UTILIZING FORCED ALIGNMENT FOR PHONETIC ANALYSIS OF SLOVENE SPEECH

Janez KRIŽAJ,[1] Jerneja ŽGANEC GROS,[2] Simon DOBRIŠEK[1]
[1] Laboratory for machine intelligence, Faculty of electrical engineering, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia
[2] Alpineon R&D, d. o. o., Ulica Iga Grudna 15, 1000 Ljubljana

This paper explores the application of the Montreal Forced Aligner (MFA) for phonetic analysis of Slovene speech. We highlight MFA's alignment effectiveness and enhance alignment outputs with additional linguistic metrics and acoustic measurements. The alignment accuracy is evaluated on the GOS2.1 corpus, analyzing different subcorpora to assess the effectiveness of MFA across various types of speech. The research results show that MFA achieves alignment accuracy comparable to state-of-the-art, confirming its usefulness for phonetic research and applications in language technologies.

Keywords: forced alignment, speech corpuses for Slovene, acoustic analysis

1 INTRODUCTION

Forced alignment is a process in speech technology where given transcriptions are automatically aligned with the corresponding audio. This technique finds extensive application in various speech processing tasks, including automatic speech recognition (ASR), speech synthesis, subtitle generation, audio anonymization and phonetic research where it is used to study phenomena like speech reduction (Adda-Decker & Lamel, 2018), historical sound changes in languages (Labov et al., 2013), diagnosing speech sound disorders (Y. Li et al., 2023), and speech disfluencies (Kouzelis et al., 2023).

Forced alignment techniques play a pivotal role in the field of phonetic analysis by providing precise and efficient tools for aligning phonetic transcriptions with audio recordings. This capability is essential for researchers who seek to analyze and understand the intricate details of speech production and variation across different languages and dialects. The Montreal Forced Aligner (MFA) is a widely used tool in this domain, known for its robustness and accuracy in aligning phonetic segments with corresponding audio (Wu et al., 2023), (Chodroff et al., 2024).

We have chosen the MFA as the primary method for forced alignment due to its comprehensive feature set and proven track record in various linguistic studies. Our contributions include the forced alignment of the Slovene speech corpus GOS2.1 (Verdonik et al., 2023) at the word, syllable, and phone levels, and its application to basic acoustic measurements. We also compare its performance with the more recent NeMo Forced Aligner (NFA) (Rastorgueva et al., 2023) to evaluate its comparability with state-of-the-art methods. The source code containing alignment procedures and acoustic measurements is freely available at https://github.com/jan3zk/forced_alignment.

The paper is structured as follows: we begin with a review of related work, then describe the forced alignment methodology and our experiments, and conclude with potential avenues for future research.

2 RELATED WORK

The application of forced alignment to phonetic research has been increasingly popular, as demonstrated by studies like (Yuan et al., 2023), which used forced alignment for phonetic segmentation and investigating speech variation. Contribution by (Young & McGarrah, 2023), focuses on the study of forced alignment on a lesser-known language with a limited amount of training data. The work of (Adda-Decker & Lamel, 2018) explores discovering speech reductions across different speaking styles and languages, underlining the versatility of forced alignment in diverse linguistic contexts. Moreover, (Kouzelis et al., 2023) emphasized the alignment of disfluent speech, eliminating the need for verbatim transcription. Additionally, Huang et al. (2024), J. Li et al. (2022) and Sun (2023) present examples of recent deep learning-based approaches, furthering the innovation in the field of forced alignment.

On the software front, several forced alignment solutions exist. PraatAlign (Lubbers & Torreira, 2013-2018), is effective but not scalable for larger datasets. Deep learning approaches such as NVIDIA NeMo (Rastorgueva et al., 2023) and WhisperX (Bain et al., 2023) offer accurate and efficient word-level alignment but lack alignments at the phone-level. For our research, we utilized the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) for its proficiency in providing phone-level alignments and its robust performance in our experimental setup.

3 FORCED ALIGNMENT FOR PHONETIC RESEARCH

3.1 Montreal Forced Aligner

In the MFA alignment process, the pronunciation dictionary is first used to convert the transcription into phones. Next, the acoustic model determines which parts of the audio recording correspond to the specific phones from the dictionary. The pronunciation dictionary provides the phones to search for, while the acoustic model assesses how well the audio patterns match these phones.

