continuous speech (including breaks) and amount to 65.000 sentences and 1.2 million words. For testing purposes, there is no need for a dataset of this size. Instead, I extracted one recording alongside its transcription and then extracted smaller portions of the recording for three speakers.

Spontaneous Conversational speech datasets are much harder to acquire in comparison to parliamentary speeches due to the lack of transcriptions. Therefore I decided to use videos from two Youtube channels for language learners, *Easy Languages*[4] and *Simple Norwegian*[5], where street interviews with native speakers in respective languages are conducted and the subtitles are provided. The interviews are typically

[2] https://www.riksdagen.se/sv/webb-tv/

[3] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-58/

[4] Easy Languages: https://www.youtube.com/channel/UCqcBu0YyEJH4vfKR--97cng. Accessed: 2022-02-25.

[5] Simple Norwegian: https://www.youtube.com/c/SimpleNorwegian. Accessed: 2022-02-25.

File	Utterance
Swedish parliamentary speech; original transcription.	*Ärade ledamöter! Den 10 november begärde statsminister [..]*
Swedish parliamentary speech; manual transcription.	***Den ehm eh*** *ärade ledamöter! Den* ***tionde*** *november begärde statsminister [...]**
Norwegian parliamentary speech; original transcription.	*Al Qaidas angrep på USA 11. september 2001 drepte ca. 3 000 mennesker.*
Norwegian parliamentary speech; manual transcription.	*Al Qaidas angrep på USA* ***ellevte september to tusen og elleve**** *drepte circa* ***tre tusen*** *mennesker.*

Table 3.2.: Comparison of utterances in the original and manually altered transcriptions. *Note that the manual transcriptions may include language errors. **The original transcription provided the correct date, however, the speaker made a mistake saying '2011'.

about topics that relate to things specific to the Scandinavian countries, such as food and traditions, but also study programmes or seasons. The audio was extracted from the videos.

Read The read speech data had to be open-source in order to be used for testing. However, open-source datasets, such as *LibriVox* (Kearns, 2014) consist of books that – for languages like Swedish and Norwegian – are mostly from the 18th and 19th centuries. In the case of Scandinavian languages, which underwent many changes in the last century (e.g. *du-reformen* in Swedish), the language of many gathered books may be too archaic. In the end, I decided on using a part of *Universal Declaration of Human Rights*, issued by the United Nations and read in Swedish and Norwegian.

3.4. Data Preprocessing

The acquired recordings had to undergo preprocessing, which consisted of chunking the audio and normalizing it. For chunking, I used the open-source software Audacity,[6] then the audio was converted to 16kHz by using the command line tool SoX.[7] In the end, the audio was segmented into chunks smaller than 40 seconds and downsampled to 16KHz mono. Altogether, the entire speech corpus amounts to 51.04 minutes. Each recording had its respective transcription, which also had to undergo preprocessing (as was mentioned in Section 3.1), since the transcripts did not completely correspond to their recordings. For example, repetitions, slips of the tongue and filler words were not included in the texts. Each transcription had to be manually annotated to meet the criteria mentioned in Section 3.1 in such a way that the text precisely corresponded with the recording. The text also had to be normalized for the aligners: numbers had to be spelt out orthographically, as well as abbreviations (only if the speaker read the expanded form instead of the abbreviation).

Semi-spontaneous Since web-scraped parliamentary recordings are most of the time part of bigger debates, I had to cut out the speech, that was to be employed. The created parliamentary dataset amounts to three recordings per language, one speaker per recording and on average one speech lasts about 4 minutes and 26 seconds. In total, the recordings of parliamentary speeches last 26 minutes and 40 seconds. As was described above, the transcriptions provided by parliamentary websites (and the National Library of Norway) had to be annotated and normalized.

[6] https://www.audacityteam.org/
[7] http://sox.sourceforge.net/Docs/Documentation

Spontaneous Since code-switching is not part of the research, I cut out utterances in other than the main language. I also decided to remove an utterance if two voices overlapped, or if the utterance was unintelligible without looking at the transcription. Music from the intro and outro of the videos was also removed. The conversational data consists of 2 recordings per language. On average one recording has around 4 minutes and 10 seconds. In total, lasting 16 minutes and 40 seconds. The transcriptions had to be transcribed from the videos and annotated in accordance with the actual utterances, that is adding mistakes and word fillers, as well as spelling out numbers. No syntax changes were encountered, probably because the videos (and the captions) are intended for language learners.

Read On account of the fact that audiobooks are read-aloud-speech, there was not much preprocessing needed. The entire corpus is 8 minutes and 24 seconds long.

4. Methodology

This Chapter is structured as follows: in Section 4.1, a short overview of how the gold standard was constructed is provided, followed by a description of the methods and settings that were employed for the forced aligners in Section 4.2 and in Section 4.3, the evaluation metrics are described.

4.1. Gold Standard

In order to assess the performance of the forced aligners, the dataset described in Chapter 3 was utilized. To compare the performance of the forced aligners, the ground truth, that is, a manually verified segmentation and labelling of speech and text, had to be established. The gold standard was created on two granularity levels: utterance alignment and word alignment. Manual alignment of audio and text is a time-consuming task, that is, prone to fatigue errors (Pettarin, 2017). Due to the time limits, instead of aligning transcripts on the word level manually from scratch, transcripts were pre-aligned with one of the forced aligners and then manually corrected. On the contrary, Swedish and Norwegian utterance-level alignments were manually segmented from scratch and – in the case of Swedish data – compared with the word-level alignments. Utterance-level alignments consist not only of full sentences but also one-word utterances, for instance, *lagom!* ("just right") in the Swedish spontaneous speech data.

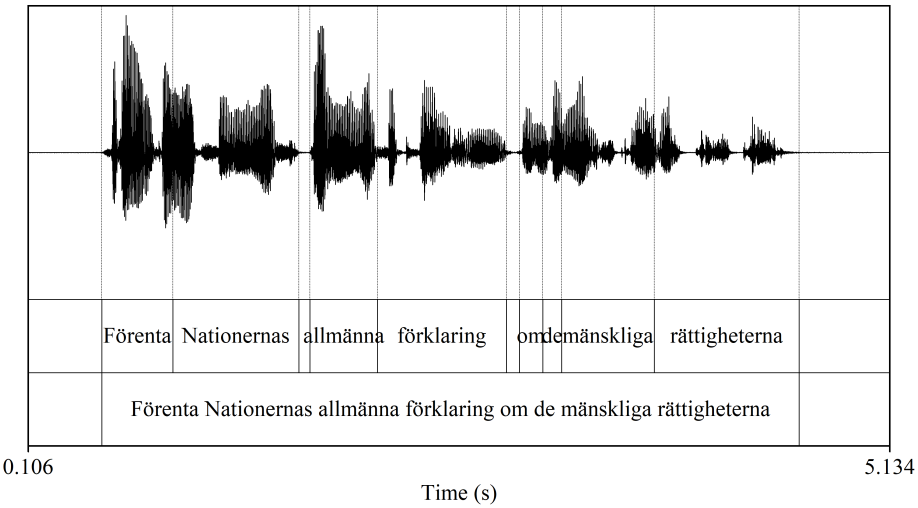


Figure 4.1.: Sample word and utterance segmentation for a sentence from the Swedish data, visualized in Praat TextGrid (Boersma, 2001). The alignment is a part of the gold standard. From top to bottom: sound wave, word alignment, utterance alignment.

WebMunich AUtomatic Segmentation (WebMAUS) (Kisler et al., 2012) with default setting was employed to pre-align transcriptions with the recordings (described in Section 3) on the word-level. WebMAUS was chosen due to the simple design of the web service and no requirement for local installation. The output files were

System	Toolkit/Algorithm	Granularity	Acoustic model	Training data
MFA	Kaldi	word, phoneme	HMM	Global Phone: read texts from newspaper articles (Swedish)
WebMAUS	HTK	word, phoneme	HMM	NB Tale: manuscript-read speech (Norwegian), NST
Wav2Vec2 + CTC s.	CTC segmentation	sentence, word*	Transformer (Wav2Vec2 (VoxRex))	P4: public radio, podcasts and audiobooks, Common Voice (read), NST (Swedish)
			Transformer (Wav2Vec2 (XLSR))	Common Voice (read), BABEL (conversational speech), MultiLingual LibriSpeech (read-speech), NST (Swedish)
Aeneas	DTW	chapter, sentence, word	TTS-dependent, here: eSpeak	N/A

Table 4.1.: Specification of the aligners used in the study. *Wav2Vec2 with CTC segmentation does not provide the specified granularity. The granularity can be adjusted by handling each segment (word, sentence etc.) as a single utterance.

loaded into Praat (Boersma, 2001) and manually realigned. Figure 4.1 shows a sample TextGrid with the first tier corresponding to word-level alignment and the second tier to utterance-level alignment. The silence – defined as audio that does not contain voice – between words or utterances was annotated, while laughter was not.

In the end, a word-level and utterance-level ground truth was created for Swedish and – due to the time limits – an utterance-level ground truth for Norwegian.

4.2. Testing the-off-shelf forced aligners

In this section, the description of off-the-shelf forced aligners, such as *WebMAUS*, *MFA*, *Aeneas* and *Wav2Vec2 with CTC segmentation* and how they were employed is provided. The specification for each aligner is presented in Table 4.1.

4.2.1. WebMAUS

Even though MAUS offers a Swedish model, it is far from optimal: there are no real Swedish acoustic models since no phonetically labelled Swedish data was used. The Swedish model performs forced alignment to cloned acoustic models from Norwegian (Schiel, 2022). The datasets were tested with the Swedish model. To compare the performance of the cloned and the original Norwegian acoustic model, the alignments for Norwegian were also generated with the Norwegian model.

WebMAUS created alignment on word and phoneme-level. Hence, as in the case of MFA, the beginning time of the first word and the end time of the last word of the given utterance were determined.

4.2.2. The Montreal Forced Aligner

The Montreal Forced Aligner (MFA) (Schultz et al., 2013) provides a user with an acoustic and a grapheme-to-phone (G2P) model for Swedish. After the installation of the Swedish acoustic model, the aligner required a pronunciation dictionary. Since there is no Swedish dictionary published for MFA, the G2P model was needed to acquire pronunciations from the orthographic transcripts of the recordings. The pretrained G2P model for Swedish provided by Schultz et al. (2013), which scored a Word Error Rate of 18.75%, was used. The G2P model was used to generate the pronunciation dictionary, as well as a document with Out Of Vocabulary (OOV) words.

To test if MFA with the Swedish acoustic model could produce alignment for another Scandinavian language, that is, Norwegian, the Swedish G2P model was applied to the Norwegian data to produce a dictionary. After that, the Swedish acoustic model was used to align the Norwegian data.

MFA created alignments on word and phoneme-level. To get utterance-level alignments from word-level alignments, the beginning time of the first word and the end time of the last word of the given utterance were extracted.

4.2.3. Aeneas

Aeneas works on various levels of granularity (paragraph, sentence, sub-sentence, word, etc.) and – in comparison to the other forced aligners – the intended granularity level should be detectable for the software in the text file. Hence, the transcription files had to be pre-processed before using them as the input for Aeneas. Since two levels of granularity (word and utterance) are tested, two scripts were written: one for the word level, which removed punctuation and created files with one word per line and another for the utterance level, which created files with one utterance per line.

As was mentioned in Section 2.4.4, even though Aeneas was not designed for word-level alignment, the software offers a few parameters to improve the accuracy of word alignment (following Pettarin (2017)),. MFCC non-speech masking was used by applying presets for word-level alignment. Although Pettarin (2017) argue for using *Festival* or *AWS/Nuance TTS API*, I decided to use the default TTS engine, *eSpeak*. The reasons were that other TTS engines are neither open-source nor have a Swedish model. For the utterance level granularity, the alignments were generated with basic settings. Other settings were tested, however, no difference between the alignments was noted. The Swedish data, as well as the Norwegian, was tested with a Swedish TTS engine.

Aeneas managed to create output in *TextGrid* format for all files.

4.2.4. Wav2Vec2+CTC segmentation

To test forced alignment with deep-network Wav2Vec2 models, a CTC-based network for segmentation proposed by Kürzinger et al. (2020), was employed. Two pre-trained Wav2Vec2 models, which were fine-tuned by the researchers from the National Library of Sweden (KBLab), were used: *Wav2Vec2 Large XLSR 53 Swedish* and *Wav2vec 2.0 large VoxRex Swedish* (Malmsten et al., 2022) (see Subsection 2.4.5).

First, a Wav2Vec2 model was used to generate labels and the label class probability of each audio frame. To test Swedish Wav2Vec2 models on transcripts in Norwegian, two additional letters (i.e., *æ* and *ø*), which do not exist in the Swedish alphabet, had to be added to the labels. However, due to the model not accepting new labels, *æ* and *ø* were transformed into *ae* and *oe*, respectively.

The transcripts had to be pre-processed before being fed to the algorithm by splitting into a list of utterances and then into a list of words. Punctuation was removed from the transcription. Next, the alignment probability (trellis) representing the probability of transcript labels occurring at each time frame, was created. Finally, the most likely path was found by backtracking. The algorithm output a confidence score for each utterance, along with a start and end time (see Fig. 4.2). All segments were merged into words, or further into utterances.

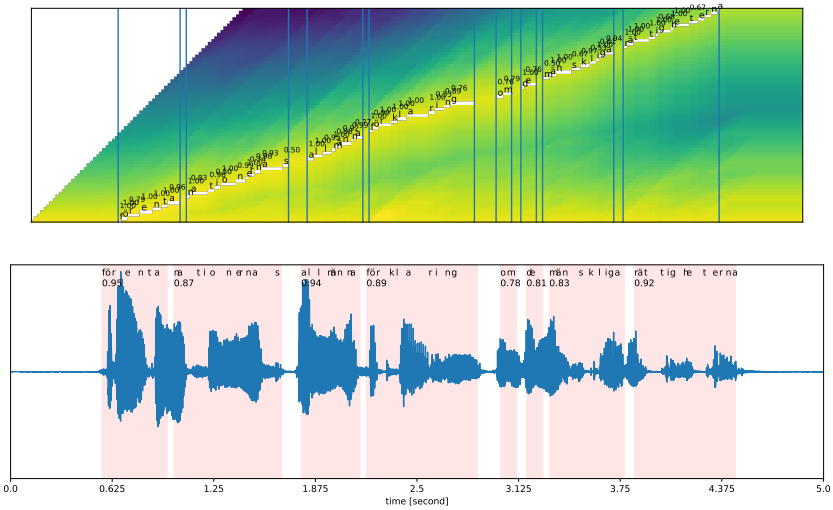


Figure 4.2.: Visualization of the path and the alignment with a Swedish Wav2Vec2 model (VoxRex)+CTC segmentation. Above: the path found by backtracing with confidence scores and labels. Below: the alignment of labels with their confidence scores.

4.3. Evaluation metrics

The aligners were evaluated by comparing the timings with the gold standard. Traditionally, the evaluation metrics, which are employed for the forced alignment task, are either the number of alignments below a specified threshold or the mean and median of the absolute timing error (Kempton, 2017).

In this research, the metrics Boundary Displacement, Accuracy and Duration Difference were employed.

Boundary Displacement Boundary Displacement is a measure of the difference between the segment boundaries allocated in the gold standard and those of a reference aligner. Boundary displacement was proposed by, among others, McAuliffe et al. (2017). The formula is provided below (4.1), where the absolute difference of S_bound_{human} denoting the timepoint in seconds of a segment's boundary placed in the gold standard and $S\_bound_{aligner}$ denoting the timepoint located in a reference aligner, is calculated. Next, all BDs per corpora are summed up. The lower the BD is, the better the alignment.

$$BD = |S\_bound_{human} - S\_bound_{aligner}| \tag{4.1}$$

To compare the employed aligners, the mean, median and standard deviation were taken into consideration. For the qualitative comparison, I extracted five words with the highest BD per segment for each dataset and aligner.

Accuracy To calculate the accuracy, the proportion of alignments below a particular threshold is measured. Two common choices are 0.01 and 0.02 seconds (Kempton, 2017; MacKenzie and Turton, 2020). In this thesis, to calculate the ratio of predictions under a specified threshold, we follow thresholds, such as 0.01, 0.025, 0.5 and 0.1 seconds, specified by MacKenzie and Turton (2020) as well as 0.5 seconds, specified by Kürzinger et al. (2020).

Duration The duration is used for calculating the time differences between the segments of the forced aligners and the segments in the gold standard. For the qualitative comparison, I extracted five words with the largest duration error for each dataset and aligner. In addition, the errors were calculated in accordance with *Weber's Law*, a hypothesis that a just-noticeable difference in a stimulus is proportional to the magnitude of the original stimulus (e.g., Stern and Johnson (2010)). In other words, if a word's duration in the gold standard is 0.1 seconds and an aligner predicts the duration of 0.2 seconds, it can be classified as a more noticeable error, than a word with the duration of 0.4 seconds, predicted by an aligner as 0.5 seconds. To calculate errors according to a just-noticeable difference (JND), the below formula is used,

$$JND = |1 - \frac{aligner\_alignment}{human\_alignment}| \tag{4.2}$$

where $aligner\_alignment$ is the estimated duration from an aligner and $human\_alignment$ is the duration from the gold standard.

5. Results for the Swedish alignment

In this Chapter, the results of the Swedish temporal alignments are presented and discussed. The chapter is divided into four sections. In Section 5.1, the general outline of the results is described. In Sections 5.2 and 5.3 evaluations and comparisons at the word and utterance level are presented. The qualitative comparison is presented in Section 5.4. The comparison is based on the evaluation metrics introduced in Section 4.3.

5.1. General outline of the results

The Montreal Forced Aligner The Montreal Forced Aligner (MFA) had no problems with aligning the semi-spontaneous or read dataset. However, it failed to align three out of 17 chunks of the spontaneous data and did not produce any output files for them. For each dataset, MFA produced a file with out of vocabulary words (OOV). For the semi-spontaneous dataset, the file consisted of 120 (OOV) words (out of 1758), while for the spontaneous dataset and the read dataset – 111 (out of 1253) and 40 (out of 457), respectively. In total, 271 OOV words were produced, not only compounds, such as *litteraturvetenskap* ("literary studies") and *yttrandefrihet* ("freedom of speech") but also common words such as *hej* ("hi"), *något* ("something"), *tolv* ("twelve"), *liv* ("life") and *kaffe* ("coffee").

Aeneas Aeneas managed to align each of the chunks. Noticeably, it did not annotate any pauses between words with the chosen settings. Moreover, it removed words from the output that could not be recognized. Hence, due to the mismatches between

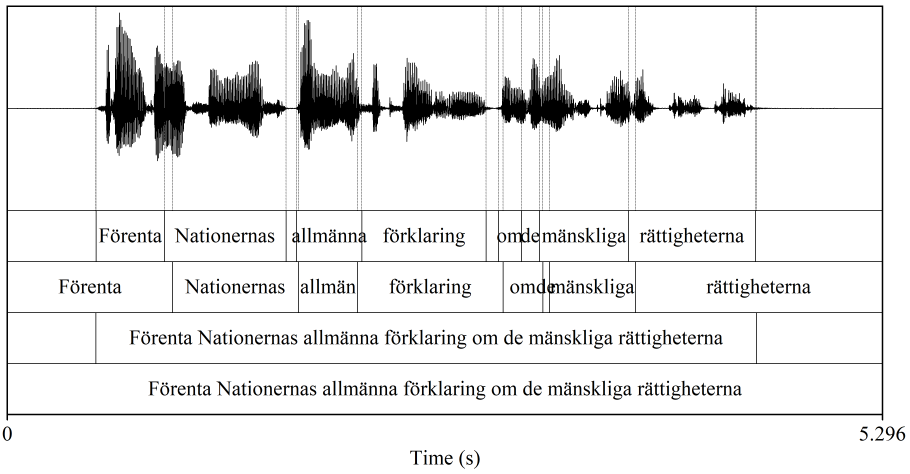


Figure 5.1.: Sample Praat TextGrid comparing the gold standard with Aeneas. From top to bottom: the gold standard (word-level), Aeneas (word-level), the gold-standard (utterance-level), Aeneas (utterance-level).

Aeneas and the gold standard and for comparison purposes, the removed words had to be re-added to the synchronization maps with no recognized time boundaries.

A visual inspection of Aeneas' alignments in Praat showed that pauses were not annotated. Figure 5.1 shows a sample TextGrid comparing the gold standard with Aeneas' output. As can be seen, Aeneas attached pauses to the preceding word, unless there was a silence at the beginning of the recording, in that case, the pause was combined with the following word.

WebMAUS In comparison to MFA and Aeneas, WebMAUS succeeded in creating the output files for the entire dataset without missing words. However, since the gold standard was manually aligned based on WebMAUS, the results are biased towards it. Hence, even though the results for MAUS are presented, they are not going to be described in greater detail.

Wav2Vec2+CTC Segmentation Both Wav2Vec2 models with CTC segmentation managed to create alignments for the entire corpus.

5.2. Evaluations and comparisons (word-level)

Figures 5.2, 5.4 and 5.6 show the distribution of manual/force aligned word boundary displacement, for each of the forced aligners. The x-axis represents the word boundary displacement (in seconds) and the y-axis represents the number of words. Tables 5.2, 5.5 and 5.8 provide mean, median and the standard deviation of manual/aligned word boundary displacement values for each aligner on the semi-spontaneous dataset, as well as the accuracy measurements for utterance boundaries displacement values at different tolerances. Figures 5.3, 5.5, 5.7 show linear regression plots between the duration of utterances in the human alignment and the duration approximated by the aligners and Tables 5.3, 5.6, 5.9 compare regression coefficients. The null hypothesis states that there is no correlation between the variables. When all p-values are smaller than .01, the null hypothesis is rejected and the independent variable is statistically significant.

5.2.1. Semi-spontaneous speech data

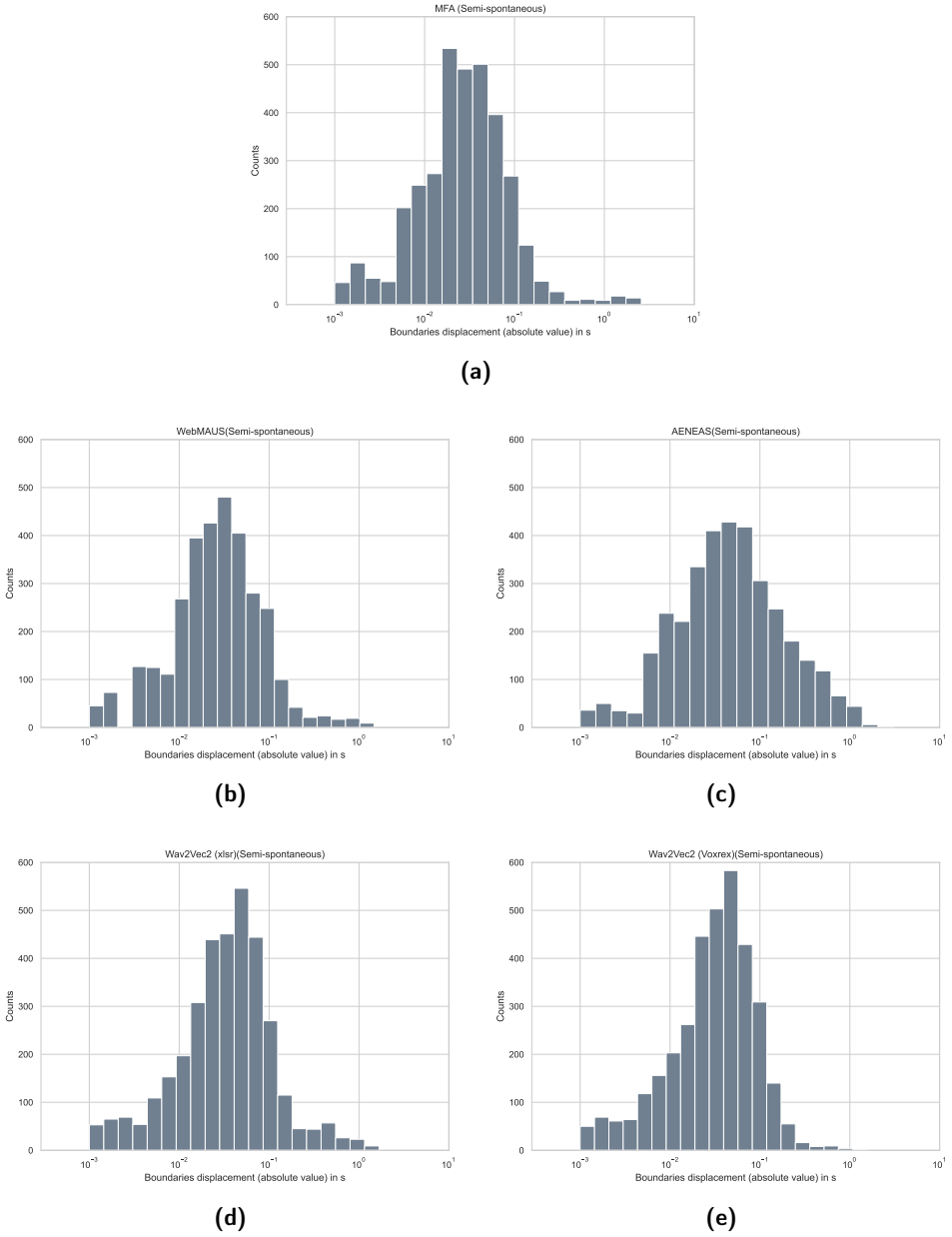


Figure 5.2.: Histograms of absolute boundary displacement (on log scale) between force-aligned word boundary and gold-standard annotations for the semi-spontaneous dataset. Due to the log scale, if a boundary displacement error equalled 0, the value was not visualized in the histogram.

Aligner	BD			Tolerance				
	Mean (s)	Median (s)	SD (s)	<0.01s	<0.025s	<0.05s	<0.1s	<0.5
MFA	0.060	0.027	0.177	22.53%	48.07%	73.69%	90.64%	98.52%
WebMAUS	**0.049**	**0.024**	0.108	**25.91%**	**51.39%**	**75.97%**	**91.61%**	98.75%
Aeneas	0.114	0.046	0.202	15.93%	33.56%	53.58%	73.61%	95.28%
Wav2Vec2 (XLSR)+CTC s.	0.068	0.036	0.123	16.84%	38.54%	65.07%	86.86%	97.87%
Wav2Vec2 (VoxRex)+CTC s.	**0.049**	0.034	**0.070**	17.01%	38.99%	66.89%	90.80%	**99.55%**

Table 5.2.: Descriptive statistics of boundary displacement and accuracies at different tolerance thresholds for boundary displacement between force-aligned boundaries and gold-standard annotations for the semi-spontaneous dataset. The best results are bolded.

When WebMAUS is excluded (due to the bias), it is visible that Wav2Vec2 with the VoxRex model performed best compared to the other forced aligners, having a mean Boundary Displacement (*BD*) error of 0.049 seconds and a standard deviation of 0.070. Even though MFA has a median of 0.027 seconds, it has one of the highest standard deviations, which indicates that the aligned data is widely spread and that the part of aligned words can be far from the mean, as can be seen in the Figure 5.8a. On the contrary, a low standard deviation indicates that the data is more tightly grouped around the mean, as in the case of Wav2Vec2 (VoxRex), which is visible in Figure 5.8e.

The Montreal Forced Aligner (MFA) (along with WebMAUS) outperforms deep learning models at 0.01, 0.025 and 0.05 tolerance thresholds. The amount of error under the mentioned tolerance thresholds positively affects the reported means and medians for MFA and WebMAUS by decreasing the final values. It can be observed that Wav2Vec2 (VoxRex) starts to exceed MFA from 0.1 seconds tolerance, however, the differences between the accuracies of these aligners are low, up to 2%. Wav2Vec2 (VoxRex) reaches almost 100% of the correct ratio of predictions which are at the maximum of 0.5 seconds apart from the gold standard.

In comparison to the VoxRex model, the Wav2Vec2 with the fine-tuned XLSR model performs worse than WebMAUS and MFA at all tolerance thresholds. However, the standard deviation is 2.5 times smaller than MFA's, which – again – demonstrates that the MFA's large number of errors within 0.05 and 0.1 thresholds affect the overall scores for the aligner. As can be seen in the histograms for both Wav2Vec2 models (Figures 5.8d and 5.8e), there are differences in the distribution of errors. However, both Wav2Vec2 models gain more than 86% of accuracy at the $<$0.1 seconds threshold.

Aeneas was outperformed by the other forced aligners almost on every level, mostly due to not annotating pauses (see Section 5.1). The standard deviations of MFA and Aeneas are comparable. However, as can be seen in Fig. 5.8a, it seems plausible that the standard deviation of MFA was caused by some outliers. The median of MFA, which is 0.027 seconds, also confirms that.

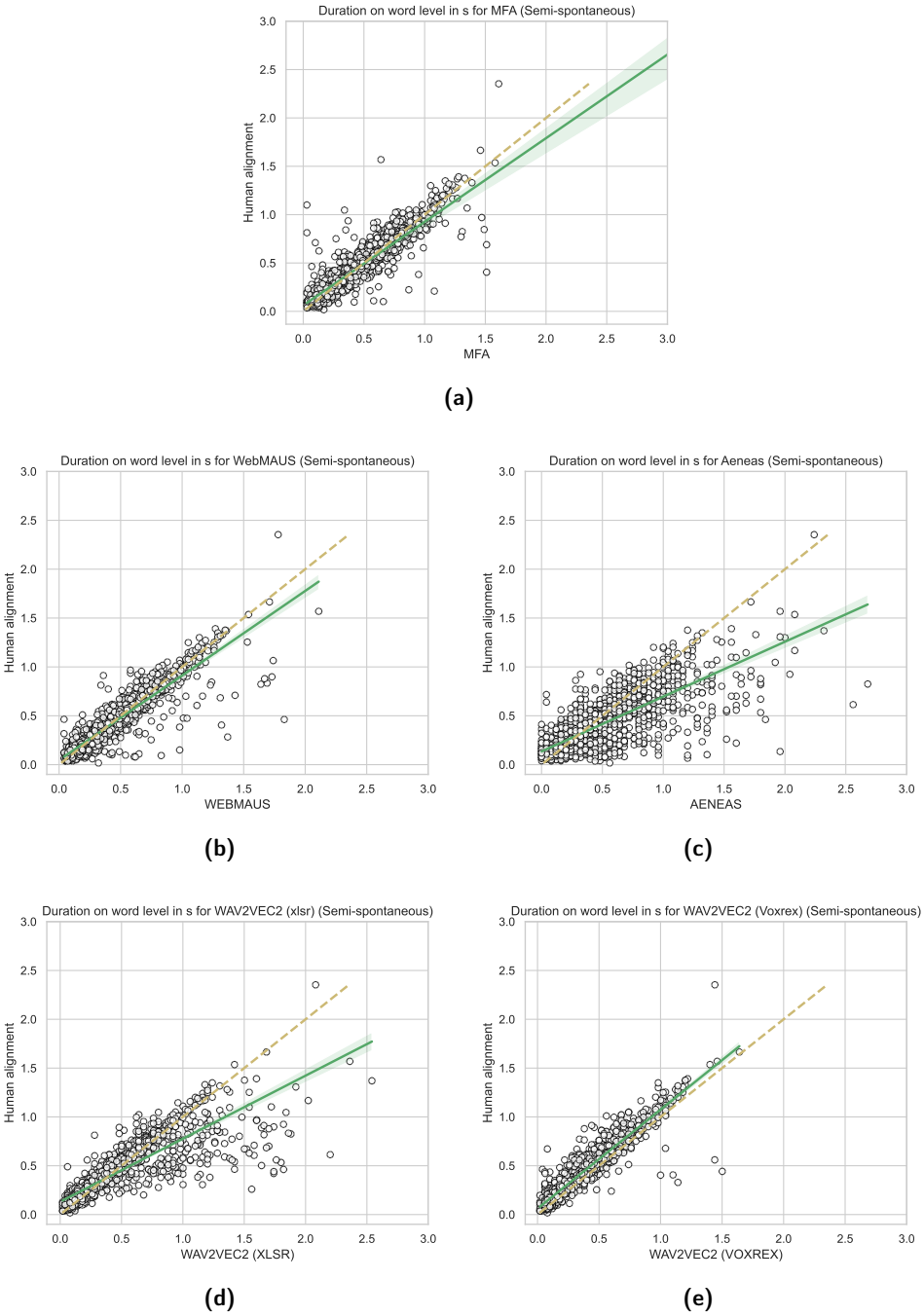


Figure 5.3.: Linear regression lines the between the duration of words in human alignment and the duration of words predicted by the aligners for the semi-spontaneous dataset. The dashed gold line represents a perfect fit. Confidence interval=95%

Aligner	**Slope**	**Intercept**	$\mathbf{R^2}$	***p*-value**
MFA	0.86	0.06	0.81	<0.01
WebMAUS	0.87	0.05	0.85	<0.01
Aeneas	0.56	0.14	0.59	<0.01
Wav2Vec2 (XLSR)+CTC s.	0.65	0.13	0.73	<0.01
Wav2Vec2 (VoxRex)+CTC s.	1.01	0.06	0.89	<0.01

Table 5.3.: Regression statistics for the semi-spontaneous dataset.

Figure 5.3 illustrates linear regression plots between the duration of words in human alignment and the duration approximated by the aligners. The regression statistics is displayed in Table 5.3.

The independent variable is statistically significant due to all p-values being smaller than .01. It can be seen in Figure 5.3 and Table 5.3 that all slopes are positive, which means that the two variables are positively related, that is, as the duration in the gold standard increases, so do the duration predicted by an aligner.

The coefficient of determination (R^2) (given in Tab. 5.3) indicates how much of the variation in the data values are covered by the model and hence how much of the observed variation can be explained by the independent variable. The larger coefficient, the better model fit. Figure 5.3 and Table 5.3 show that Wav2Vec2 (VoxRex) had the largest R^2 of 0.89, which means that the model expected that 89% of the observed variation could be explained by the independent variable WAV2VEC2 (VOXREX). Meanwhile, Aeneas has the lowest R^2 of 0.59, meaning that 41% of the variation was not explained by the model. The slope for Wav2Vec2 (VoxRex) was 1.06, that is, for every step of increase (i.e., 1 second of word duration) on the x-axis (WAV2VEC2 (VOXREX)), the linear regression model expected an increase of 1.06 seconds on the y-axis (THE HUMAN ALIGNMENT). The intercept was 0.06, that is, the model expected that a word that had a duration of zero seconds in the human alignment would have a duration of 0.06 seconds for Wav2Vec2 (VoxRex). The same intercept was observed for MFA, while it was 0.05 seconds for WebMAUS.

Aeneas produced the highest number of outliers (Fig. 5.3c) and similar behaviour can be observed in the predictions of Wav2Vec2 (XLSR) (Fig. 5.3d).

5.2.2. Spontaneous speech data

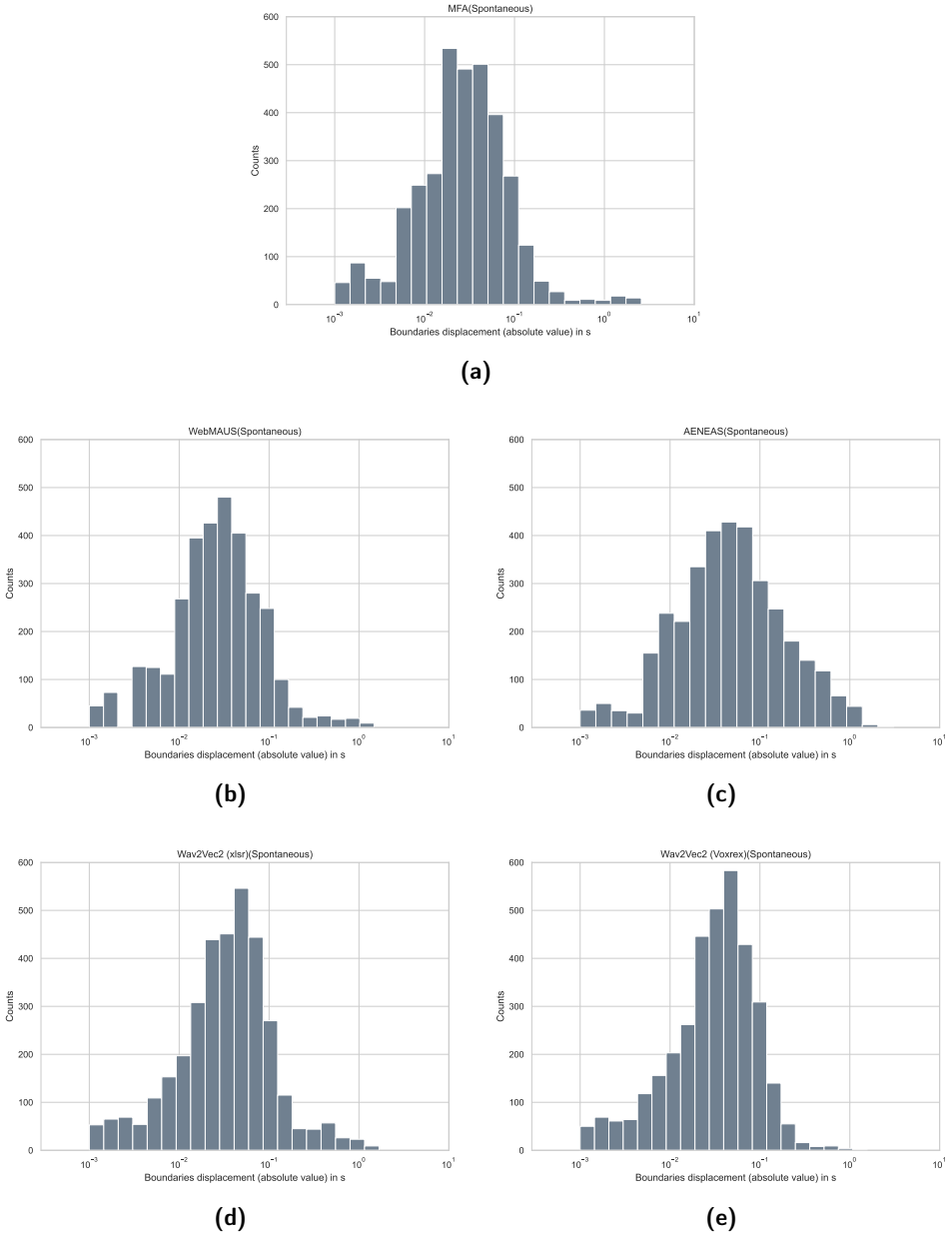


Figure 5.4.: Histograms of absolute boundary displacement (on log scale) between force-aligned word boundary and gold-standard annotations for the spontaneous dataset. Due to the log scale, if a boundary displacement error equalled 0, the value was not visualized in the histogram.

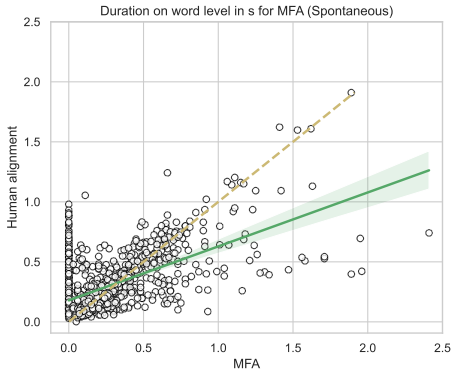
Aligner	BD			Tolerance				
	Mean (s)	Median (s)	SD (s)	<0.01s	<0.025s	<0.05s	<0.1s	<0.5
MFA	2.595	0.042	6.535	19.14%	36.84%	53.83%	65.31%	77.55%
WebMAUS	0.152	**0.035**	0.317	**23.60%**	**39.23%**	57.70%	70.18%	91.80%
Aeneas	0.253	0.072	0.407	12.44%	25.27%	41.91%	56.98%	83.77%
Wav2Vec2 (XLSR)+CTC s.	0.090	0.043	0.155	13.76%	31.22%	56.86%	79.43%	96.45%
Wav2Vec2 (VoxRex)+CTC s.	**0.065**	0.038	**0.109**	15.31%	35.65%	**60.73%**	**84.77%**	**98.84%**

Table 5.5.: Descriptive statistics of boundary displacement and accuracies at different tolerance thresholds for boundary displacement between force-aligned boundaries and gold-standard annotations for the spontaneous dataset. The best results are bolded.

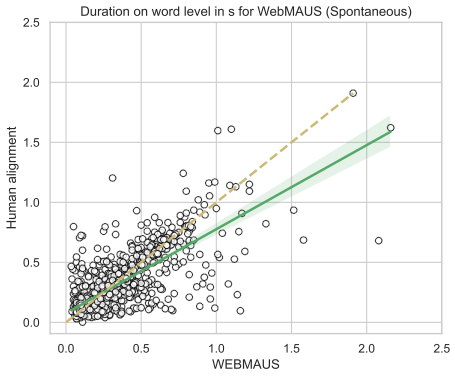
Table 5.5 presents the summary statistics for the spontaneous dataset. As was mentioned in Section 5.1, MFA did not produce any alignment for three out of 17 chunks of the spontaneous data, which affected the results (see Fig. 5.4a).

Another aspect of Table 5.5 is that even though the timings of the gold standard are biased towards WebMAUS, the deep learning models achieve better scores (except for the median of the XLSR model). Since the median difference is comparable for both WebMAUS and Wav2Vec2 (VoxRex), it implies that the main difference between these aligners is significantly bigger misalignments for WebMAUS. The difference between the standard deviations of the traditional forced alignment methods (MFA: 6.535, WebMAUS: 0.317, Aeneas: 0.407) and the deep learning models (Wav2Vec2 (VoxRex): 0.109, Wav2Vec2 (XLSR): 0.155) illustrates that data points from the traditional methods are not consistent, but more spread out over a large range of values.

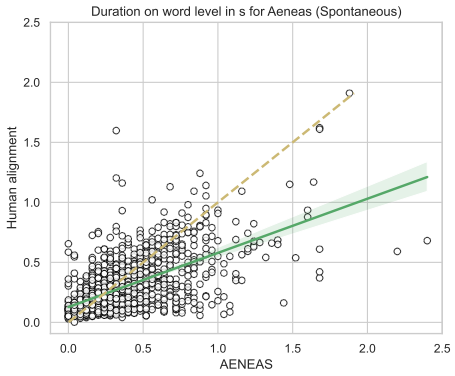
Table 5.5 presents also comparisons in the percentage of BD errors of the forced aligners at different thresholds. It can be seen that all forced aligners, except Aeneas, have similar accuracy scores within the range 30%-40% at 0.025 seconds tolerance and within the range 50%-60% at 0.05 seconds tolerance. It is worth noticing that accuracies of Aeneas were worse than MFA up to 0.5 seconds tolerance. The deep learning models were the only ones that had an accuracy larger than 95% at 0.5 seconds tolerance. Overall, these results indicate that the deep learning models can still produce excellent alignments despite noisy data.



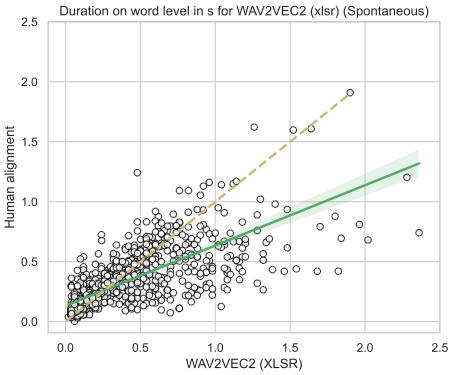
(a)



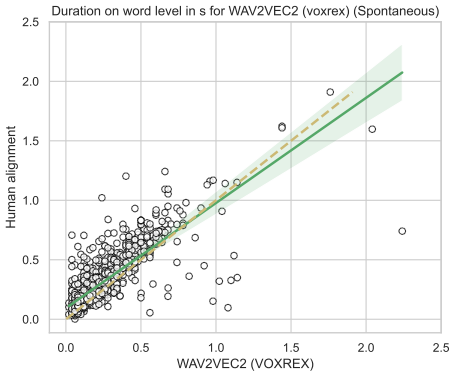
(b)



(c)



(d)



(e)

Figure 5.5.: Linear regression lines the between the duration of words in human alignment and the duration of words predicted by the aligners for the spontaneous dataset. The dashed gold line represents a perfect fit. Confidence interval=95%.