The pronunciation dictionary contains a list of words along with their phonetic transcriptions, indicating which phones compose each word. This transcription is crucial, as it informs the acoustic model of the sounds to search for in the audio signal. The acoustic model is trained to recognize specific phones from audio features such as Mel-Frequency Cepstral Coefficients (MFCCs). It models how individual phones sound in different contexts, such as monophones and triphones, as well as with varying speech styles or accents. The acoustic model is based on Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs), which determine how likely a given set of audio features corresponds to a specific phone or sequence of phones.

In finding the most likely path through these models, which allows aligning the audio signal with the text transcription, the MFA employs the Viterbi algorithm, optimized for efficient search of the most probable sequences of phones that align with the observed audio features.

The MFA process outputs a detailed information about the timing of each phone and word in the input speech audio, as depicted in the flowchart (Figure 1).

3.2 Additional Tiers for Phonetic Research

The output of MFA initially provides only two tiers in the output files, containing alignments at the word and phone levels. To enhance the utility of these files for linguistic and phonetic research, we have introduced additional tiers into the annotation files, as shown in Figure 2. These tiers are designed to facilitate

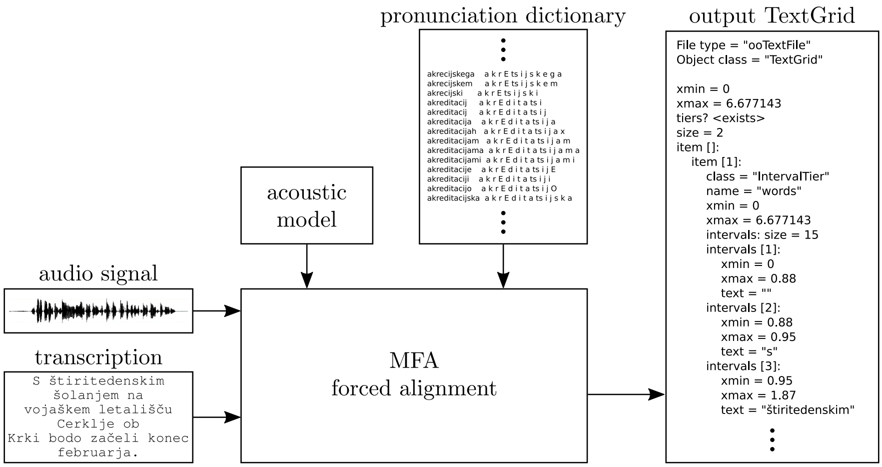

Figure 1: Schematic representation of the forced alignment process. On the left are the input files: a speech audio and its transcription. Given the acoustic model and the pronunciation dictionary, the MFA produces an output TextGrid file, which contains time intervals at both the word and phone levels.

more detailed analyses of Slovene speech allowing researchers to delve into various aspects of speech, including prosody, discourse, and conversational dynamics, which are pivotal for comprehensive linguistic research.

Several tiers are derived from data in the transcript files, which are provided as corpus metadata, and combined with time intervals obtained through forced alignment. These tiers offer information regarding speaker IDs, word IDs, speaker changes, and conversational text transcriptions.

Other tiers result from further processing of the corresponding audio file and/or data within the existing tiers. These tiers include information on discourse markers, syllable intervals, pitch resets, intensity resets, speech rate reductions, and pauses, enriching the scope of phonetic analysis.

3.3 Acoustic Measurements

Utilizing the synchronized audio and transcript files, our study involves the computation of a variety of acoustic metrics essential for subsequent speech analysis. For each phone, we calculate various acoustic features including the