Aligner	**Slope**	**Intercept**	$\mathbf{R^2}$	***p*-value**
MFA	0.45	0.18	0.34	<0.01
WebMAUS	0.7	0.08	0.58	<0.01
Aeneas	0.45	0.13	0.36	<0.01
Wav2Vec2 (XLSR)+CTC s.	0.5	0.13	0.56	<0.01
Wav2Vec2 (VoxRex)+CTC s.	0.88	0.09	0.67	<0.01

Table 5.6.: Regression statistics for the spontaneous dataset.

Figure 5.5 shows the linear regression plots for each of the aligners. The duration of words predicted by an aligner is on the x-axis, while the duration of words in the gold standard is plotted on the y-axis. The regression statistics are displayed in Table 5.6, where it can be seen that all p-values are smaller than .01, hence independent variable is statistically significant.

As can be seen in Table 5.6, the model for MFA expected the lowest R^2 of 0.34, which was considerably lower compared to Wav2Vec2 (VoxRex), where R^2 was 0.67. In other words, according to the model, 68% of the observed variation was not explained by the data produced by MFA. The words, for which MFA did not produce any alignments, can be noticed in Figure 5.5a.

Once again, Wav2vec (VoxRex) had the biggest slope, which was 0.88. The linear regression model expected the lowest increase of 0.45 seconds on the y-axis, for every step of increase on the x-axis, for both MFA and Aeneas.

The intercept for MFA was 0.18, meaning that for a word with a duration of zero seconds in the gold standard, the model expected 0.18 seconds for MFA. Wav2Vec2 (VoxRex), along with WebMAUS had the lowest intercept, that is 0.09 and 0.08, respectively.

5.2.3. Read speech data

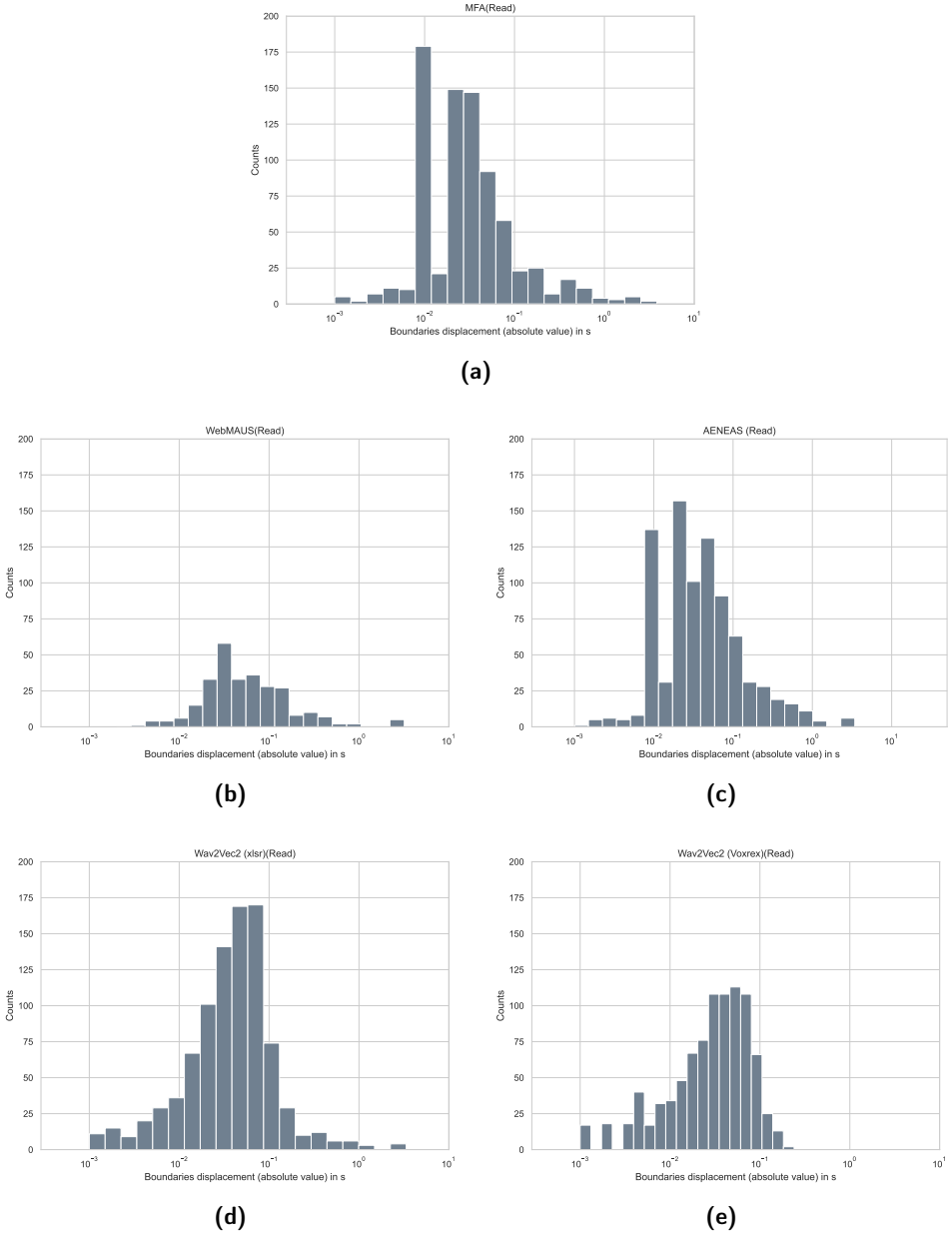


Figure 5.6.: Histograms of absolute boundary displacement (on log scale) between force-aligned word boundary and gold-standard annotations for the read dataset. Due to the log scale, if a boundary displacement error equalled 0, the value was not visualized in the histogram.

Aligner	BD			Tolerance				
	Mean (s)	Median (s)	SD (s)	<0.01s	<0.025s	<0.05s	<0.1s	<0.5
MFA	0.074	0.020	0.269	38.07%	55.36%	79.65%	90.04%	97.48%
WebMAUS	0.043	**0.000**	0.232	**70.79%**	**76.04%**	**84.90%**	91.25%	99.02%
Aeneas	0.094	0.030	0.277	23.63%	45.19%	69.37%	83.48%	96.72%
Wav2Vec2 (XLSR)+CTC s.	0.077	0.042	0.229	12.14%	31.73%	58.43%	88.73%	98.25%
Wav2Vec2 (VoxRex)+CTC s.	**0.041**	0.034	**0.033**	18.16%	39.28%	67.62%	**94.64%**	**100%**

Table 5.8.: Descriptive statistics of boundary displacement and accuracies at different tolerance thresholds for boundary displacement between force-aligned boundaries and gold-standard annotations for the read dataset. The best results are bolded.

It can be seen in Table 5.8 that Wav2Vec2 (VoxRex) again outperforms the other forced aligners, given the lower mean and standard deviation. MFA, along with WebMAUS,[1] had the lowest medians for *BD* errors, of 0.02 and 0 seconds, respectively. Since histograms are plotted on the log scale and the median of WebMAUS is 0, most of the values are not plotted and the distribution of errors is much smaller (see Fig. 5.6b). It is worth noticing that Aeneas had a lower median than the deep learning models, at the same time having the highest mean and standard deviation. Even though MFA managed to align the whole read dataset, Aeneas has a lower standard deviation of 0.08 compared to that of MFA. Moreover, it has a lower median than Wav2Vec2 (XLSR). Aeneas managed to exceed 95% of accuracy at the 0.5 seconds threshold and created better alignments up to the 0.05 threshold than both of the deep learning models.

The deep learning models produced exceptionally better alignments at 0.5 seconds tolerance as opposed to 0.01 seconds. The accuracy for the XLSR and VoxRex models rose by more than 80 percentage points. Moreover, Wav2Vec2 (VoxRex) managed to score an accuracy of 100% at the 0.5 seconds tolerance threshold, which can be seen in Figure 5.6e.

[1] The high median of WebMAUS is most likely the consequence of using the aligner for the gold standard.

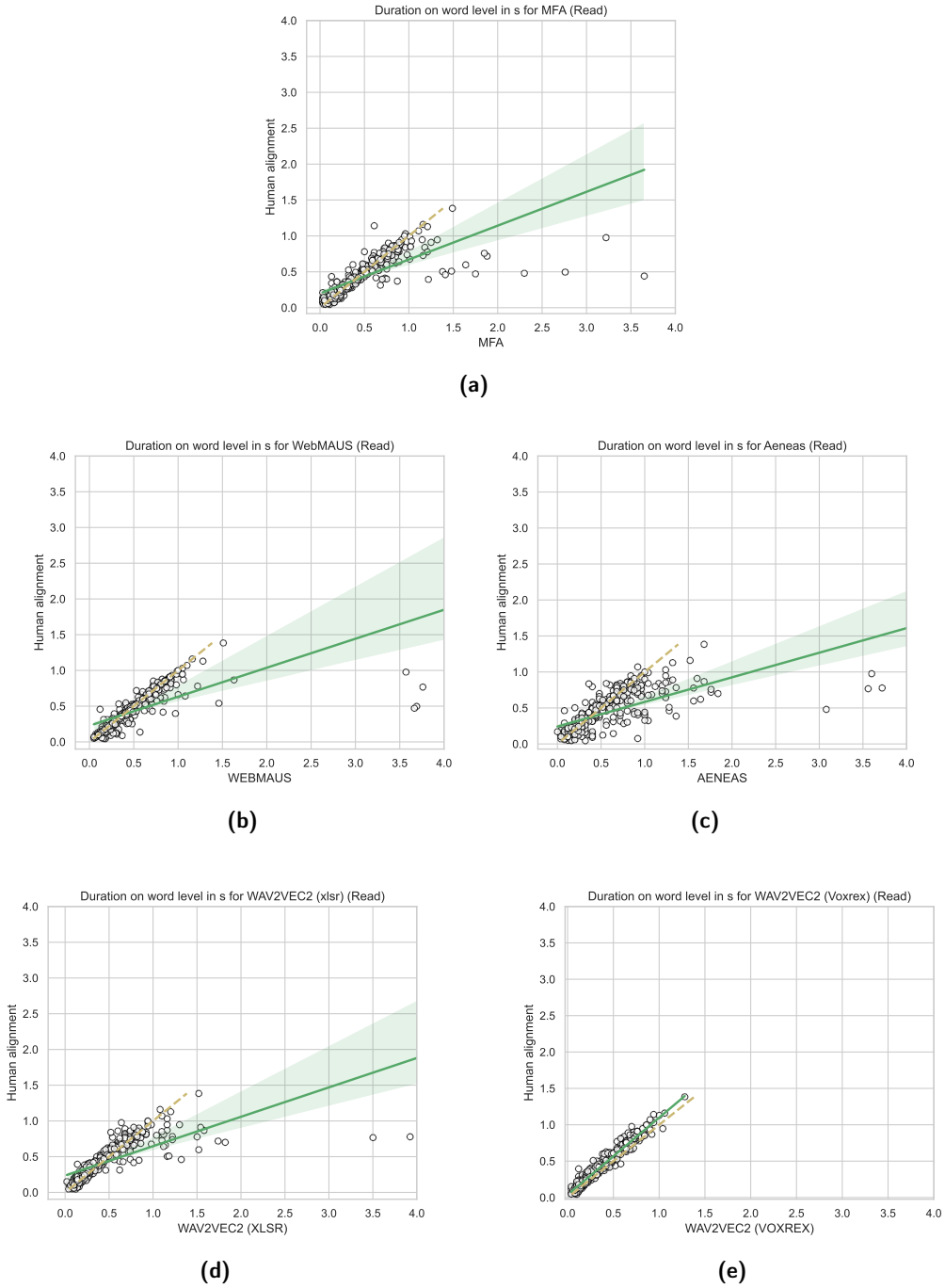


Figure 5.7.: Linear regression lines the between the duration of words in human alignment and the duration of words predicted by the aligners for the read dataset. The dashed gold line represents a perfect fit. Confidence interval=95%.

Aligner	**Slope**	**Intercept**	$\mathbf{R^2}$	***p*-value**
MFA	0.47	0.20	0.53	<0.01
WebMAUS	0.41	0.23	0.48	<0.01
Aeneas	0.34	0.24	0.46	<0.01
Wav2Vec2 (XLSR)+CTC s.	0.41	0.24	0.53	<0.01
Wav2Vec2 (VoxRex)+CTC s.	1.06	0.04	0.96	<0.01

Table 5.9.: Comparison of regression coefficients for the read dataset.

Figure 5.7 illustrates the linear regression plots for each of the aligners. The duration of words predicted by an aligner is on the x-axis, while the duration of words in the gold standard is plotted on the y-axis. The null hypothesis states that there is no correlation between the variables. The regression statistics is displayed in Table 5.9. The independent variable is statistically significant due to all p-values being smaller than .01.

In Figure 5.3 and Table 5.3, it can be noticed that once again Wav2Vec2 (VoxRex) had the largest R^2 of 0.96, which means 96% of the variation was explained by the model. When compared with other coefficients of determination (R^2), the difference is striking.

The slope for Wav2Vec2 (VoxRex) was 1.06, that is, for every step of increase (i.e., 1 second of word duration) on the x-axis (WAV2VEC2 (VOXREX)), the linear regression model expected an increase of 1.06 seconds on the y-axis (THE HUMAN ALIGNMENT). The intercept was 0.04, that is, the model expected that a word that had a duration of zero seconds in the human alignment would have a duration of 0.04 seconds for Wav2Vec2 (VoxRex). The highest intercept was observed for Aeneas and Wav2Vec2 (XLSR) with a score of 0.24. The intercept for WebMAUS is only one percentage point lower than Aeneas and Wav2Vec2 (XLSR).

5.2.4. Summary

Table 5.11 presents the descriptive statistics and tolerance thresholds for the entire corpus at the word level.

Aligner	**BD**			**Tolerance**				
	Mean (s)	**Median (s)**	**SD (s)**	**<0.01s**	**<0.025s**	**<0.05s**	**<0.1s**	**<0.5**
MFA	0.978	0.030	4.116	23.50%	44.97%	67.30%	81.41%	90.80%
WebMAUS	0.086	**0.025**	0.228	**30.99%**	**50.25%**	**70.54%**	83.81%	96.27%
Aeneas	0.161	0.050	0.309	15.68%	32.26%	51.44%	68.90%	91.31%
Wav2Vec2 (XLSR)+CTC s.	0.077	0.039	0.156	15.11%	35.00%	61.23%	84.42%	97.41%
Wav2Vec2 (VoxRex)+CTC s.	**0.054**	0.035	**0.084**	16.55%	37.82%	64.76%	**88.87%**	**99.35%**

Table 5.11.: Descriptive statistics of boundary displacement and accuracies at different tolerance thresholds for boundary displacement between force-aligned boundaries and gold-standard annotations for the whole Swedish dataset. The best results are bolded.

As can be seen, both Wav2Vec2 models outperformed the other forced aligners, concerning mean and standard deviation. Wav2Vec2 (VoxRex) has the lowest mean and the standard deviation (0.054 and 0.084, respectively). WebMAUS got the lowest median, which was most likely caused by being biased towards the gold standard. MFA got one of the lowest medians (0.030 seconds), however, mean and standard deviation were visibly affected by the chunks of data that were not aligned. Aeneas got the highest median of 0.050 seconds.

Table 5.11 also shows comparisons in the percentage of BD errors of the forced aligners at different thresholds for the whole corpus. In lower thresholds, that is 0.01, 0.025 and 0.05 seconds, MFA and WebMAUS surpassed the other forced aligners. From the 0.1 seconds tolerance threshold, the deep learning models achieved better accuracy than the other forced aligners. Wav2Vec2 (XLSR) achieved an accuracy of almost 100% at the 0.5 seconds threshold.

5.3. Evaluations and comparisons (utterance-level)

This section presents results for the Swedish alignments at the utterance level. Figures 5.8, 5.10 and 5.12 show the distribution of manual/force aligned boundary

displacement, for each of the forced aligners. The x-axis represents the word boundary displacement (in seconds), while the y-axis represents the number of words. Tables 5.13, 5.16 and 5.19 provide mean, median and the standard deviation of manual/aligned boundary displacement values for each aligner, as well as the accuracy measurements for utterance boundary displacement values at different tolerances. Figures 5.9, 5.11, 5.13 show linear regression plots between the duration of utterances in human alignment and the duration approximated by the aligners and Tables 5.14, 5.17, 5.20 compare regression coefficients. The null hypothesis states that there is no correlation between the variables. When all p-values are smaller than .01, the null hypothesis is rejected and the independent variable is statistically significant.

5.3.1. Semi-spontaneous speech data

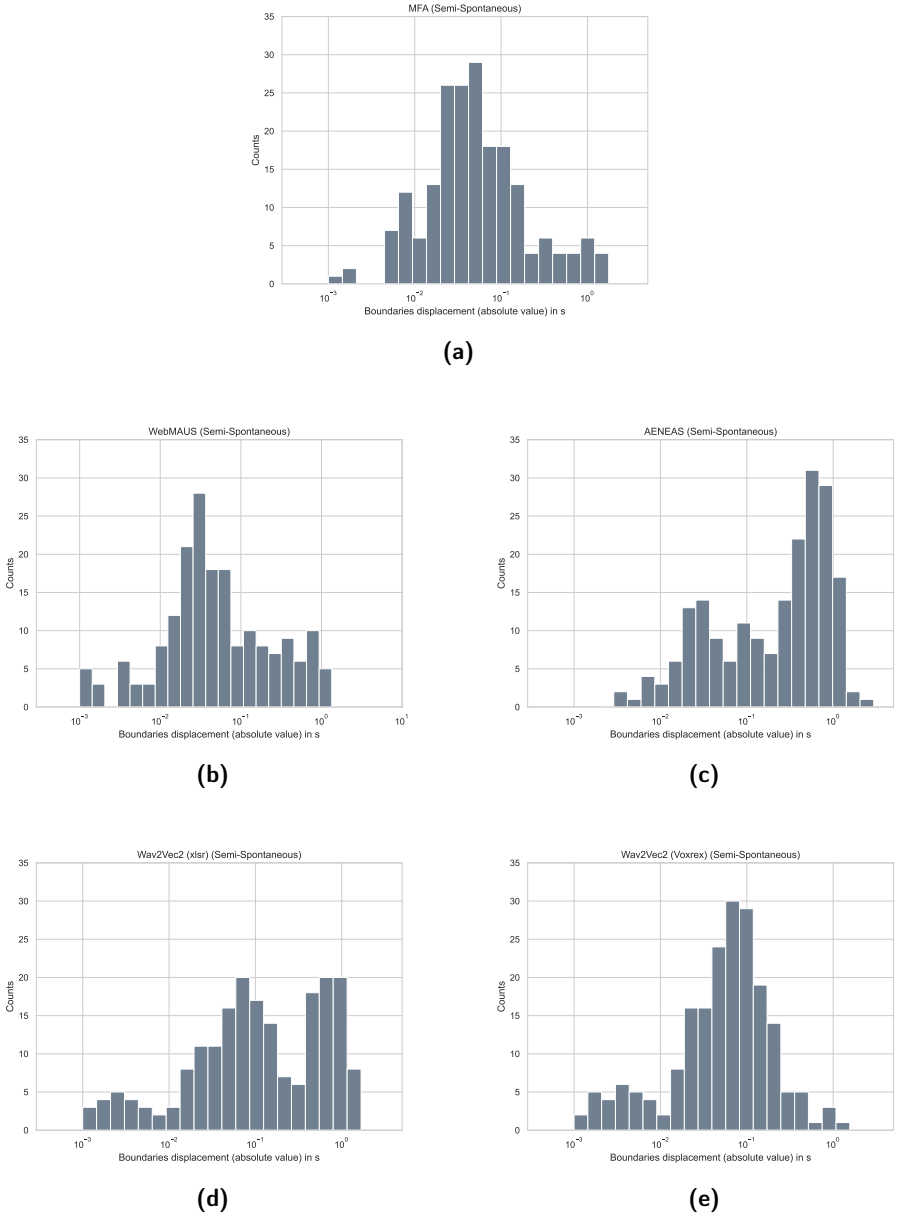


Figure 5.8.: Histograms of absolute utterance boundary displacement (on log scale) between force-aligned boundaries and gold-standard annotations for the semi-spontaneous dataset at the utterance level. Due to the log scale, if a boundary displacement error equalled 0, the value was not visualized in the histogram.

Aligner	**BD**			**Tolerance**				
	Mean (s)	**Median (s)**	**SD (s)**	**<0.01s**	**<0.025s**	**<0.05s**	**<0.1s**	**<0.5**
MFA	0.139	0.044	0.283	12.87%	32.18%	57.43%	75.25%	92.57%
WebMAUS	0.149	**0.036**	0.264	**18.32%**	**37.13%**	**59.41%**	71.29%	89.60%
Aeneas	0.425	0.350	0.430	4.95%	14.85%	24.75%	33.66%	63.86%
Wav2Vec2 (XLSR)+CTC s.	0.307	0.104	0.377	11.39%	20.79%	32.67%	49.50%	73.27%
Wav2Vec2 (VoxRex)+CTC s.	**0.105**	0.060	**0.174**	14.26%	26.24%	42.57%	**72.28%**	**97.52%**

Table 5.13.: Descriptive statistics of boundary displacement and accuracies at different tolerance thresholds for boundary displacement between force-aligned boundaries and gold-standard annotations for the semi-spontaneous dataset. The best results are bolded.

Table 5.13 compares the descriptive statistics of boundary displacement for the semi-spontaneous dataset at the utterance level.

If all three measures are taken into consideration, Wav2Vec2 (VoxRex) seems best with a mean of 0.105 seconds and a standard deviation of 0.174, even though it has a lower median than WebMAUS. The results of Wav2Vec2 (VoxRex) at the utterance-level overlap with the results at the word level, that is, the lowest mean and standard deviation and a median higher than MFA and WebMAUS. MFA has a higher median than WebMAUS while having a lower mean.

Despite having a lower standard deviation than MFA, Wav2Vec2 (XLSR) has a very high mean and median (0.425 and 0.350 seconds, respectively). The results of Wav2Vec2 (XLSR) at the utterance level are visibly worse than at the word level. The median of Aeneas is more than seven times larger than the median of MFA.

Table 5.13 also presents the accuracy measurements for boundary displacement values at different tolerances. Once again, the Montreal Forced Aligner (MFA) (along with WebMAUS), outperforms deep learning models at small tolerance thresholds. In comparison to the results at the word level, MFA has a higher accuracy at 0.1 seconds tolerance than Wav2Vec2 (XLSR) and outperforms WebMAUS at 0.1 and 0.5 seconds tolerance. Wav2Vec2 (VoxRex) reaches 97.52% of predictions which are less than 0.5 seconds apart from the gold standard. WebMAUS has the highest accuracy at 0.01 and 0.025 seconds. Aeneas and Wav2Vec2 (XLSR) did not manage to exceed 75% at 0.5 seconds tolerance.

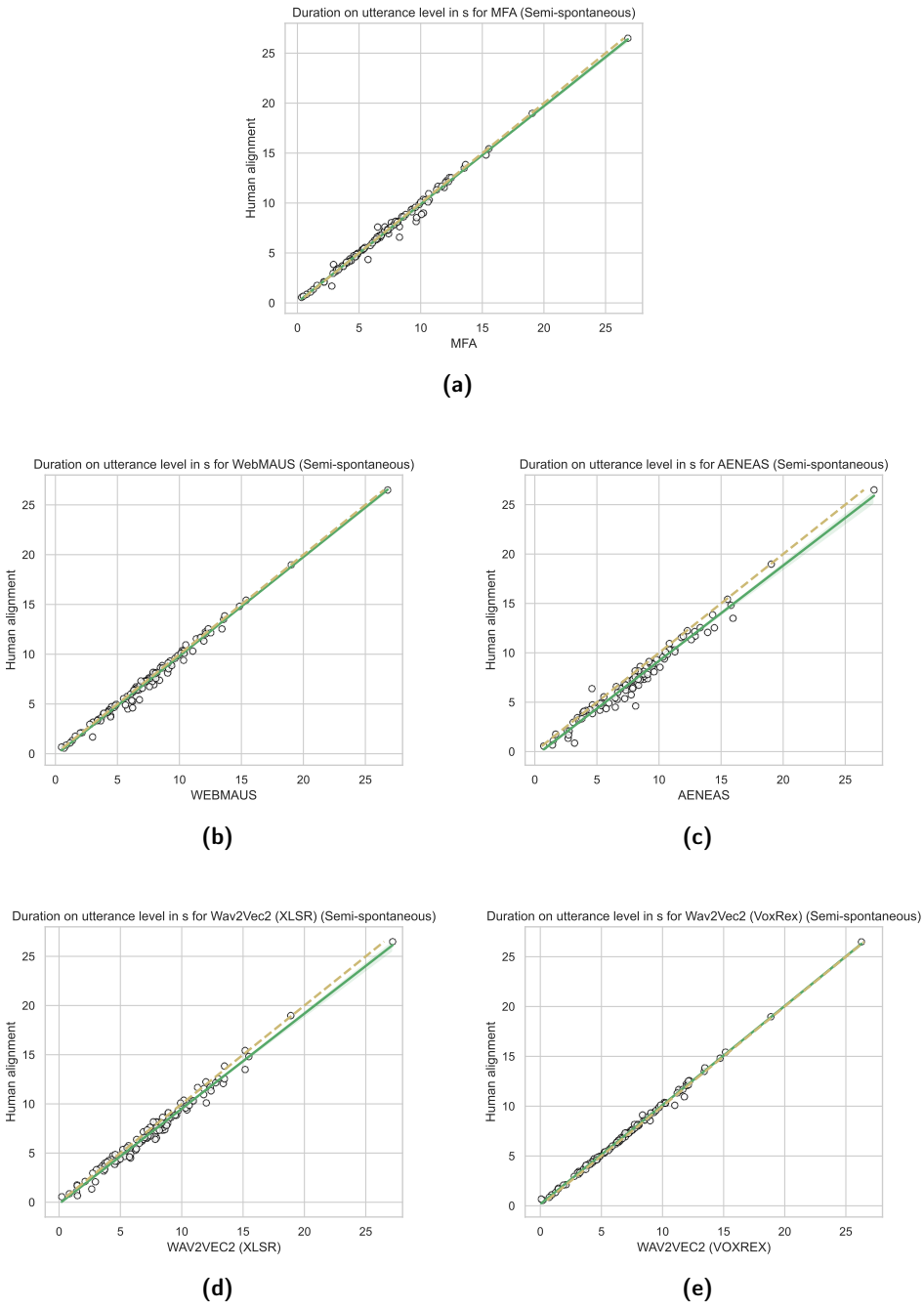


Figure 5.9.: Linear regression lines the between the duration of words in human alignment and the duration of utterances predicted by the aligners for the semi-spontaneous dataset. The dashed gold line represents a perfect fit. Confidence interval=95%

Aligner	**Slope**	**Intercept**	$\mathbf{R^2}$	***p*-value**
MFA	0.98	0.08	0.99	<0.01
WebMAUS	1.00	-0.14	0.99	<0.01
Aeneas	0.97	-0.47	0.97	<0.01
Wav2Vec2 (XLSR)+CTC s.	0.97	-0.16	0.98	<0.01
Wav2Vec2 (VoxRex)+CTC s.	0.99	0.17	1.00	<0.01

Table 5.14.: Regression statistics for the semi-spontaneous dataset at the utterance level.

Figure 5.9 illustrates linear regression plots between the duration of utterances in human alignment and the duration approximated by the aligners. The regression statistics is displayed in Table 5.14. The independent variable is statistically significant due to all p-values being smaller than .01.

The coefficient of determination (R^2) (given in Tab. 5.14) indicates how much of the variation in the data values is covered by the model and hence how much of the observed variation can be explained by the independent variable. The larger coefficient, the better model fit. Figure 5.9 and Table 5.14 show that Wav2Vec2-VoxRex had the largest R^2 of 1, which means that the model expected that 100% of the variation in the data could be explained. Meanwhile, Aeneas had the lowest R^2 of 0.97, meaning that 3% of the variation was not explained by the model. Wav2Vec2 (Voxrex) and MFA had the positive values of intercept whereas the other forced aligners had negative values.

The slope for Wav2Vec2-VoxRex was 1, that is, for every step of increase (i.e., 1 second of word duration) on the x-axis (WAV2VEC2), the linear regression model expected an increase of 1 second on the y-axis (the human alignment).

5.3.2. Spontaneous speech data

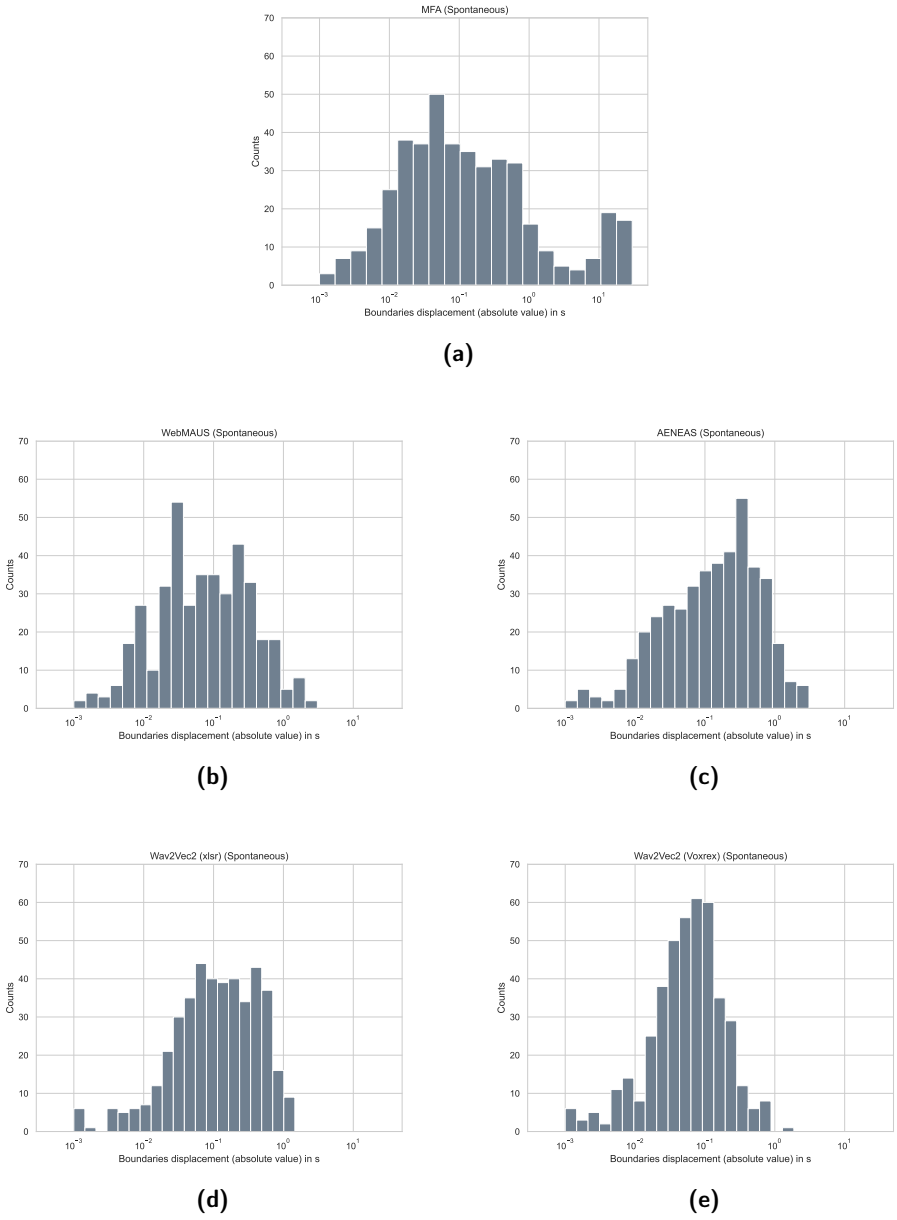


Figure 5.10.: Histograms of absolute utterance boundary displacement (on log scale) between force-aligned utterance boundaries and gold-standard annotations for the spontaneous dataset at the utterance level. Due to the log scale, if a boundary displacement error equalled 0, the value was not visualized in the histogram.

Aligner	BD			Tolerance				
	Mean (s)	Median (s)	SD (s)	<0.01s	<0.025s	<0.05s	<0.1s	<0.5
MFA	1.984	0.093	5.558	11.98%	26.50%	40.32%	51.84%	75.81%
WebMAUS	0.189	**0.060**	0.341	**18.20%**	**29.26%**	**46.77%**	59.91%	90.55%
Aeneas	0.301	0.150	0.410	6.45%	18.66%	29.26%	40.78%	80.88%
Wav2Vec2 (XLSR)+CTC s.	0.224	0.116	0.264	7.37%	14.98%	29.26%	46.77%	85.71%
Wav2Vec2 (VoxRex)+CTC s.	**0.105**	0.063	**0.157**	10.83%	21.66%	42.17%	**67.97**	**97.47%**

Table 5.16.: Descriptive statistics of boundary displacement and accuracies at different thresholds for utterance boundary displacement between force-aligned boundaries and gold-standard annotations for the spontaneous dataset. The best results are bolded.

Table 5.16 compares the descriptive statistics of boundary displacement for the spontaneous dataset at the utterance level. Since MFA did not produce any alignment for three out of 17 chunks of the spontaneous data (see Section 5.1), the results are visibly affected, especially the standard deviation (5.558) and mean (1.984 seconds). Surprisingly, the median of MFA is smaller than the medians of Aeneas (0.301 seconds) and Wav2Vec2 (XLSR) (0.224 seconds). Even though the median and mean of WebMAUS are smaller than Wav2Vec2 (XLSR)'s, its standard deviation is much higher. Wav2Vec2 (VoxRex) outperformed the other forced aligners in every metric, by scoring the median of 0.063 seconds, the mean of 0.105 seconds and the standard deviation of 0.157 seconds.

The results for tolerance at 0.01 seconds are consistent with results obtained in Section 5.3.1 – WebMAUS, along with MFA obtained better results than the other forced aligners. What is worth noticing, Wav2Vec2 (VoxRex) outperformed the other forced aligners at the 0.1 seconds tolerance threshold, scoring 97.47% at the tolerance threshold of 0.5 seconds. It is worth noticing that the accuracies of Aeneas were worse than MFA up to 0.5 seconds tolerance.

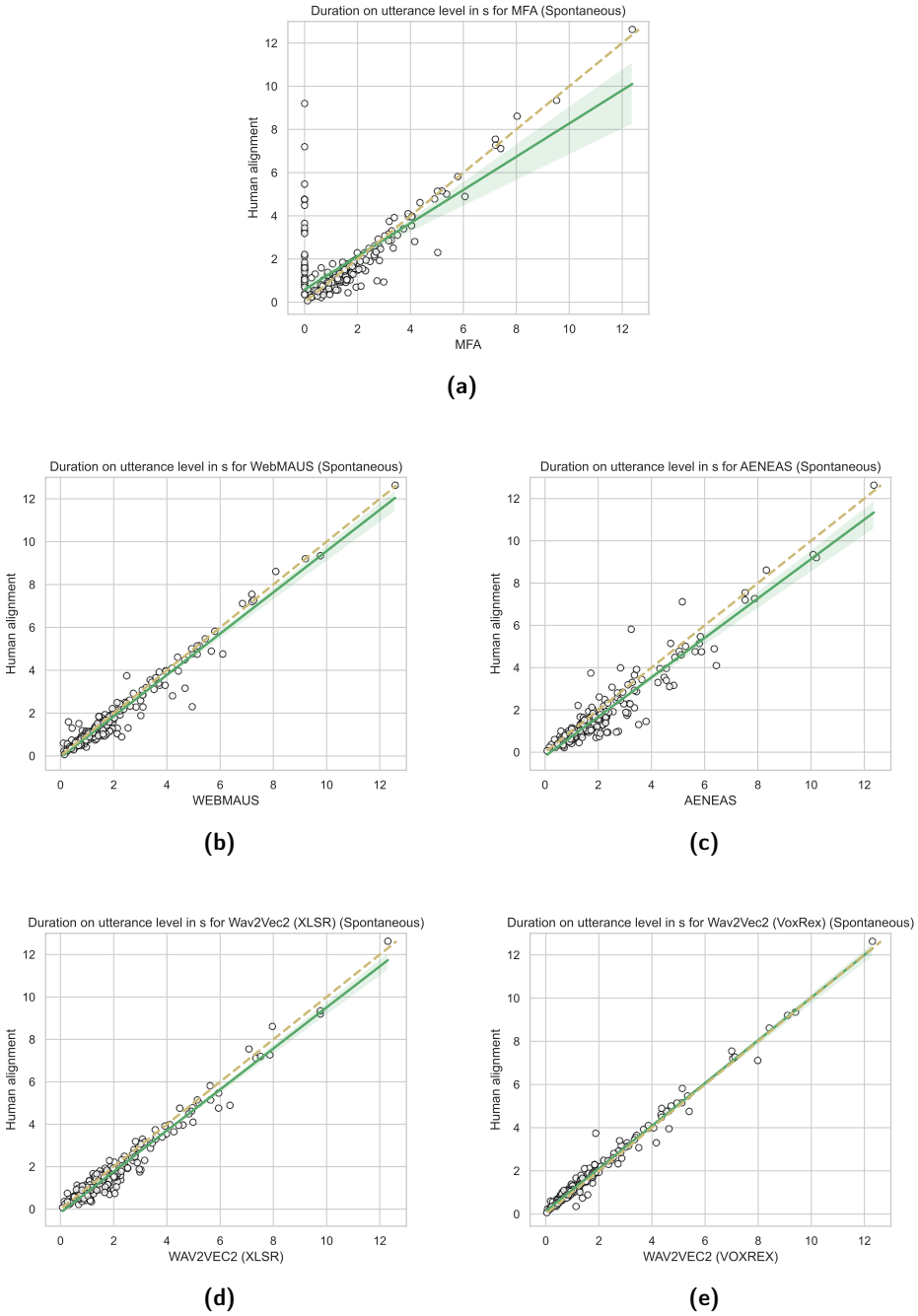


Figure 5.11.: Linear regression lines the between the duration of utterances in human alignment and the duration of utterances predicted by the aligners for the spontaneous dataset. The dashed gold line represents a perfect fit. Confidence interval=95%

Aligner	**Slope**	**Intercept**	$\mathbf{R^2}$	***p*-value**
MFA	0.77	0.58	0.55	<0.01
WebMAUS	0.96	-0.07	0.95	<0.01
Aeneas	0.93	-0.19	0.90	<0.01
Wav2Vec2 (XLSR)+CTC s.	0.97	-0.16	0.98	<0.01
Wav2Vec2 (VoxRex)+CTC s.	0.99	0.13	0.98	<0.01

Table 5.17.: Regression statistics for the spontaneous dataset at the utterance level.

Figure 5.11 illustrates linear regression plots between the duration of utterances in human alignment and the duration approximated by the aligners. The regression statistics is displayed in Table 5.17. The null hypothesis states that there is no correlation between the variables.

As listed in Tab. 5.17, all p-values are smaller than .01. In consequence, the null hypothesis, that there is no correlation between the variables, is rejected and the independent variable is statistically significant.

Figure 5.11 and Table 5.17 show that Wav2Vec2 (VoxRex), along with Wav2Vec2 (XLSR) had the largest R^2 of 0.98, which means that the model expected that 98% of the observed variation could be explained by the independent variable. Meanwhile, MFA has the lowest R^2 of 0.55, meaning that 45% of the variation was not explained by the model. The intercept of MFA was 0.58 that is, the model expected that a word that had a duration of zero seconds in the human alignment would have a duration of 0.58 seconds for MFA. The slope for Wav2Vec2-VoxRex was 0.99, while the intercept was 0.13.

MFA produced a number of outliers, which are visible in Figure 5.11a.

5.3.3. Read speech data

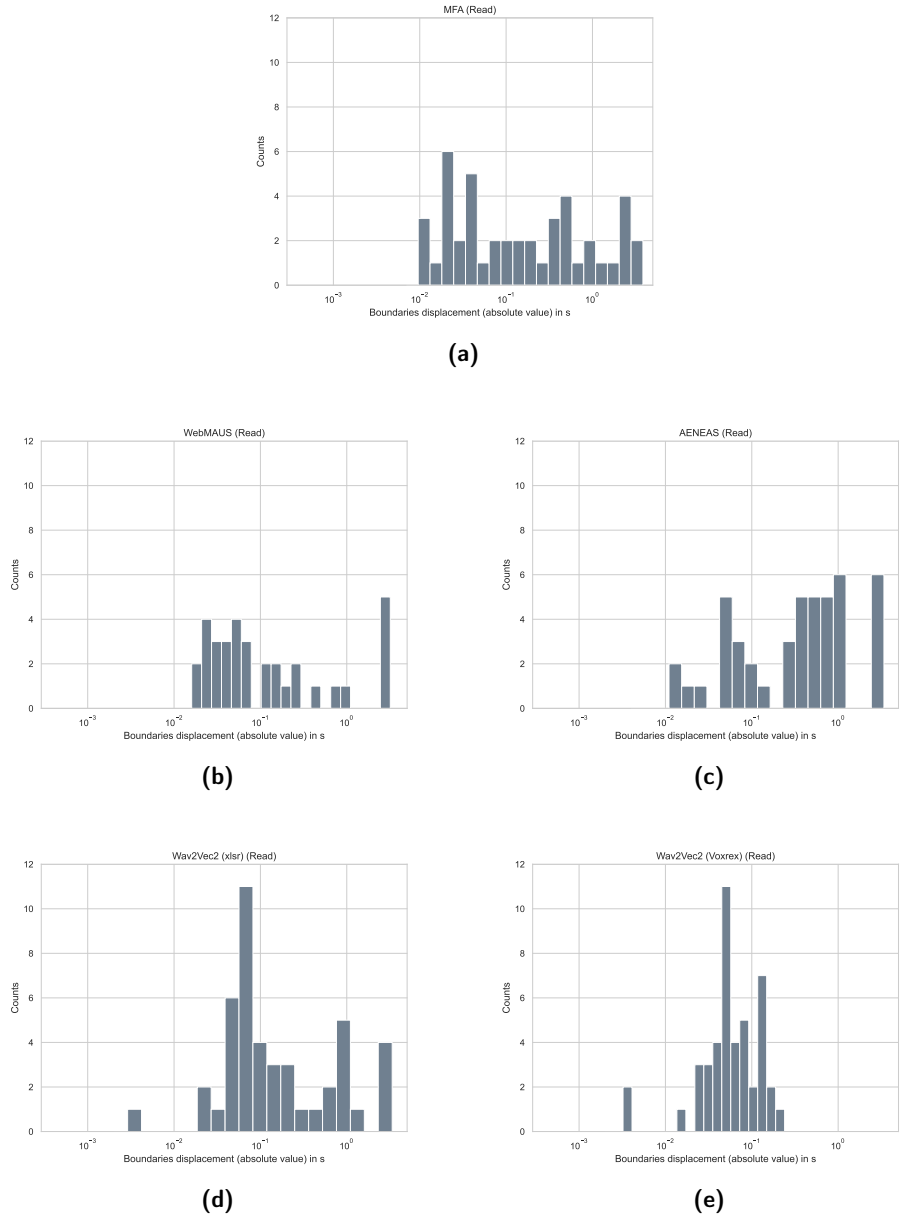


Figure 5.12.: Histograms of absolute utterance boundary displacement (on log scale) between force-aligned utterance boundaries and gold-standard annotations for the read dataset at the utterance level. Due to the log scale, if a boundary displacement error equalled 0, the value was not visualized in the histogram.

Aligner	BD			Tolerance				
	Mean (s)	Median (s)	SD (s)	<0.01s	<0.025s	<0.05s	<0.1s	<0.5
MFA	0.597	0.124	0.970	4.35%	23.91%	41.30%	45.65%	71.74%
WebMAUS	0.414	**0.037**	0.932	**26.09%**	**36.96%**	**54.35%**	67.39%	84.78%
Aeneas	0.747	0.434	0.948	2.17%	8.70%	21.74%	28.26%	54.35%
Wav2Vec2 (XLSR)+CTC s.	0.503	0.086	0.885	4.35%	8.70%	19.57%	56.52%	73.91%
Wav2Vec2 (VoxRex)+CTC s.	**0.071**	0.054	**0.048**	6.52%	13.04%	39.13%	**76.09%**	**100%**

Table 5.19.: Descriptive statistics of boundary displacement and accuracies at different thresholds for utterance boundary displacement between force-aligned boundaries and gold-standard annotations for the read dataset. The best results are bolded.