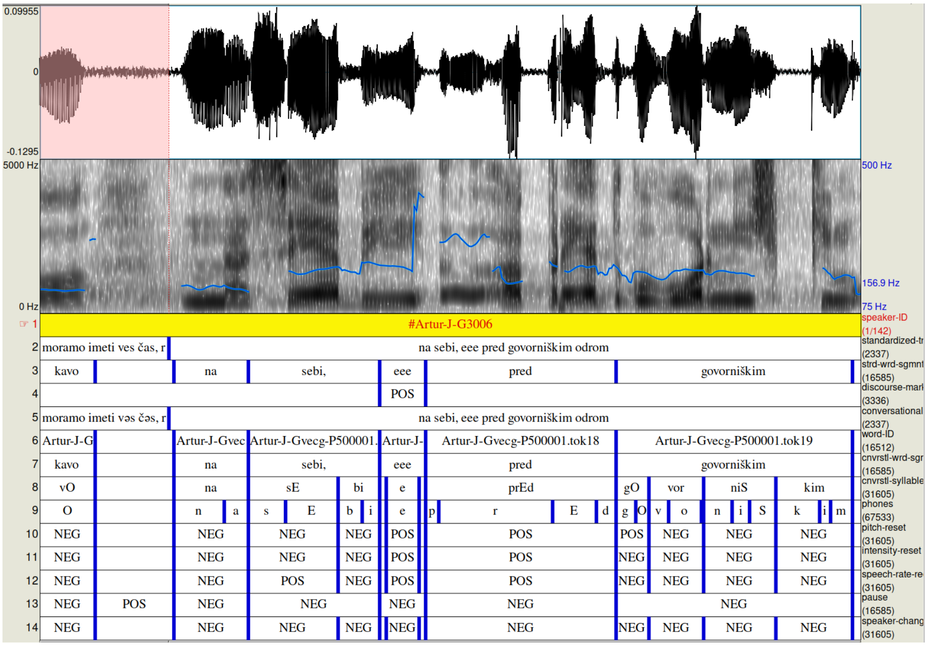

Figure 2: Praat software displaying additional TextGrid tiers for an in-depth analysis of Slovene speech.

duration of each phone, average pitch values, pitch trend, formant frequencies, intensity levels, voice onset time, and the center of gravity.

4 EXPERIMENTS

4.1 Databases

The Gos 2.1 corpus (Verdonik, Zwitter Vitez, et al., 2023) is a comprehensive reference speech corpus for the Slovenian language, comprising about 300 hours of speech (2.4 million words and 127 thousand utterances from 1,500 texts), enhanced with word-level temporal information where possible. It amalgamates data from three sources: the Gos 1.1 corpus (112 hours, 1 million words), the Gos VideoLectures 4.2 (22 hours, 179,000 words), and selections from the ARTUR 1.0 ASR database (185 hours, 1.2 million words), including varied content like media recordings, dialogues, and transcriptions from the Slovene National Assembly. Transcriptions are provided in two forms: pronunciation-based and standardised orthographic transcriptions. This edition features improvements such as unified casing and normalization, re-introduction of punctuation, additional temporal data, and uniform encoding across subcorpora. Corpus transcriptions are presented in TEI (XML) format and include part-of-speech tagging and lemmatisation, automated by CLASSLA (Ljubešić & Dobrovoljc, 2019).

4.2 Experimental Setup

We adopted the training approach outlined by (McAuliffe et al., 2017) for MFA alignment, training the acoustic model on the Artur-B subset (Verdonik et al., 2023). This subset includes 485 hours of read speech and is distinct from the GOS2.1 dataset, which is used for subsequent alignment evaluations.

4.3 Alignment Accuracy Evaluation

In the absence of ground truth time intervals, we conducted a comparative analysis of the MFA with two other forced alignment methods: Nemo Forced Aligner (NFA) (Rastorgueva et al., 2023) and temporal intervals obtained by Kaldi recipe which are part of GOS2.1 database metadata. The NFA uses a pretrained Conformer Connectionist Temporal Classification (CTC) ASR model, trained on the Artur corpus as described by (Lebar Bajec et al., 2022), while the Kaldi approach utilizes the Artur-P subset for training (Verdonik et al., 2024).

For our evaluation, each method was alternately used as the reference (pseudo-ground truth) to assess the relative performance of the others. The models were assessed by calculating the mean and median absolute errors between the predicted and reference temporal boundaries at the word level. Additionally, we computed the proportion of boundaries falling within specific tolerance thresholds (10 ms, 50 ms, 100 ms). These metrics are restricted to the word level due to the lack of phone-level alignments in the Kaldi and NFA methods.

Each of the assessed methods served as the reference (pseudo-ground truth) in turn allowing us to evaluate the relative performance of the other methods. The evaluation of the models is conducted by assessing the mean and median of absolute errors between the predicted temporal boundaries and reference values at the word-level temporal data as well as share of boundaries that fall within specific tolerance thresholds (10 ms, 50 ms, 100 ms). These evaluation

Table 1: Performance assessment of the evaluated methods at the word level across different subsets of the GOS2.1 database.