Table 5.19 compares the descriptive statistics of boundary displacement for the read dataset at the utterance level. It can be seen in Table 5.19 that Wav2Vec2 (VoxRex) again outperforms the other forced aligners, given the lowest mean and standard deviation. The statistics for the other forced aligners are characterized by a large standard deviation. For instance, the standard deviation of WebMAUS is almost twenty times larger than the standard deviation of Wav2Vec2 (VoxRex), even though the median of WebMAUS is lower than the median of Wav2Vec2 (VoxRex). Once again, Aeneas got the worst results with a median of 0.434 seconds and a mean of 0.747 seconds. However, the standard deviation of Aeneas is lower than the standard deviations of MFA.

Wav2Vec2 (VoxRex) outperformed the other forced aligners at two tolerance thresholds: 0.1 and 0.5 seconds. The accuracy of the VoxRex model increased by 36.96 percentage points at the 0.1 seconds threshold, when compared to the 0.05 seconds threshold. Moreover, it managed to score an accuracy of 100% at the 0.5 seconds tolerance threshold. WebMAUS got the best results within the range of 0.01 and 0.05 seconds tolerance and the second-best result at 0.5 seconds with an accuracy of 84.78%. The accuracy achieved by MFA at 0.01 seconds threshold is lower more than six times in comparison to the accuracy of WebMAUS. Moreover, Wav2Vec2 (VoxRex) outperformed MFA at the 0.01 seconds tolerance threshold by more than 2 percentage points. Aeneas did not exceed 55% of the accuracy at 0.5 seconds threshold.

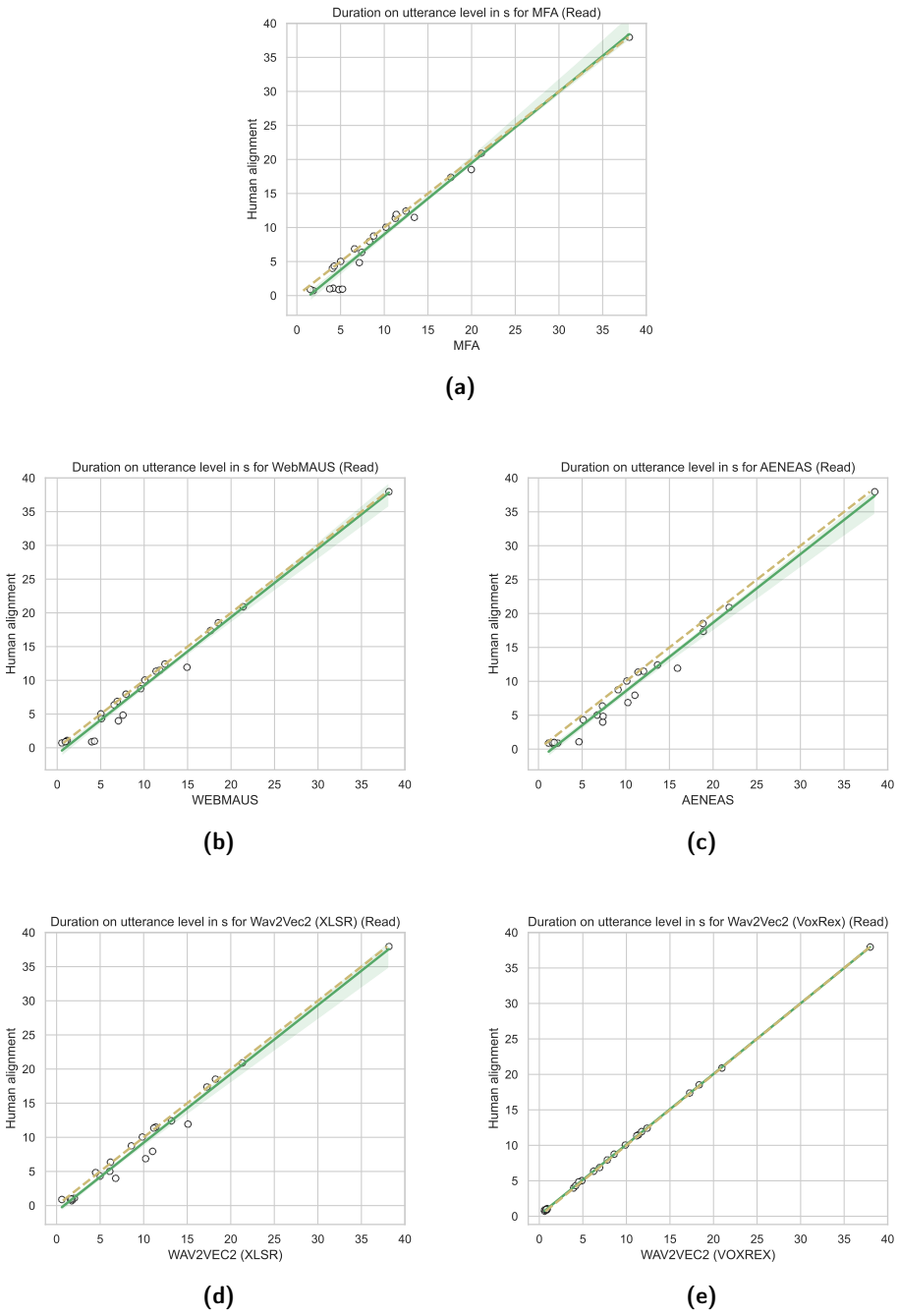


Figure 5.13.: Linear regression lines between the duration of utterances in human alignment and the duration of utterances predicted by the aligners for the read dataset. the dashed gold line represents perfect fit. Confidence interval=95%

Aligner	**Slope**	**Intercept**	$\mathbf{R^2}$	***p*-value**
MFA	1.05	-1.48	0.98	<0.01
WebMAUS	1.02	-0.95	0.98	<0.01
Aeneas	1.01	-1.54	0.98	<0.01
Wav2Vec2 (XLSR)+CTC s.	1.01	-0.85	0.98	<0.01
Wav2Vec2 (VoxRex)+CTC s.	0.99	0.17	1.00	<0.01

Table 5.20.: Regression statistics for read dataset at the utterance level.

Figure 5.13 illustrates linear regression plots between the duration of utterances in human alignment and the duration approximated by the aligners. The regression statistics is displayed in Table 5.20. The independent variable is statistically significant due to all p-values being smaller than .01.

Wav2Vec2 (VoxRex) had the largest R^2 of 1, which means that the model expected that 100% of the observed variation could be explained by the independent variable (WAV2VEC2-VOXREX). Meanwhile, all other forced aligners had R^2 of 0.98. MFA had the biggest slope, The intercept of Wav2Vec2 (VoxRex) was 0.06, that is, the model expected that a word that had a duration of zero seconds in the human alignment would have a duration of 0.06 seconds for Wav2Vec2 (VoxRex), whereas the intercepts of other forced aligners were negative.

Aeneas, along with Wav2Vec2 (XLSR) produced similar outliers (see Fig. 5.3c and 5.3d).

5.3.4. Summary

Table 5.22 presents the descriptive statistics for the entire Swedish corpus at the utterance level.

Aligner	BD			Tolerance				
	Mean (s)	Median (s)	SD (s)	<0.01s	<0.025s	<0.05s	<0.1s	<0.5
MFA	1.344	0.064	4.523	11.73%	28.01%	45.45%	58.36%	80.50%
WebMAUS	0.193	**0.047**	0.395	**18.77%**	**32.11%**	**51.03%**	63.78%	89.88%
Aeneas	0.419	0.248	0.507	4.55%	12.61%	21.85%	33.72%	69.50%
Wav2Vec2 (XLSR)+CTC s.	0.267	0.108	0.378	8.36%	16.28%	29.62%	48.24%	81.23%
Wav2Vec2 (VoxRex)+CTC s.	**0.103**	0.061	**0.157**	11.58%	22.43%	42.08%	**69.79%**	**97.65%**

Table 5.22.: Descriptive statistics of boundary displacement and accuracies at different tolerance thresholds for boundary displacement between force-aligned boundaries and gold-standard annotations for the whole Swedish dataset at the utterance level. The best results are bolded.

As in the case of word-level alignment, Wav2Vec2 (VoxRex) outperformed the other forced aligners in mean and standard deviation. Wav2Vec2 (VoxRex) had the lowest mean of 0.103 and a standard deviation of 0.157. Once again, WebMAUS had the lowest median, which was most likely caused by being biased towards the gold standard. MFA and Wav2Vec2 (VoxRex) got comparable medians (0.064 and 0.061 seconds), however, the mean and standard deviation of MFA were visibly affected by the non-aligned data. Aeneas got the worst median of 0.248 seconds. In lower thresholds, that is, 0.01, 0.025 and 0.05 seconds, MFA and WebMAUS surpassed the other forced aligners. From 0.1 seconds tolerance threshold, Wav2Vec2 (VoxRex) achieved better accuracy than the other forced aligners, scoring 69.79% at 0.1 seconds and 97.65% at 0.5 seconds.

5.4. Qualitative comparison

In order to compare the performance of the aligners at the word level, I examined a set number of words, which were outliers. Moreover, I also extracted a number of words with the biggest duration difference, calculated by *Weber's Law of Just Noticeable Difference* explained in Section 4.3. Using the methods described above, I obtained 45 words.

By analyzing the extracted words according to Weber's Law, it can be noticed that in the majority of cases the same words were problematic for the various aligners, such as:

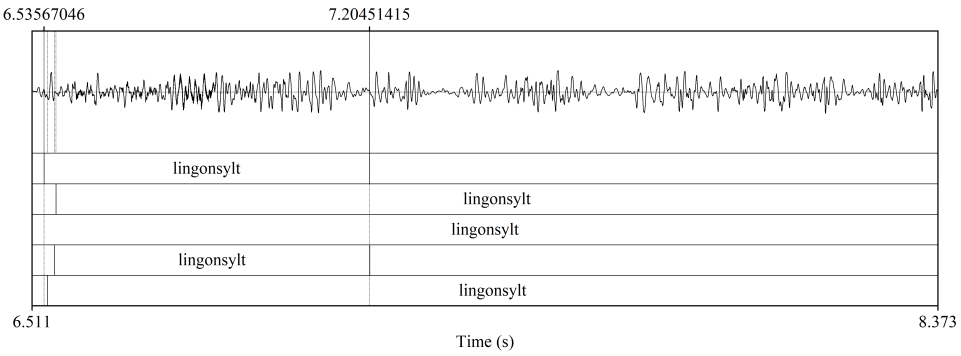


Figure 5.14.: Sample word alignment from the spontaneous dataset visualized in Praat TextGrid (Boersma, 2001). From top to bottom: gold standard, Aeneas, WebMAUS, Wav2Vec2-VoxRex and Wav2Vec2-XLSR. MFA did not produce the alignment.

- prepositions: *till*, *på*, *i*, *och*;
- pronouns: *han*, *hans*, *vad*;
- interjections: *hej*;
- conjunctions: *att*, *för*;
- articles: *en*, *ett*, *den*, *det*;
- verbs: *är*, *har*.[2]

The outliers of Wav2Vec2-VoxRex were mostly word fillers (so-called *filled pauses*), such as *ehm*, *eh* for all three metrics. In the case of the read dataset, other words were the worst outliers, since word fillers were not in the data. For the read dataset, the biggest misalignments were: *att* ("to") for *BD* with the misalignment of 0.278 seconds and *säkerhet* ("safety") for the duration difference with the misalignment of 0.276 seconds. However, according to Weber's Law, *på* ("on") along with words, which were listed above, had the biggest temporal errors. For comparison, for MFA, the biggest misalignment according to *BD* errors was *artikel* ("article") with the misalignment of 4.906 seconds, the same word was listed as the biggest misalignment according to the duration difference and Weber's Law. *Artikel* ("Article"), along with *suveränitet* ("sovereignty") and *säkerhet* ("safety") were one of the most problematic words for other aligners in the read dataset.

For the spontaneous dataset, besides the words listed above, I noticed reoccurring words, such as nouns (*lingonsylt* ("lingonberry jam"), *brunsås* ("brown sauce")) or adjectives (*svenskt* ("Swedish"), *svart* ("black")). The alignment of the word *lingonsylt* ("lingonberry jam") is visualized in Figure 5.14.

For the semi-spontaneous dataset, besides the listed words, the outliers vary from aligner to aligner. The biggest misalignments are – among others – nouns such as *svaret* ("the answer") or *året* ("the year"). For Wav2Vec2 (VoxRex), as in the case of the semi-spontaneous dataset, the biggest misalignments were word fillers.

Figure 5.14 illustrates the temporal alignment for a Swedish word *lingonsylt* ("lingonberry jam"), which is a part of the spontaneous dataset that MFA did not manage to annotate. The first tier represents the ground truth, while the dotted lines represent the actual timing of the word. Comparing Fig. 5.14 and 4.1 shows that the sound wave illustrated in Fig 5.14 is exceptionally affected by the noise (in

[2]Some of the mentioned words belong to more than one part of speech.

this case: wind), while the sound wave from the read data (see Fig 4.1) is cleaner and words can be differentiated only by looking at it.

To assess the performance of the aligners at the utterance level, I retrieved utterances with the biggest boundary displacement error. The retrieved utterances correlate with outliers at the word level. In most cases, the first or last word of an utterance is one of the words listed above. For instance, the biggest misalignments of Wav2Vec2 (VoxRex) for the semi-spontaneous and spontaneous datasets start with word fillers, such as *ehm* and *eh*.

For the read dataset, the utterances retrieved by Wav2Vec2 (VoxRex) finish with words such as *suveränitet* ("sovereignty") or *säkerhet* ("safety"), or start with a word *artikel* ("article"). The same pattern can be noticed in the results for WebMAUS and Aeneas.MFA has mostly issues with aligning utterances starting with *artikel*. Wav2Vec2 (XLSR) has more unique biggest misalignments, however, the utterances finishing with *suveränitet* ("sovereignty") is also one of them. The biggest misalignment of 3.99 seconds was created by MFA, while the biggest misalignment of Wav2Vec2 (VoxRex) was 0.333 seconds.

For the semi-spontaneous dataset, the biggest misalignment was generated by Aeneas (3.508 seconds), while for the spontaneous dataset the biggest misalignments were created by MFA, which did not produce any synchronization map for some data.

6. Results for the Norwegian alignment

In this Chapter, the results of the Norwegian temporal alignment are presented and discussed. The chapter is divided into three sections. In Section 6.1, the general outline of the results is described. In Section 6.2 evaluations and comparisons at the utterance level are presented. The qualitative comparison is presented in Section 6.3. The comparison is based on the evaluation metrics introduced in Section 4.3.

6.1. General outline of the results

WebMAUS WebMAUS succeeded in creating the output files for the entire dataset without missing words. WebMAUS was not used to create the gold standard for Norwegian at the utterance level, however, since it uses a cloned Norwegian acoustic model, the results still may be biased towards it. The descriptive statistics, as well as accuracy under tolerance thresholds for alignment generated with the Norwegian model, can be found in Figure A.2.

The Montreal Forced Aligner The Montreal Forced Aligner (MFA) managed to create alignments for the entire Norwegian read dataset. However, it failed to align one out of 27 chunks of the Norwegian semi-spontaneous data and one out of 17 chunks of the Norwegian spontaneous data. Consequently, it did not produce two output files for them. As expected, it created much bigger files with out of vocabulary words (OOV) than in the case of Swedish data. OOV dictionary consisted not only of Norwegian words with vowels, that does not exist in Swedish language, namely *æ*, *ø*, such as *økonomi* ("economy") or *væpnet* ("armed") but also words, which are written the same way in Swedish language, such as *ute* ("out"), *tematisk* ("thematic"), *område* ("area"), *språk* ("language") or *Sverige* ("Sweden").

Aeneas Aeneas managed to create an alignment for each of the chunks. As in the case of the Swedish data, it did not annotate any pauses between utterances with the chosen settings. Pauses were attached to a preceding utterance. If silence started the recording, the pause was attached to an utterance following it.

Wav2Vec2+CTC Segmentation Both Wav2Vec2 models managed to create alignments for the entire corpus.

6.2. Evaluations and comparisons

Figures 5.2, 5.4 and 5.6 show the distribution of manual/force aligned utterance boundary displacement, for each of the forced aligners. The x-axis represents the utterance boundary displacement (in seconds), while the y-axis represents the number of utterances. Tables 6.2, 6.7 and 6.10 provide mean, median and the standard deviation, as well as the accuracy measurements for utterance boundary displacement values at different tolerances. Figures 6.2, 6.4, 6.6 depict linear regression plots between the duration of utterances in human alignment and the duration approximated by the aligners and Tables 6.5, 6.8, 6.11 compare regression coefficients.

The null hypothesis states that there is no correlation between the variables. When all p-values are smaller than .01, the null hypothesis is rejected and the independent variable is statistically significant.

6.2.1. Semi-spontaneous speech data

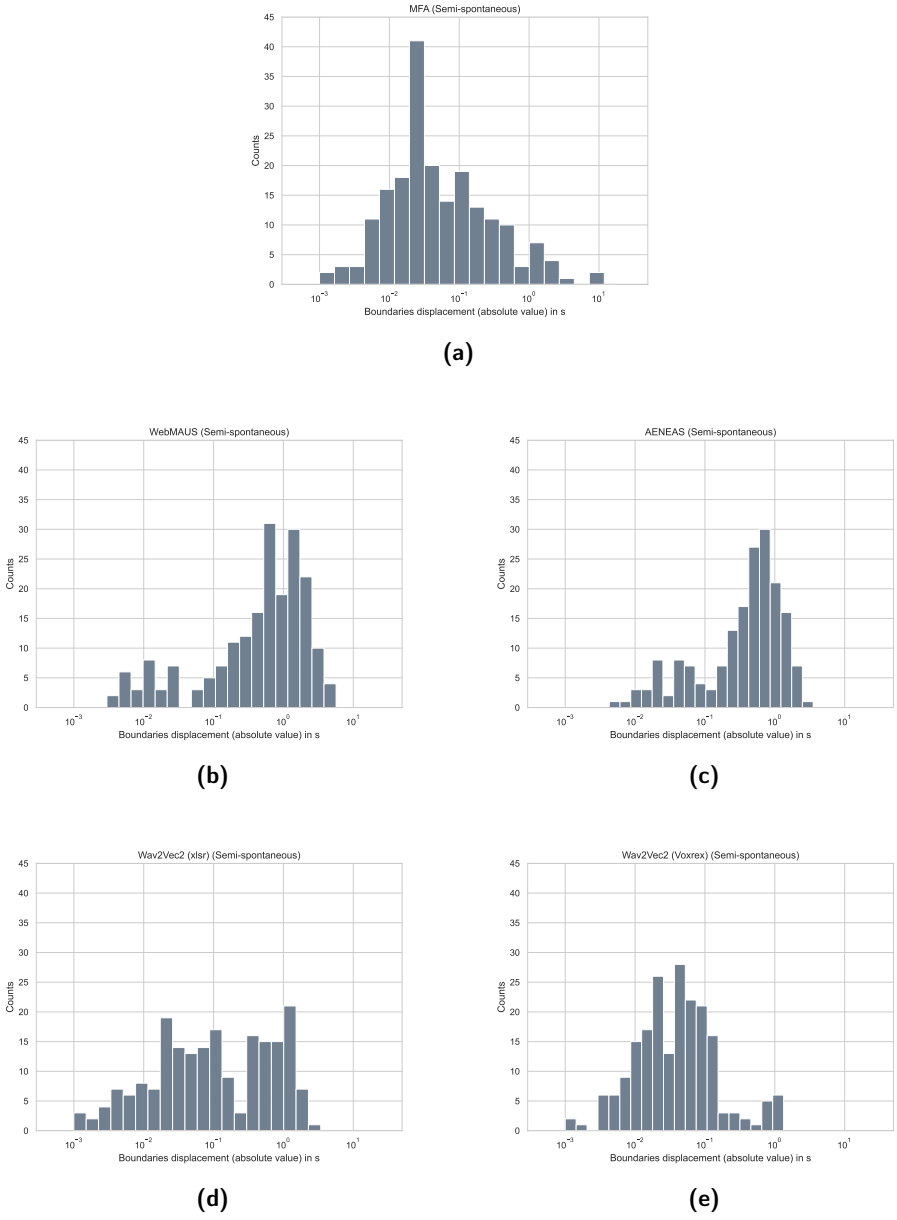


Figure 6.1.: Histograms of absolute utterance boundary displacement (on log scale) between force-aligned boundaries and gold-standard annotations for the Norwegian semi-spontaneous dataset at the utterance level. Due to the log scale, if a boundary displacement error equalled 0, the value was not visualized in the histogram.

Aligner	BD			Tolerance				
	Mean (s)	Median (s)	SD (s)	<0.01s	<0.025s	<0.05s	<0.1s	<0.5
MFA	0.367	**0.033**	1.634	**18.32%**	36.14%	57.43%	67.82%	89.11%
WebMAUS	0.950	0.621	1.003	7.43%	14.36%	15.84%	19.31%	41.09%
Aeneas	0.546	0.432	0.560	13.37%	19.31%	24.26%	29.70%	56.44%
Wav2Vec2 (XLSR)+CTC s.	0.373	0.094	0.533	14.85%	28.22%	38.61%	52.48%	71.78%
Wav2Vec2 (VoxRex)+CTC s.	**0.108**	0.038	**0.236**	16.34%	**40.59%**	**60.40%**	**81.19%**	**94.06%**

Table 6.2.: Descriptive statistics of boundary displacement and accuracies at different thresholds for boundary displacement between force-aligned boundaries and gold-standard annotations for the Norwegian semi-spontaneous dataset. The best results are bolded.

The semi-spontaneous speech data consist of two speeches written in *Bokmål* and one speech written in *Nynorsk*. First, the overall results are presented and then the results are divided by a written language, to see if a variant of written language has any influence on the alignments. Table 6.2 compares the descriptive statistics of boundary displacement for the semi-spontaneous dataset at the utterance level. As was mentioned in Section 6.1, MFA did not produce any alignment for one out of 17 chunks of the Norwegian semi-spontaneous data, which affected the results, in particular the standard deviation.

Wav2Vec2 (VoxRex) had – as the only one – a mean lower than 0.11 seconds, as well as one of the lowest mean (0.038 seconds) and standard deviation (0.236). MFA had the second-lowest median and mean and the highest standard deviation of 1.634. On the contrary, WebMAUS had median almost seventeen times higher than the median of Wav2Vec2 (VoxRex) and was the only aligner, in which a median exceeded 0.5 seconds. Wav2Vec2 (XLSR) achieved a mean comparable to the mean of MFA, despite the median being higher by 0.468 seconds. Overall, it can be noticed that Wav2Vec2 had the lowest descriptive statistics of boundary displacement error, while WebMAUS had the highest.

Table 6.2 also presents the accuracy measurements for boundary displacement values at different tolerances. MFA got the highest ratio of predictions with boundary displacement error lower than 0.01 seconds and the second-highest ratio for all other thresholds. Except for the <0.01 seconds threshold, Wav2Vec2 achieved the highest accuracy, for instance, 94.06% at the 0.5 seconds threshold. Both WebMAUS and Aeneas failed to exceed 60% of predictions with errors lower than 0.5 seconds. In the alignments produced by Wav2Vec2 (XLSR), almost 29% of utterances have a difference of at least 0.5 seconds, while about 85% have a difference of at least 0.05 seconds.

Table 6.4 presents the summary of descriptive statistics and one tolerance threshold for the Norwegian semi-spontaneous speech data broken down into two subtables based on the written language. Table 6.4 was created to see if one of the languages significantly influences the entire data.

Aligner	Bokmål				Nynorsk			
	Mean (s)	**Median (s)**	**SD (s)**	**<0.5s**	**Mean (s)**	**Median (s)**	**SD (s)**	**<0.5s**
MFA	0.180	0.032	0.454	92.50%	1.078	0.074	3.411	76.19%
WebMAUS	1.174	0.815	1.009	27.50%	**0.097**	**0.014**	**0.207**	**92.86%**
Aeneas	0.562	0.438	0.601	55.62%	0.485	0.373	0.360	59.52%
Wav2Vec2 (XLSR)+CTC s.	0.378	0.058	0.565	70.62%	0.354	0.161	0.388	76.19%
Wav2Vec2 (VoxRex)+CTC s.	**0.090**	**0.031**	**0.223**	**95.00%**	0.180	0.080	0.269	90.48%

Table 6.4.: Descriptive statistics of boundary displacement for the semi-spontaneous dataset in Bokmål and Nynorsk. Means, medians and standard deviations are over differences between aligned and gold-standard boundaries. <0.5s is a ratio of predictions with boundary displacement error lower than 0.5 seconds. The best results are bolded.

There are differences between the results for each of the aligners. The reason for the high standard deviation and mean of MFA on the Nynorsk data is the lack of alignment for one of the chunks, which resulted in several outliers with high boundary displacement errors. The median of MFA is much lower than the medians of WebMAUS, Wav2Vec2 (XLSR) and Aeneas.

However, the biggest difference is between the results of WebMAUS. In the data written in Bokmål, 72.50% of utterances have a boundary difference of at least 0.5 seconds, while in the alignments created for the utterances in Nynorsk only 7.14%. Moreover, there is also a big difference between descriptive statistics, the median for the Nynorsk data is 0.014, while for the Bokmål it exceeds 0.8 seconds of boundary displacement error. Thanks to these results, WebMAUS is the best aligner for the Nynorsk data, while being the worst aligner for the utterances written in Bokmål. The poor results on Bokmål data, influence the results on the whole semi-spontaneous dataset, making WebMAUS the worst-performing aligner at the 0.5 seconds threshold. When the same data is aligned by a model of WebMAUS intended for Norwegian, the distribution of displacement errors is smaller (see Table A.2).

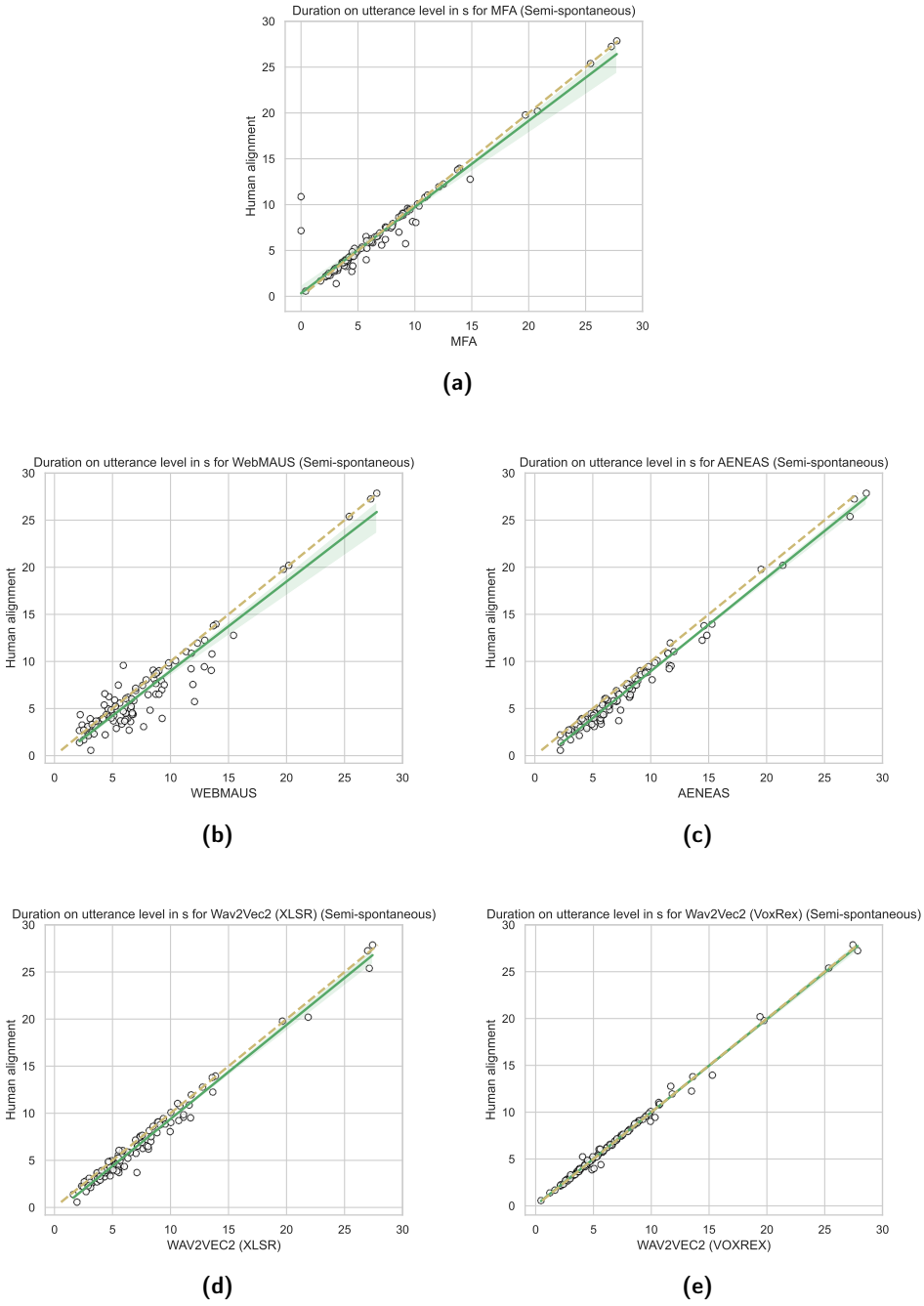


Figure 6.2.: Linear regression lines the between the duration of utterances in human alignment and the duration of utterances predicted by the aligners for the semi-spontaneous dataset. The dashed gold line represents a perfect fit. Confidence interval=95%

Aligner	**Slope**	**Intercept**	$\mathbf{R^2}$	***p*-value**
MFA	0.94	0.34	0.92	<0.01
WebMAUS	0.95	-0.50	0.90	<0.01
Aeneas	0.99	-0.97	0.98	<0.01
Wav2Vec2 (XLSR)+CTC s.	1.00	-0.59	0.98	<0.01
Wav2Vec2 (VoxRex)+CTC s.	0.99	0.06	0.99	<0.01

Table 6.5.: Regression statistics for the semi-spontaneous dataset.

Figure 6.2 illustrates linear regression plots between the duration of utterances in human alignment and the duration approximated by the aligners. The regression statistics is displayed in Table 6.5. The independent variable is statistically significant due to all p-values being smaller than .01.

Figure 6.2 and Table 6.5 illustrate that Wav2Vec2 (VoxRex) had the largest R^2 of 0.99, which means that the model expected that 99% of the observed variation could be explained by the independent variable (WAV2VEC2 (VOXREX)), whereas WebMAUS had the lowest R^2 of 0.90, meaning that 10% of the variation not explained by the model. The slope for Wav2Vec2 (XLSR) was 1, that is, for every step of increase (i.e., 1 second of word duration) on the x-axis (WAV2VEC2 (XLSR)), the linear regression model expected an increase of 1 second on the y-axis (THE HUMAN ALIGNMENT). The intercept of MFA and Wav2Vec2 (VoxRex) was positive, while for Aeneas, WebMAUS and Wav2Vec2 (XLSR) was negative.

6.2.2. Spontaneous speech data

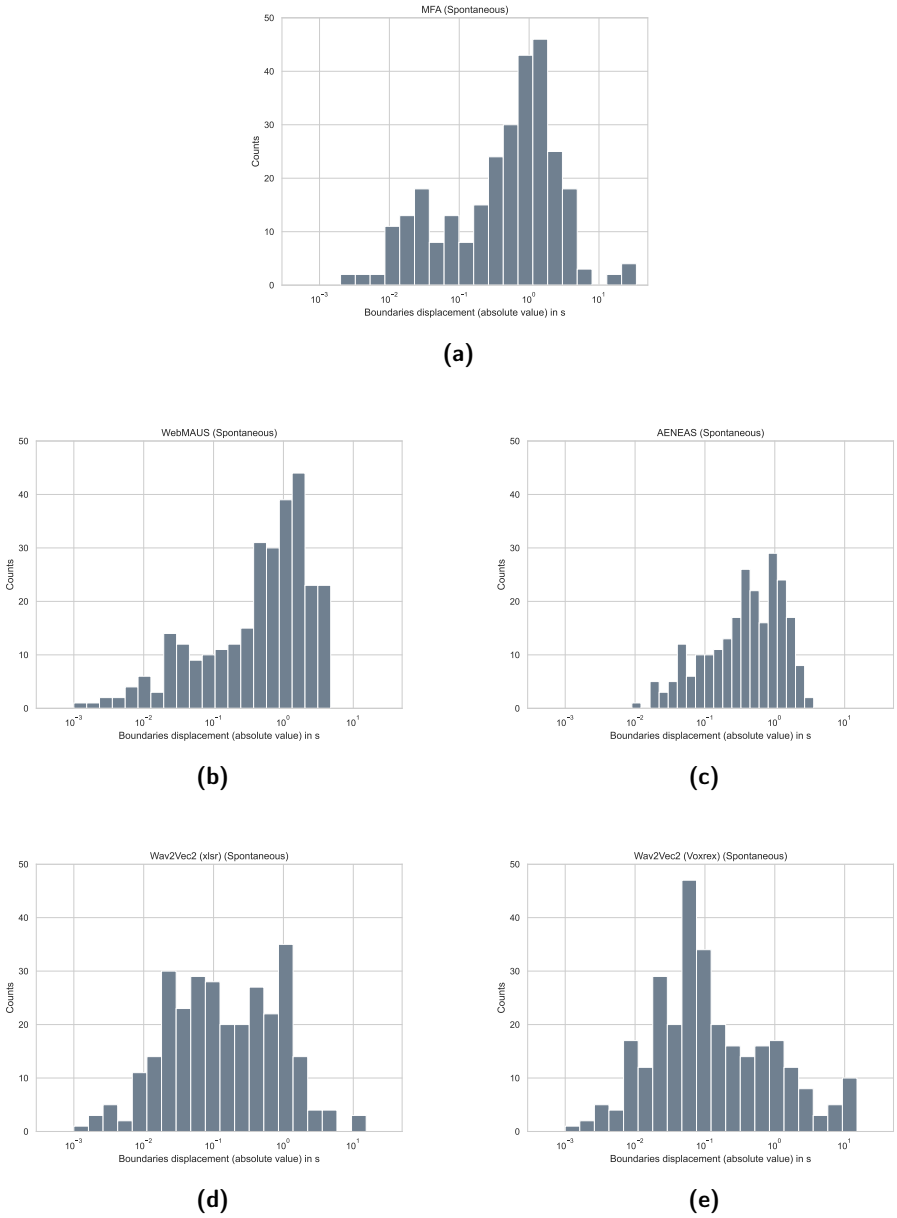


Figure 6.3.: Histograms of absolute utterance boundary displacement (on log scale) between force-aligned boundaries and gold-standard annotations for the Norwegian spontaneous dataset at the utterance level. Due to the log scale, if a boundary displacement error equalled 0, the value was not visualized in the histogram.

Aligner	BD			Tolerance				
	Mean (s)	Median (s)	SD (s)	<0.01s	<0.025s	<0.05s	<0.1s	<0.5
MFA	1.563	0.635	4.235	7.43%	13.51%	21.28%	26.69%	46.62%
WebMAUS	1.029	0.672	1.083	6.08%	10.47%	17.23%	22.64%	43.24%
Aeneas	**0.545**	0.300	**0.650**	**19.59%**	**22.30%**	25.68%	33.45%	63.18%
Wav2Vec2 (XLSR)+CTC s.	0.601	0.126	1.616	6.76%	19.93%	31.08%	45.95%	72.30%
Wav2Vec2 (VoxRex)+CTC s.	0.971	**0.079**	2.751	8.45%	**22.30%**	**33.78%**	**55.74%**	**76.01%**

Table 6.7.: Descriptive statistics of boundary displacement and accuracies at different tolerance thresholds for boundary displacement between force-aligned boundaries and gold-standard annotations for the Norwegian spontaneous dataset. The best results are bolded.

Table 6.7 compares the descriptive statistics of boundary displacement for the Norwegian spontaneous dataset at the utterance level. The spontaneous speech data has transcripts written in *Bokmål*. It can be noticed that the standard deviation is high for all aligners, especially for MFA. As was mentioned in Section 6.1, MFA did not produce any alignment for one out of 17 chunks of the Norwegian spontaneous data, which affected the results, in particular the standard deviation (SD=4.235). Wav2Vec2 (VoxRex) has the lowest median of 0.083 seconds and the second-highest standard deviation, suggesting that it created some large outliers, which is visible in Figure 6.3e. Aeneas scored the lowest mean and standard deviation and a lower median than WebMAUS and MFA. Despite the standard deviation of 1.616 and the mean higher than Aeneas, Wav2Vec2 (XLSR) got the second-lowest result for median (0.126 seconds). The distribution of boundary displacement errors is more right-skewed for MFA, WebMAUS and Aeneas than for the Wav2Vec2 models (see Figure 6.3).

Concerning the accuracy measurements, both two deep learning models surpassed the other forced aligners. Wav2Vec2 (VoxRex) got the best score at almost every tolerance threshold, except for 0.01 seconds, where it was exceeded by Aeneas. The good results of Aeneas can also be seen in Figure 6.3c, where a smaller number of values is depicted, meaning that a higher number of boundary displacement errors was 0. Surprisingly, Aeneas got better results than MFA and WebMAUS at every tolerance threshold. Both MFA and WebMAUS achieved similar results that differed by less than five percentage points.

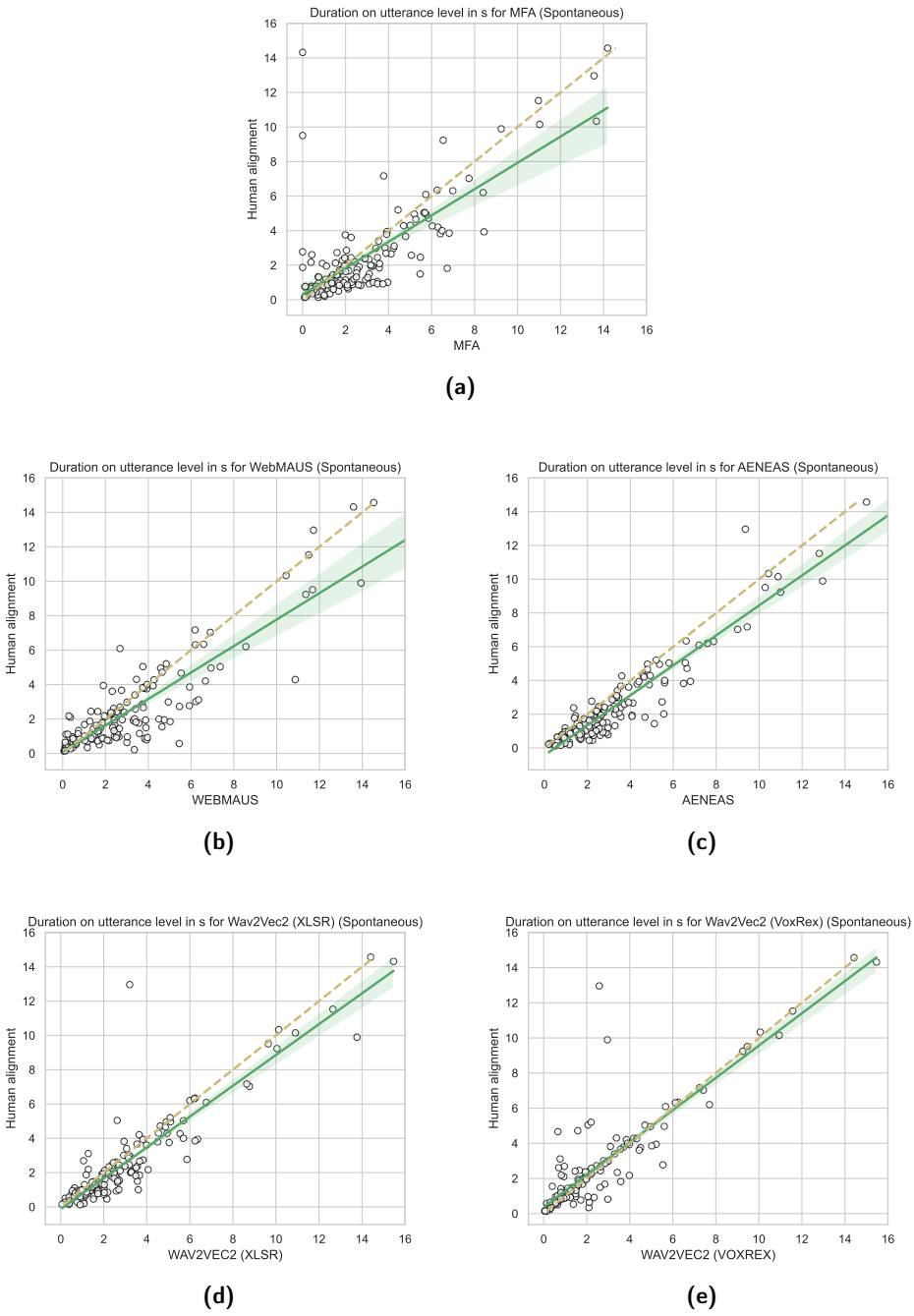


Figure 6.4.: Linear regression lines the between the duration of utterances in human alignment and the duration of utterances predicted by the aligners for the spontaneous dataset. The dashed gold line represents perfect fit. Confidence interval=95%

Aligner	**Slope**	**Intercept**	$\mathbf{R^2}$	***p*-value**
MFA	0.76	0.31	0.57	<0.01
WebMAUS	0.77	0.09	0.78	<0.01
Aeneas	0.89	-0.43	0.9	<0.01
Wav2Vec2 (XLSR)+CTC s.	0.90	-0.12	0.83	<0.01
Wav2Vec2 (VoxRex)+CTC s.	0.92	0.38	0.77	<0.01

Table 6.8.: Regression statistics for the spontaneous dataset.

Figure 6.4 shows linear regression plots between the duration of utterances in human alignment and the duration approximated by the aligners. The regression statistics is displayed in Table 6.8. The independent variable is statistically significant due to all p-values being smaller than .01.

All slopes are positive, which means that the two variables are positively related, that is, as the duration in the gold standard increases, so does the duration predicted by an aligner.

The model for MFA expected the lowest R^2 of 0.57, which was considerably lower than, for instance, Aeneas, where R^2 was 0.83. According to the model, 43% of the observed variation was not explained by the data produced by MFA.

Wav2Vec2 (VoxRex) had the highest slope of 0.88.

The intercept for MFA, WebMAUS and Wav2Vec2 (VoxRex) was positive, whereas for Wav2Vec2 (XLSR) and Aeneas the intercept was negative.

6.2.3. Read speech data

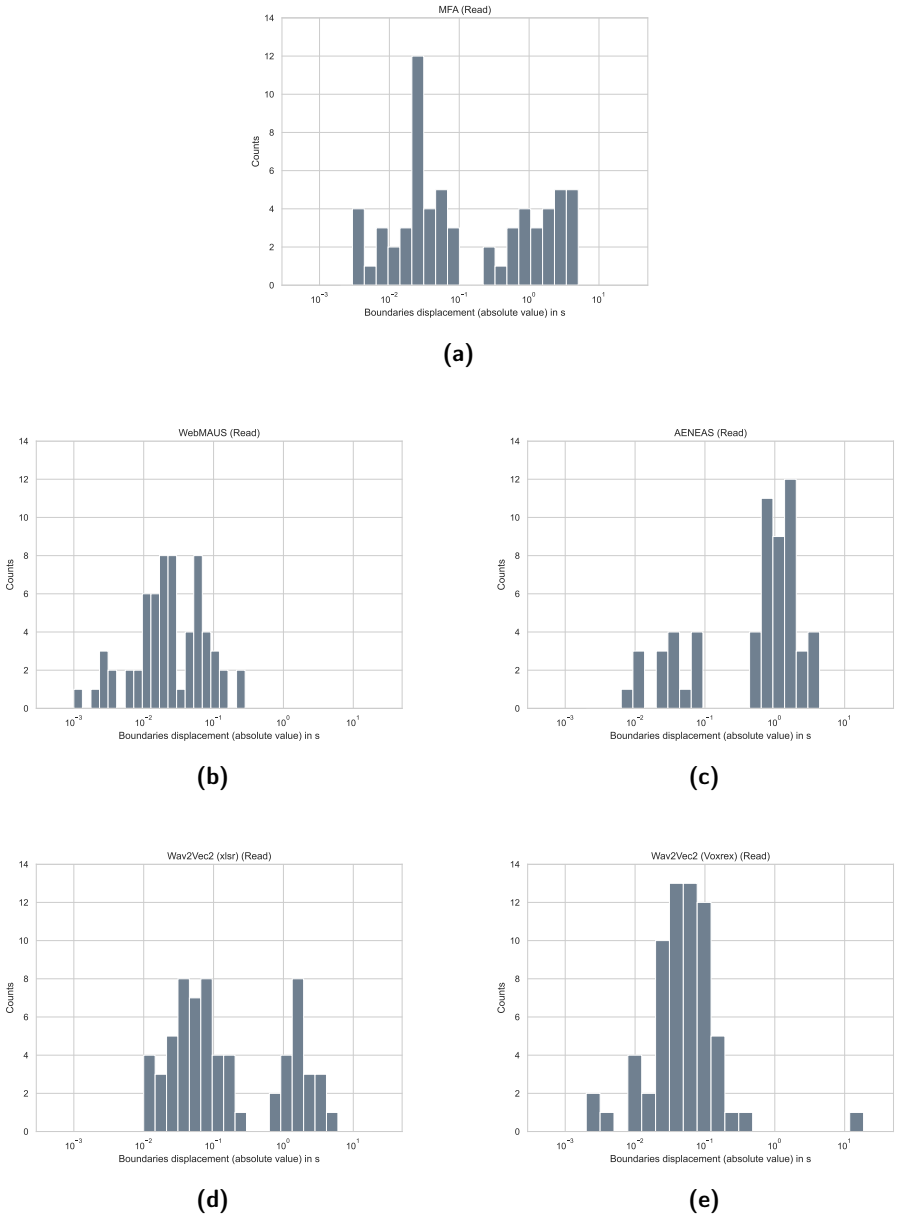


Figure 6.5.: Histograms of absolute utterance boundary displacement (on log scale) between force-aligned boundaries and gold-standard annotations for the Norwegian read dataset at the utterance level. Due to the log scale, if a boundary displacement error equalled 0, the value was not visualized in the histogram.

Aligner	BD			Tolerance				
	Mean (s)	Median (s)	SD (s)	<0.01s	<0.025s	<0.05s	<0.1s	<0.5
MFA	0.871	0.061	1.346	13.64%	31.82%	45.45%	57.58%	62.11%
WebMAUS	**0.042**	**0.022**	**0.054**	**25.76%**	**57.58%**	**71.21%**	**90.91%**	**100%**
Aeneas	0.948	0.753	0.963	12.12%	18.18%	27.27%	34.85%	34.85%
Wav2Vec2 (XLSR)+CTC s.	0.938	0.086	2.427	1.52%	13.64%	34.85%	53.03%	66.67%
Wav2Vec2 (VoxRex)+CTC s.	0.348	0.052	2.287	9.09%	19.70%	50.50%	78.79%	98.48%

Table 6.10.: Descriptive statistics of boundary displacement and accuracies at different tolerance thresholds for boundary displacement between force-aligned boundaries and gold-standard annotations for the Norwegian read dataset. The best result are bolded.

Table 6.10 compares the descriptive statistics of word boundary displacement for the Norwegian read dataset at the utterance level. The read speech data has a transcript written in *Nynorsk*. WebMAUS got the best results of all forced aligners, that is, a median of 0.022 seconds, a mean of 0.042 seconds and a standard deviation of 0.054. Wav2Vec2 (VoxRex) got the second-lowest median and the highest standard deviation. A low standard deviation indicates the data clustered around the mean. In the case of Wav2Vec2 (VoxRex), the highest standard deviation, along with one of the lowest medians, suggests a small number of outliers with significantly high boundary displacement error (see Fig. 6.5e), which influenced the mean and standard deviation of the aligner.

Table 6.10 also presents the accuracy measurements for utterance boundary displacement values at different tolerances. WebMAUS outperformed the other forced aligners at each tolerance threshold – 100% of alignments were at the maximum of 0.49 seconds apart from the ground truth. Wav2Vec2 (VoxRex) managed to align 98.48% of the data with a BD error under 0.5 seconds. MFA obtained higher accuracy at a 0.025 seconds threshold than Wav2Vec2 (VoxRex). Noticeably, Aeneas performs the worst – 65.15% of data has a BD error equal to or higher than 0.5 seconds. Wav2Vec2 (XLSR) created only 1.52% of alignments with a BD error under 0.01 seconds, while 33.33% of data has a BD error equal to or higher than 0.5 seconds. 37.89% of MFA's predictions has a boundary displacement equal to or higher than 0.5 seconds.

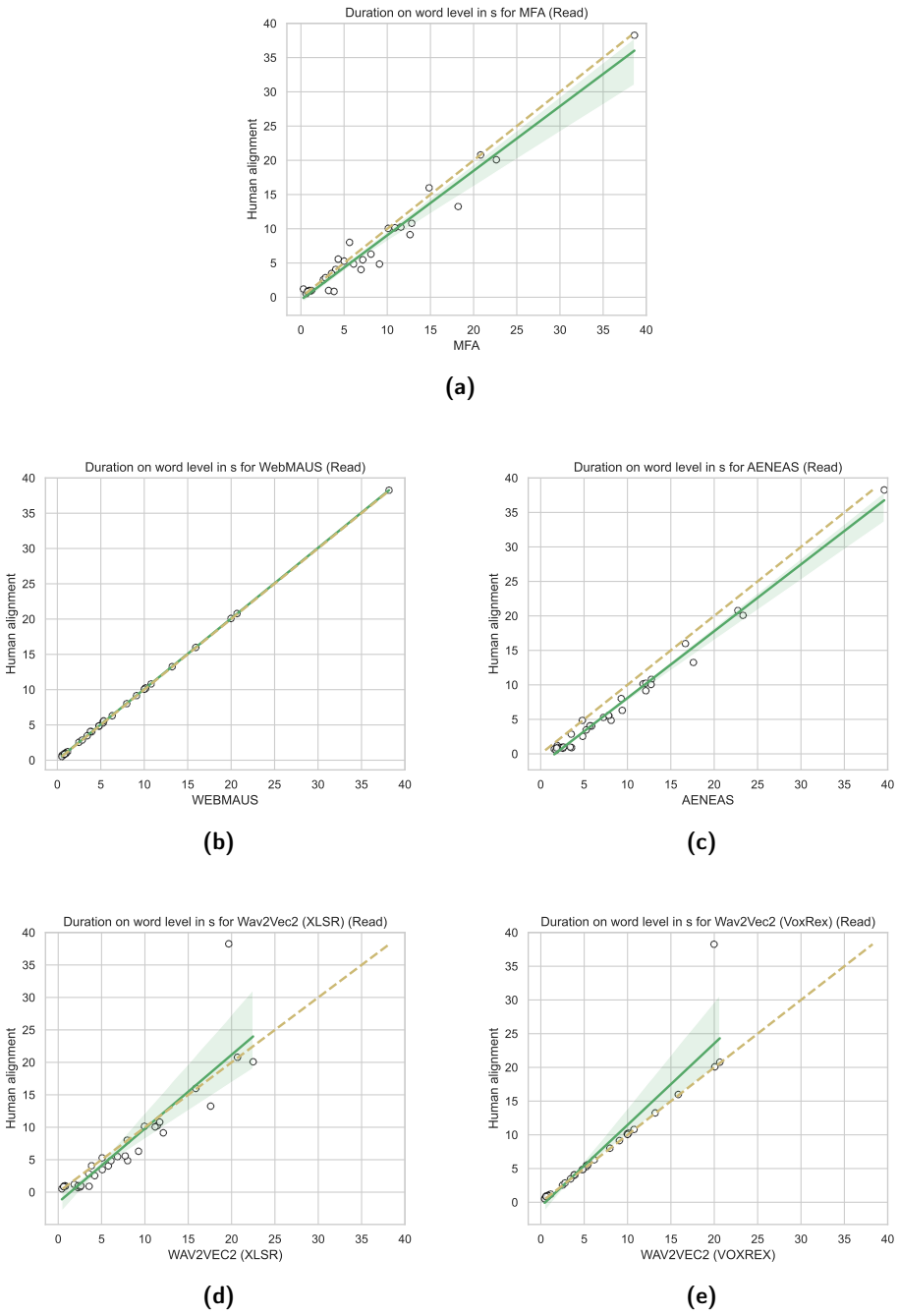


Figure 6.6.: Linear regression lines the between the duration of utterances in human alignment and the duration of utterances predicted by the aligners for the read dataset. The dashed gold line represents perfect fit. Confidence interval=95%

Aligner	**Slope**	**Intercept**	$\mathbf{R^2}$	***p*-value**
MFA	0.94	-0.38	0.96	<0.01
WebMAUS	1.00	0.06	1.00	<0.01
Aeneas	0.97	-1.58	0.99	<0.01
Wav2Vec2 (XLSR)+CTC s.	1.14	-1.58	0.80	<0.01
Wav2Vec2 (VoxRex)+CTC s.	1.21	-0.62	0.87	<0.01

Table 6.11.: Regression statistics for the Norwegian read dataset.

Figure 6.6 illustrates linear regression plots between the duration of utterances in human alignment and the duration approximated by the aligners. The regression statistics is displayed in Table 6.11. The independent variable is statistically significant due to all p-values being smaller than .01.

It can also be seen in Figure 6.6 and Table 6.11 that all slopes are positive, which means that the two variables are positively related, that is, as the duration in the gold standard increases, so do the duration predicted by an aligner. Figure 6.6 and Table 6.11 show that WebMAUS had the largest R^2 of 1, which means that the model expected that 100% of the observed variation could be explained by the independent variable (WEBMAUS). Meanwhile, Wav2Vec2 (XLSR) has the lowest R^2 0.8, meaning that 20% of the variation is not explained by the model. The slope for Wav2Vec2 (VoxRex) was 1.21, that is, for every step of increase on the x-axis (WAV2VEC2 (VOXREX)), the linear regression model expected an increase of 1.21 seconds on the y-axis (THE HUMAN ALIGNMENT). For all of the aligners except for WebMAUS, the intercept was negative. The same intercept was observed for Wav2Vec2 (XLSR) and Aeneas, that is, -1.58. Wav2Vec2 (XLSR) (Fig. 6.6d), along with Wav2Vec2 (VoxRex) (Fig. 6.6e) produced the biggest outliers.

6.2.4. Summary

Table 6.13 presents the descriptive statistics for the entire corpus.

Aligner	**BD**			**Tolerance**				
	Mean (s)	**Median (s)**	**SD (s)**	**<0.01s**	**<0.025s**	**<0.05s**	**<0.1s**	**<0.5**
MFA	1.054	0.170	3.297	12.06%	23.76%	37.06%	45.04%	63.95%
WebMAUS	0.885	0.521	1.035	8.87%	17.38%	23.05%	29.43%	49.11%
Aeneas	0.592	0.388	**0.676**	**16.49%**	20.74%	25.35%	32.27%	57.45%
Wav2Vec2 (XLSR)+CTC s.	**0.559**	0.104	1.476	9.04%	22.16%	34.22%	49.11%	71.45%
Wav2Vec2 (VoxRex)+CTC s.	0.589	**0.058**	2.181	11.35%	**28.55%**	**45.21%**	**67.55%**	**85.11%**

Table 6.13.: Descriptive statistics of boundary displacement and accuracies at different tolerance thresholds for boundary displacement between force-aligned boundaries and gold-standard annotations for the whole Norwegian dataset. The best results are bolded.

Transformer models outperformed the other forced aligners concerning median and mean boundary displacement. Wav2Vec2 (VoxRex) had the best median of 0.058 seconds, while Wav2Vec2 (XLSR) had the best mean of 0.559 seconds. What can be noticed in Table 6.13, Wav2Vec2 (VoxRex) got one of the highest standard deviations, while Aeneas had the lowest standard deviation of 0.676. WebMAUS had the worst median of 0.521 seconds.

Aeneas achieved the best accuracy at the lowest threshold of 0.01 seconds. From the 0.025 seconds threshold, Wav2Vec2 (VoxRex) had the best accuracy, scoring 67.55% at 0.1 seconds and 85.11% at 0.5 seconds. Moreover, MFA outperformed Wav2Vec2 (XLSR) up to 0.05 seconds threshold. WebMAUS got the lowest accuracy at all tolerance thresholds.

6.3. Qualitative comparison

As in the case of the Swedish dataset, in order to compare the performance of the aligners at the utterance level, the utterances with the highest boundary displacement errors were extracted. The extracted utterances were compared among the forced aligners, to see if the extracted outliers were similar for a higher number of aligners.

For the spontaneous and semi-spontaneous datasets, the highest outliers were created by MFA due to the lack of alignments. On the contrary, both Wav2Vec2

models created the highest outliers for the read dataset, which generated a misalignment of around 19 seconds on the same utterance. The second highest misalignment of Wav2Vec2 (VoxRex) is 63 times smaller, lasting 0.3 seconds. One of the biggest Wav2Vec2 (VoxRex) misalignments for the read dataset were utterances starting with a word *Artikkel* ("Article"). Generally, the extracted utterances vary from aligner to aligner and no clear pattern is visible.

7. Discussion

In this thesis work, I have evaluated the performance of the state-of-the-art forced aligners for Swedish speech. I have also investigated whether a Swedish-dependent forced alignment (Language Dependent Force Alignment, LDFA) system could be used for aligning Norwegian (Cross-language forced alignment, CLFA) and how the performance was affected given it was used on a language, on which it was not trained. Furthermore, it was investigated whether the spontaneousness of the speech affected the performance. In this Chapter, I will provide an analysis of the results and discuss them in relation to the literature. Moreover, I will point to some limitations of the work. Below the research questions are addressed.

1. What is the accuracy of the aligners relative to the alignments of a manually created gold standard?
First, I consider the results for the combined Swedish dataset. The Swedish alignment results at the utterance level are comparable to the results achieved by the authors of the CTC segmentation algorithm (Kürzinger et al., 2020), who evaluated the alignments produced by MAUS, Aeneas, Gentle (DNN based on Kaldi; Ochshorn and Hawkins (2017)) and three neural network models (one BLSTM and two Transformers) with CTC segmentation at the utterance level. Due to the differences in size between test sets, the scores cannot be compared one-to-one. However, it can be noticed that the scores achieved on the Swedish dataset are in line with the results from Kürzinger et al. (2020). The highest ratio for boundary displacement errors smaller than 0.5 seconds was generated by CTC segmentation algorithm (Kürzinger et al. (2020): 85-90%, thesis results: 81-97.7%), then Gentle and MAUS (Kürzinger et al. (2020): 82% & 74.1%, thesis results: 80.50% (MFA) & 89.99%), while Aeneas scored the lowest accuracy (Kürzinger et al. (2020): 64.7%, thesis results: 69.50%).

Even though in accordance with my results WebMAUS achieved better accuracy than MFA (based on Kaldi), the gold standard is biased towards it, which is explained in detail later in the chapter.

Concerning the results for Swedish on word level, they can be compared to Leinonen et al. (2021), who created a Finnish forced aligner based on Kaldi, as well as to McAuliffe et al. (2017), the authors of MFA (also based on Kaldi). McAuliffe et al. (2017) and Leinonen et al. (2021) calculated the accuracies at the word level within the thresholds of 0.01, 0.025, 0.05 and 0.1 seconds. The results of MFA with the Swedish acoustic model at 0.01 seconds threshold (23.35%) are comparable with the ones achieved by the Finnish forced aligner (Leinonen et al., 2021, 21 %) and around 10% lower than the results of MFA. However, at the higher tolerance thresholds, the Finnish model, as well as English MFA, exceeds all Swedish forced alignment systems by around 10-30%. The different amounts of boundary errors might be caused by the method of annotation, training data, audio quality and speaking style (Leinonen et al., 2021), which is explored later in this Chapter.

Overall, the accuracy, that is, the ratio of prediction at a maximum of 0.1 or 0.5 seconds apart from the gold standard, was between 33.72%-69.79% and 69.5%-97.65% at the utterance level and at the word level 68.90%-88.87% and 90.8%-99.35%, respectively. When it comes to the descriptive statistics, the mean was in the range of 0.054-0.978 seconds (word-level) and 0.103-1.344 seconds (utterance-level), while the

median was between 0.025-0.050 (word-level) and 0.061-0.248 seconds (utterance-level). The high difference between the means indicates that some of the aligners produced more spread-out segmentation.

2. How is the performance of a forced alignment system affected when it is used to align a language, on which it was not trained?
The results for Norwegian data could be discussed in relation to Leinonen et al. (2021), who examined cross-language forced alignment at the word level for Estonian and Northern Sami by using a language specific forced aligner for Finnish. However, once again, the results cannot be compared one-to-one, due to the Norwegian data being aligned at the utterance level. As can be noticed in the results on the Swedish dataset, the results at the word level are better than the results at the utterance level. The Estonian results are comparable to Finnish alignments, except for smaller ranges that are actually better than Finnish alignments (e.g., at 0.1 threshold – Finnish: 21%, Estonian: 33%) (Leinonen et al., 2021). On the contrary, the results for Northern Sami are worse than the results for Estonian. The researchers suggest that better results of Estonian can be explained by the similarities in how the speech recognizer mainly aligns the speech (Leinonen et al., 2021).

The scores for the whole Norwegian corpus were significantly worse than the ones achieved on the Swedish data. Both Wav2Vec2 models achieved the best results at higher thresholds – at 0.1 and 0.5 tolerance threshold, Wav2Vec2 (VoxRex) got 67.55% and 85.11%, while Wav2Vec2 (XLSR) achieved 49.11% and 74.45%, respectively. On average, there was a difference of 10-20% between the scores on the Swedish data and Norwegian data. What is interesting, WebMAUS got the lowest accuracy scores at all thresholds, even though it could be assumed that WebMAUS should be biased towards the Norwegian data. As was mentioned before, the Swedish model of WebMAUS uses a cloned Norwegian acoustic model and no phonetically labelled Swedish data was used for training, which according to the authors, is far from optimal (Schiel, 2022). When analyzing the results, it can be noticed that WebMAUS created alignments of good quality for the recordings transcribed in Nynorsk, while the recordings transcribed in Bokmål were problematic. The huge differences between the accuracies occurred within the semi-spontaneous speech data, which excludes speaking style or background noise (specific to the type of data) to be the reason for the differences. Even though the results are consistent with the mutual intelligibility between Scandinavian countries, that is, Nynorsk is easier to understand for Swedes than Bokmål (Delsing, 2005), I assume that the reason for the differences between these results is the training data, or the usage of a Swedish pronunciation lexicon, because when the same data is tested with the Norwegian acoustic model, the distribution of displacement errors is much smaller. Moreover, it can also be hypothesized that these results might be related not only to the written language but also to the dialects of the speakers.

Since Scandinavian languages belong to a dialect continuum (Gooskens, 2020) and Wav2Vec2 (VoxRex) was trained on different dialects of Swedish (Malmsten et al., 2022), it can thus be suggested that the training data of Wav2Vec2 (VoxRex) is the reason for its accuracy on Norwegian data.

As in the case of the Swedish dataset, the difference in boundary errors might be caused, not only by using a different target language but also by a method of annotation, audio quality and speaking style (Leinonen et al., 2021), which is explored later in this chapter.

Overall, the accuracy, that is, the ratio of prediction at a maximum of 0.1 or 0.5 seconds apart from the gold standard, was between 29.43%-67.55% and 49.11%-85.11% at the utterance level. When it comes to the descriptive statistics, the mean

difference was in the range of 0.589-1.054 seconds, while the median disagreement was between 0.058-0.521 seconds. Higher mean and median scores in the Norwegian dataset suggest that the distribution of boundary displacement errors is more right-skewed than the distribution of Swedish boundary displacement errors.

3. Are there any differences in performance between aligners for 1-2?

It's worth noting that MFA and WebMAUS almost always outperformed the Wav2Vec2 models, as well as Aeneas, in low thresholds, such as 0.05 and 0.1 seconds. These results suggest that traditional ASR algorithms (HTK and Kaldi) are able to perform better at small tolerance levels. A possible explanation for this might be that these algorithms were developed for studies on phonemes, hence they offer the phoneme-level of granularity.

Prior studies have noted that WebMAUS tends to yield worse alignments than MFA (Gonzalez et al., 2020; Mahr et al., 2021; Meer, 2020). Gonzalez et al. (2020) and Mahr et al. (2021) suggest that the high performance of MFA is caused by more advanced ASR techniques implemented in Kaldi. Contrary to HTK, a speech recognition toolkit used in WebMAUS, MFA generates alignments at the triphone level, taking surrounding phonetic context into account (McAuliffe et al., 2017), while WebMAUS generates the alignment on the monophone level. The results of WebMAUS on the Norwegian dataset correlate with the findings presented by Gonzalez et al. (2020), where MAUS presented the poorest performance despite being trained on the aligned variety of English. In the case of the Norwegian dataset, even though WebMAUS uses a cloned Norwegian acoustic model trained on recordings of Norwegian speakers from 24 different dialect areas (Schiel, 2022), it had poorer performance on the part of Norwegian data, which affected the overall accuracy. WebMAUS with a cloned Norwegian model was outperformed by the Wav2Vec2 model trained solely on Swedish (Malmsten et al., 2022). For instance, on the whole, for the Norwegian dataset, Wav2Vec2 (VoxRex) had an accuracy of 85.11% at the 0.5 tolerance threshold, while WebMAUS had an accuracy of 49.11%. In addition, even though Gonzalez et al. (2020) tried to improve the performance of WebMAUS by using an acoustic model based on the spontaneous training data, no significant improvements were noted. These findings suggest that the algorithm may be responsible for its limitations.

Wav2Vec2 (VoxRex) consistently achieved one of the best results not only on the whole Swedish and Norwegian datasets but also on the divided data. Wav2Vec2 (XLSR), on the other hand, while achieving the second-best result for the whole datasets, was outperformed by, for instance, MFA. These differences in performance between Wav2Vec2 (XLSR) and Wav2Vec2 (VoxRex) strongly suggest that the choice of a fine-tuned model for a deep learning approach has a big influence on the performance of the aligner. The temporal alignment scores for Wav2Vec2 (VoxRex) and Wav2Vec2 (XLSR) were parallel to their Word Error Rate (WER) scores, reported by KBLab, that is 8.49 and 14.30, respectively (Malmsten et al., 2022).

Overall, it is apparent that the performance of Aeneas was inferior in comparison to other tested forced aligners. The algorithmic approach behind Aeneas is not as effective as other aligners for word-level. Even though the aligner offers parameters to improve word-level synchronization, according to Pettarin (2017) the limitations of Aeneas are built into the TTS+DTW algorithm. To produce a word-level granularity, Aeneas synthesizes single words and as a result, it does not consider coarticulation. Moreover, with the selected settings, pauses between words were not annotated, which primarily affects the accuracy and the estimated duration of the segments. Since pauses are added to the segments, estimated durations are longer in comparison

to the gold standard and other aligners, which is especially visible in intercepts of the regression models. It should, however, be noted that Aeneas most of the time produced the correct start points of the words. The big misalignments occurred mostly due to the lack of annotated pauses, which affects the end time of the words.

It can be observed that the quality of automatic alignment with traditional HMM and DTW methods is affected by the deviations from the audio signal since the traditional forced aligners – as the name suggests – *force* align the transcript with the audio recording. If a transcribed segment does not exist in the audio, the aligner will still try to fit the segment into timings. According to Strunk et al. (2014), WebMAUS depends more on the quality of the transcription than the quality of an acoustic model. According to Beňovič (2020), MFA does not create any alignment in case of a bigger amount of background noise or a number of errors in the transcript. Since MFA did not align some of the data, what undeniably influenced the results (especially standard deviation and mean), this study supports the findings from Beňovič (2020). Even though the performance of the traditional forced aligners decreases dramatically in case of the deviation between the audio and transcript, CTC segmentation retains the alignment abilities (Kürzinger et al., 2020), which is visible in the overall performance of Wav2Vec2 (XSLR).

4. How does the degree of spontaneousness in the speech affect the performance of a forced alignment system?

With respect to the fourth research question, it was found that the type of speech had an effect on the performance of the forced aligner. It was shown that the distribution of misalignments differs between types of speech. The spontaneous dataset was the most challenging for the forced aligners. Not only because it has a higher degree of coarticulation, or it is not as faithful to the written form as is read speech, but also due to the background noise, such as laughter or wind. Therefore, it is difficult to estimate, whether the poorer performance was caused by noise or by the type of speech data. For Swedish data, MFA failed to annotate a part of the spontaneous speech data set, while for Norwegian data it did not annotate not only a part of spontaneous speech but also a part of the semi-spontaneous dataset. As was mentioned before, MFA does not create any alignments in case of a great amount of background noise, not because of more conversational speech type.

Temporal alignments generated for the spontaneous dataset by the CTC segmentation algorithm had a higher correlation with the gold standard compared to the alignments produced by all other tested alignment algorithms. The specification of CTC segmentation strongly suggests the Wav2Vec2 models are more robust in case of unknown parts of the audio (e.g., a word that had not been annotated) by not annotating it, while the other forced aligners would try to force align one of the utterances from the transcript (Kürzinger et al., 2020). The differences between the standard deviations of the traditional forced alignment methods (MFA: 6.535, WebMAUS: 0.317, Aeneas: 0.407) and the deep learning models (Wav2Vec2 (VoxRex): 0.109 and Wav2Vec2 (XLSR): 0.155) illustrate that data points from the traditional methods are not consistent, but more spread out over a large range of values. As MacKenzie and Turton (2020) highlight, massively reduced speech, along with laughter, can be the source of errors for the forced aligners. Moreover, by analyzing common outliers, it can be noticed that words that end with aspirated consonants, such as [k], [t] and [p], are another source of errors, as well as retroflex consonants across words (e.g., *för sen* ("because then")). Plosive sounds have been shown to be among the most difficult classes for forced aligners (Mahr et al., 2021).

For the read data set, the median of displacement was the lowest in comparison to other data sets. For the Swedish data set, all aligners managed to achieve accuracy

at the 0.5 threshold bigger than 96% at the word level. What is interesting, the distribution of displacement errors at the utterance level is bigger than in the semi-spontaneous or spontaneous data.

A strong relationship between the training data and the data to be aligned has been reported in the literature. Cassidy and Schmidt (2017) suggested that a forced aligner obtains higher accuracy if the aligned data has a similar type of speech as the training data. This differs from the findings of Fromont and Watson (2016) who claimed that a forced aligner trained on spontaneous data produces a better alignment for both read and spontaneous speech. The differences in the performance of Wav2Vec2 models reflect those of Kürzinger et al. (2020) who also found that the CTC segmentation algorithm with a Transformer model trained on a dataset that contains more reverberation and noise from an audience yields better scores than with a Transformer model trained on a dataset consisting of sentences recorded in isolation. Wav2Vec2 (XLSR) was trained on read and more controlled datasets, such as Common Voice (Ardila et al., 2019), that consist of distant sentences recorded in isolation. In contrast to the XLSR model, Wav2Vec2 (VoxRex) was trained on a mix of Swedish local public radio, podcasts and audiobooks. According to Malmsten et al. (2022), the dataset used for the training was a more representative and democratic dataset, due to including different types of speech, dialects and a bigger number of speakers. Moreover, even though MFA was trained on read texts from newspaper articles, the training data also included markers for spontaneous effects like stuttering, false starts and even non-verbal effects such as breathing, laughing (Schultz et al., 2013). The fact that MFA, in most cases, outperforms Wav2Vec2 (XLSR) indicates that the type of training data has more influence on the performance than more modern architecture.

Memory constraints

Despite the fact that according to the results, Wav2Vec2 (VoxRex) might be assumed to be one of the best choices of the Swedish forced aligners, the final choice of an algorithm should be dependent on its intended use.

The biggest limitation of Wav2Vec2 (VoxRex), as well as Wav2Vec2 (XLSR), is the inference complexity of Transformer-based architectures. Since inference complexity increases on long audio files quadratically, Wav2Vec2 does not accept audio files longer than 40 seconds as the input and fails at longer audio files because of excessive memory consumption (Baevski et al., 2020). Hence, the audio needs to be chunked into several parts. In the case of a large data set (over about NNN tokens), segmentation can be problematic, as well as time- and resource-consuming. Moreover, the CTC segmentation algorithm is not parallelizable for batch processing, hence a CPU with a good single-thread performance needs to be used. Enough RAM is needed to align multiple files in parallel (Kürzinger et al., 2020). Since the memory consumption of the algorithm strongly depends on the model architecture and the used toolkit, the segmentation of the audio may be avoided or reduced by using a different CTC-based end-to-end network. However, a pre-trained and fine-tuned Swedish model other than Wav2Vec2, could not be found.

In comparison to Wav2Vec2, WebMAUS allows longer sound files with up to 3000 words. However, if the input files exceed this boundary, the data needs to be pre-segmented into chunks, since MAUS will also not create any alignments, due to the processing time increasing quadratically with the number of words.

If the adequate amount of RAM is assured, Aeneas can take as the input a single audio file with a duration of 10 hours. Moreover, the alignments are produced faster by Aeneas than by the other forced aligners.

Limitations

The alignments of WebMAUS cannot be easily compared to the alignments of the other forced aligners, since the Swedish gold standard is biased towards it. It can be noticed that the results from WebMAUS were mostly biased at lower thresholds, such as 0.01 or 0.25 seconds and were outrun at bigger thresholds. If a boundary in the gold standard was not manually adjusted, then the boundary displacement was equal to 0 seconds. Consequently, the number of 0 seconds displacements was overestimated, especially when compared to the numbers of displacement errors between 0.001-0.010 and 0.010-0.25 seconds. To exemplify this, at the word level, the WebMAUS ratio of segments with boundary displacement of 0 seconds was 16%, while MFA's and Wav2Vec2 (XLSR)'s ratios were 4% and 1%, respectively. At the same time, WebMAUS had 14% of errors between 0.001 and 0.010 seconds of displacement, while MFA and Wav2Vec2 (XLSR) had only 1% less displacement at the same range. Moreover, the number of 0.010-0.025 seconds displacements is even higher for MFA and Wav2Vec2 (XLSR). This behaviour is parallel to the results in the study by Mahr et al. (2021), who created the gold-standard manual alignments by correcting alignments created by the Prosodylab-Aligner (Gorman et al., 2011). As Mahr et al. (2021) reported, the smallest time differences were biased toward the Prosodylab-Aligner, however, at the higher time differences, Prosodylab was outperformed.

Moreover, a Swedish model in WebMAUS is not a language-dependent model, since the Swedish model of WebMAUS uses a cloned Norwegian acoustic model and no phonetically labelled Swedish data was used for training.

In addition, it can be noticed that the results for MFA and WebMAUS, the forced aligners that both use HMM algorithms, are similar. Hence it is difficult to estimate whether WebMAUS should be taken into consideration or not. It can be safely assumed that without the bias, the scores for WebMAUS would be similar to or lower than the ones achieved by MFA, as it is also known from the literature that in general MFA yields better scores than MAUS or other forced aligners based on HTK (Gonzalez et al., 2020; Mahr et al., 2021).

In order to overcome this limitation, most researchers: a) create a segmentation manually from scratch, or b) use a segmentation, pre-aligned by a forced aligner and adjusted by multiple transcribers, often reporting the percentage of human-human agreement (Gonzalez et al., 2020; MacKenzie and Turton, 2020). Due to the limited amount of time and resources, neither of these options was feasible for this thesis.

Summary

To summarise, the results for both Swedish and Norwegian are in agreement with those obtained by Kürzinger et al. (2020). When considering the entire dataset jointly, the Transformer models with CTC segmentation create alignments that usually correspond closer to the gold standard, in comparison to the models based on DTW and HMM. Even though the quality of alignments tends to differ between different datasets, the alignments produced by the CTC algorithm are more robust and consistent across the domains. However, it is worth noting that for lower thresholds, the traditional ASR forced aligners such as MFA, outperforms CTC segmentation. Moreover, in the results for the specific datasets, it can be noticed that in the semi-spontaneous dataset, MFA outperforms Wav2Vec2 (XLSR) at all thresholds. This is most likely due to MFA being used for producing alignments on the phoneme level in sociolinguistics studies.

The distribution of displacement errors for both Wav2Vec2 models suggests that the quality of training data has more influence on the quality of alignments than the algorithm itself.

Despite the fact that WebMAUS results cannot be easily compared to the other aligners due to the bias, it should be highlighted that the architecture of the Swedish model in MAUS itself is a confirmation that other Scandinavian languages can be used in cross-language forced alignment for Scandinavian languages and achieve satisfactory results.

8. Conclusions

This study was conducted to evaluate the off-the-shelf forced alignment systems available in the Swedish language for temporal alignment of speech and transcription. It was examined how effectively the problem of temporal alignment can be resolved given audio and text and how the different types of speech in the data affect the performance of a forced alignment system. Moreover, it was examined whether a forced aligner trained on Swedish can be used to create valuable alignments for another close-related language, such as Norwegian.

Three approaches for forced alignment available for the Swedish language were employed and evaluated: the Montreal Forced Aligner (MFA) and WebMAUS based on Hidden Markov Models, Aeneas based on a TTS engine and Dynamic Time Warping algorithm and two pre-trained and fine-tuned Wav2Vec2 models (VoxRex and XLSR) with forced alignment based on the CTC segmentation.

It was shown that for Swedish, MFA and WebMAUS tend to outperform the Wav2Vec2 models and Aeneas in low thresholds. Wav2Vec2 (VoxRex) yielded the best results under 0.5 seconds tolerance. More than 97% of the segments had a boundary displacement smaller than 0.5 seconds, while more than 77% smaller than 0.1 seconds.

Concerning the aligners' performance on various types of speech data, it is visible that the off-the-shelf forced aligners perform better on the read or semi-spontaneous speech without background noise and good recording quality. The performance of the aligners significantly dropped on the spontaneous dataset with background noise, however, the aligners based on CTC segmentation retained their performance.

One of my original motivations for carrying out this study was to test whether a forced alignment algorithm can take advantage of similarities between two Scandinavian languages and produce a good quality alignment. Findings from the thesis have demonstrated promising results for cross-language forced alignment using Swedish models to align related languages, such as Norwegian. As in the case of the Norwegian dataset, Wav2Vec2 (VoxRex) + CTC segmentation outperformed the other forced aligners with more than 85% of boundary predictions being at a maximum of 0.49 seconds apart from the ground truth.

It can be concluded that the performance of a forced aligner is influenced by a few aspects such as the data it was trained on, its algorithm and the type of the aligned data. The huge differences in the performance of Wav2Vec2 models may suggest that the training data has a bigger influence on the accuracy of a forced aligner than its algorithm. Moreover, I also highlight that the accuracy of the model should not be the main factor influencing the choice of an aligner, rather than the amount of computational power, type of the data, the number of errors in the transcripts and type of the expected alignment.

To summarise, the contributions of this thesis are the following:

1. A small speech corpus in Norwegian and Swedish with temporally aligned audio and text was created for evaluating forced aligners.

2. The extensive evaluation of the off-the-shelf forced alignment algorithms available for the Swedish language. The aligners were evaluated on datasets with

different properties. It has been shown which aligner works the best in which setting.

3. It has been shown that Swedish and Norwegian can be used in cross-language forced alignment. An aligner trained on Swedish can be used for aligning data in Norwegian and produce valuable alignments with 90% of utterances having boundary displacement errors of less than 0.5 seconds.

The thesis provides a lot of space for future work. A natural progression of this work is to analyse whether an off-the-shelf forced aligner trained on Swedish is able to create valuable alignments for Danish. Moreover, it would be interesting to see if aligners trained on Norwegian data are able to create better alignments for Swedish and Danish, than an aligner trained on Swedish for Norwegian and Danish. With this direction in the research, we would be able to see if forced aligners can imitate the reported level of mutual intelligibility between native speakers of Scandinavian languages. Further research might also explore the cross-lingual forced alignment for more distant-related languages and how far away we can go from the source language. For instance, it could be examined whether a Swedish forced aligner might be used to generate valuable alignments for Insular Scandinavian languages: Icelandic and Faroese, or other Germanic languages, such as English or German.

Bibliography

Amin, Talal Bin and Iftekhar Mahmood (2008). "Speech Recognition using Dynamic Time Warping". In: *2008 2nd International Conference on Advances in Space Technologies*, pp. 74–79. DOI: 10.1109/ICAST.2008.4747690.

Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber (2019). "Common voice: A massively-multilingual speech corpus". *arXiv preprint arXiv:1912.06670*.

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

Benesty, J, J Chen, and Y Huang (2008). *Automatic speech recognition: A deep learning approach*.

Beňovič, Marek (2020). "Forced Alignment via Neural Networks".

Birkenes, M. B. (2020). *NST Swedish Dictation (22 kHz)*. URL: https://www.nb.no/sprakbanken/en/%20resource-catalogue/oai-nb-no-sbr-17/.

Boersma, Paul (2001). "Praat, a system for doing phonetics by computer". *Glot. Int.* 5.9, pp. 341–345.

Braunschweiler, Norbert, Mark JF Gales, and Sabine Buchholz (2010). "Lightly supervised recognition for automatic alignment of large coherent speech recordings". In: *Eleventh Annual Conference of the International Speech Communication Association*.

Cassidy, Steve and Thomas Schmidt (2017). "Tools for multimodal annotation". *Handbook of linguistic annotation*, pp. 209–227.

Chambers, Jack K and Peter Trudgill (1998). *Dialectology*. Cambridge University Press.

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

Delsing, Lars-Olof (2005). *Håller språket ihop Norden?: en forskningsrapport om ungdomars förståelse av danska, svenska och norska*. Nordic Council of Ministers.

Fromont, Robert and Kevin Watson (2016). "Factors influencing automatic segmental alignment of sociophonetic corpora". *Corpora* 11.3, pp. 401–431.

Furtună, Titus Felix (2008). "Dynamic programming algorithms in speech recognition". *Revista Informatica Economică nr* 2.46, p. 94.

Furui, Sadaoki (1986). "Speaker-independent isolated word recognition based on emphasized spectral dynamics". In: *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 11. IEEE, pp. 1991–1994.

Gonzalez, Simon, James Grama, and Catherine E Travis (2020). "Comparing the performance of forced aligners used in sociophonetic research". *Linguistics Vanguard* 6.1.

Gooskens, Charlotte (2007). "The contribution of linguistic factors to the intelligibility of closely related languages". *Journal of Multilingual and multicultural development* 28.6, pp. 445–467.

Gooskens, Charlotte (2020). "The North Germanic Dialect Continuum". *The Cambridge Handbook of Germanic Linguistics*, pp. 761–782.

Gorman, Kyle, Jonathan Howell, and Michael Wagner (2011). "Prosodylab-aligner: A tool for forced alignment of laboratory speech". *Canadian Acoustics* 39, p. 192.

Graves, Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber (2006). "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376.

Haugen, Einar (1966). "Semicommunication: The language gap in Scandinavia". *Sociological inquiry* 36.2, pp. 280–297.

Hoffmann, Sarah and Beat Pfister (2013). "Text-to-speech alignment of long recordings using universal phone models." In: *INTERSPEECH*. Citeseer, pp. 1520–1524.

Kamath, Uday, John Liu, and James Whitaker (2019). *Deep learning for NLP and speech recognition*. Vol. 84. Springer.

Kearns, Jodi (2014). "Librivox: Free public domain audiobooks". *Reference Reviews*.

Kempton, Timothy (2017). "Cross-language forced alignment to assist community-based linguistics for low resource languages". In: *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 165–169.

Kisler, Thomas, Uwe Reichel, and Florian Schiel (2017). "Multilingual processing of speech via web services". *Computer Speech & Language* 45, pp. 326–347.

Kisler, Thomas, Florian Schiel, and Han Sloetjes (2012). "Signal processing via web services: the use case WebMAUS". In: *Digital Humanities Conference 2012*.

Kürzinger, Ludwig, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll (2020). "Ctc-segmentation of large corpora for german end-to-end speech recognition". In: *International Conference on Speech and Computer*. Springer, pp. 267–278.

Leinonen, Juho et al. (2015). "Automatic speech recognition for human-robot interaction using an under-resourced language". MA thesis.

Leinonen, Juho, Sami Virpioja, Mikko Kurimo, et al. (2021). "Grapheme-Based Cross-Language Forced Alignment: Results with Uralic Languages". In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Linköping University Electronic Press.

Lingit (2015). *NB Tale - Speech Database for Norwegian*. https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-31/. Accessed: 2022-05-07.

MacKenzie, Laurel and Danielle Turton (2020). "Assessing the accuracy of existing forced alignment software on varieties of British English". *Linguistics Vanguard* 6.s1.

Mahr, Tristan J, Visar Berisha, Kan Kawabata, Julie Liss, and Katherine C Hustad (2021). "Performance of forced-alignment algorithms on children's speech". *Journal of Speech, Language, and Hearing Research* 64.6S, pp. 2213–2222.

Malmsten, M (2021). *KBLAB/WAV2VEC2-large-xlsr-53-swedish. Hugging Face*. https://huggingface.co/KBLab/wav2vec2-large-xlsr-53-swedish/. Accessed: 2022-04-25.

Malmsten, Martin, Chris Haffenden, and Love Börjeson (2022). "Hearing voices at the National Library–a speech corpus and acoustic model for the Swedish language". *arXiv preprint arXiv:2205.03026*.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." In: *Interspeech*. Vol. 2017, pp. 498–502.

Meer, Philipp (2020). "Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English". *The Journal of the Acoustical Society of America* 147.4, pp. 2283–2294.

Morgan, Douglas O'Shaughnessy (2010). *Research Developments and Directions in Speech Recognition and Understanding, Part 1.*

Ochshorn, Robert M and Max Hawkins (2017). *Gentle: A robust yet lenient forced aligner built on Kaldi.*

Permanasari, Yurika, Erwin H Harahap, and Erwin Prayoga Ali (2019). "Speech recognition using dynamic time warping (DTW)". In: *Journal of Physics: Conference Series*. Vol. 1366. 1. IOP Publishing, p. 012091.

Pettarin, Alberto (2017). *Aeneas*. URL: https://www.readbeyond.it/aeneas/.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. (2011). "The Kaldi speech recognition toolkit". In: *IEEE 2011 workshop on automatic speech recognition and understanding*. CONF. IEEE Signal Processing Society.

Reddy, Sravana and James N Stanford (2015). "Toward completely automated vowel extraction: Introducing DARLA". *Linguistics Vanguard* 1.1, pp. 15–28.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, and Jiahong Yuan (2011). "FAVE (forced alignment and vowel extraction) program suite". *URL http://fave. ling. upenn. edu.*

Sahlgren, Magnus, Fredrik Carlsson, Fredrik Olsson, and Love Börjeson (2021). "It's Basically the Same Language Anyway: the Case for a Nordic Language Model". In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 367–372.

Sakoe, Hiroaki and Seibi Chiba (1978). "Dynamic programming algorithm optimization for spoken word recognition". *IEEE transactions on acoustics, speech, and signal processing* 26.1, pp. 43–49.

Schiel, Florian (1999). "Automatic phonetic transcription of non-prompted speech".

Schiel, Florian (2022). personal communication. Apr. 12, 2022.

Schultz, Tanja, Ngoc Thang Vu, and Tim Schlippe (2013). "Globalphone: A multilingual text & speech database in 20 languages". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 8126–8130.

Shrawankar, Urmila and Vilas M Thakare (2013). "Techniques for feature extraction in speech recognition system: A comparative study". *arXiv preprint arXiv:1305.1145*.

Språkrådet (2017). *Aksentteikn*. https://www.sprakradet.no/sprakhjelp/Skriveregler/tegn/Aksentteikn/. Accessed: 2022-05-07.

Stern, Melissa K and James H Johnson (2010). "Just noticeable difference". *The Corsini Encyclopedia of Psychology*, pp. 1–2.

Strunk, Jan, Florian Schiel, Frank Seifart, et al. (2014). "Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS." In: *LREC*. Citeseer, pp. 3940–3947.

Tang, Kevin and Ryan Bennett (2019). "Unite and conquer: Bootstrapping forced alignment tools for closely-related minority languages (mayan)". In: *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, pp. 1719–1723.

Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. (2002). "The HTK book". *Cambridge university engineering department* 3.175.

A.

Aligner	BD			Tolerance				
	Mean (s)	**Median (s)**	**SD (s)**	**<0.01s**	**<0.025s**	**<0.05s**	**<0.1s**	**<0.5**
Semi-spontaneous	0.302	0.034	0.668	18.81%	41.58%	60.40%	69.80%	82.18%
Spontaneous	1.311	0.616	1.691	8.45%	17.91%	27.03%	33.78%	47.30%
Read	0.329	0.028	2.280	19.70%	42.42%	69.70%	90.91%	98.48%
Total	0.835	0.072	1.584	13.48%	29.26%	43.97%	53.37%	65.78%

Table A.2.: Descriptive statistics of boundary displacement with accuracies at different tolerance thresholds for the Norwegian dataset created by WebMAUS with the Norwegian model.