Corpus	Method	Mean abs. err. [ms]	Median abs. err. [ms]	Share within 10 ms [%]	Share within 50 ms [%]	Share within 100 ms [%]
Artur-P	MFA vs Kaldi	28	11	47	90	96
	NFA vs Kaldi	71	49	11	51	86
	MFA vs NFA	77	50	10	50	84
Artur-J	MFA vs Kaldi	48	14	40	84	92
	NFA vs Kaldi	86	50	11	50	83
	MFA vs NFA	86	52	10	48	52
Artur-N	MFA vs Kaldi	208	14	41	83	90
	NFA vs Kaldi	450	50	11	50	80
	MFA vs NFA	457	52	10	48	79
GosVL	MFA vs Kaldi	212	12	44	84	90
	NFA vs Kaldi	184	91	3	22	56
	MFA vs NFA	190	91	4	22	57

metrics are calculated at the word level due to the absence of phone level annotations in case of Kaldi and NFA methods.

The results, summarized in Table 1, reveal significant differences in alignment accuracy across different subcorpora. Not surprisingly, we observed higher error rates in the Artur-N and GosVL corpora, which feature more complex audio elements like spontaneous conversations, overlapping speech, and speakers wearing masks. When compared against the Kaldi reference, the MFA predicts over 90% of time boundaries within a 100 ms margin across all examined subcorpora. While the MFA method generally demonstrated higher accuracy compared to the NFA when evaluated against the Kaldi references, this may be attributed to the close methodological similarities between the MFA and Kaldi approaches. Upon manual inspection, we observed that NFA has favourable performance in several challenging scenarios, particularly in cases of overlapping indistinct speech and poor transcriptions.

Considering the effective performance of the Montreal Forced Aligner (MFA) and the requirement for significant adaptations in the NeMo Forced Aligner (NFA) tokens to align with phones from our pronunciation dictionary for phone and syllable level alignment, we opted to employ MFA for our subsequent phonetic studies.

5 CONCLUSION

This study demonstrates the effective application of the Montreal Forced Aligner for phonetic segmentation and acoustic analysis of Slovene speech. Our alignment of the GOS2.1 corpus, along with additional tier annotations and acoustic measurements, contributes to Slovene phonetics by providing a faster alternative to manual alignments.

Future research will focus on automatic detection of speech disfluencies and examining phonetic variations in relation to corpus metadata, different speech styles, and different dialects. This effort aims to deepen our understanding of speech dynamics and phonetic characteristics in Slovene language.

6 ACKNOWLEDGMENTS

This research was made possible through the support of the MEZZANINE project (teMeljnE raZiskave Za rAzvoj govorNih vIrov in tehNologij za slovEnščino – Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language), project ID: J7-4642, financed by the Slovenian Research Agency.

REFERENCES

Adda-Decker, M., & Lamel, L. (2018). Discovering speech reductions across speaking styles and languages. In F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, & M. Zellers (Eds.), *Rethinking reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation* (pp. 101–128). Berlin, Boston: De Gruyter Mouton. doi: doi:10.1515/9783110524178-004

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Chodroff, E., Ahn, E., & Dolatian, H. (2024). Comparing language-specific and cross-language acoustic models for low-resource phonetic forced alignment. *Language Documentation & Conservation*.

Huang, R., Zhang, X., Ni, Z., Sun, L., Hira, M., Hwang, J., Manohar, V., Pratap, V., Wiesner, M., Watanabe, S., Povey, D., & Khudanpur, S. (2024). Less peaky and more accurate ctc forced alignment by label priors. In *Icassp 2024 - 2024 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 11831-11835). doi: 10.1109/ICASSP48485.2024.10446111

Kouzelis, T., Paraskevopoulos, G., Katsamanis, A., & Katsouros, V. (2023. Weakly-supervised forced alignment of disfluent speech using phoneme-level modeling. In (p. 1563-1567). doi: 10.21437/Interspeech.2023-1887

Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, *89*(1), 30–65. Retrieved 2024-04-08, from http://www.jstor.org/stable/23357721

Lebar Bajec, I., Bajec, M., Bajec, Ž., & Rizvič, M. (2022). *Slovene conformer CTC BPE E2E automated speech recognition model RSDO-DS2-ASR-E2E 2.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle .net/11356/1737

Li, J., Meng, Y., Wu, Z., Meng, H., Tian, Q., Wang, Y., & Wang, Y. (2022). Neufa: Neural network based end-to-end forced alignment with bidirectional attention mechanism. In *Icassp 2022 - 2022 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 8007-8011). doi: 10.1109/ICASSP43922.2022.9747085

Li, Y., Wohlan, B. J., Pham, D.-S., Chan, K. Y., Ward, R., Hennessey, N., & Tan, T. (2023). Improving text-independent forced alignment to support speech-language pathologists with phonetic transcription. *Sensors*, *23*(24). https://www.mdpi.com/1424-8220/23/24/9650 doi: 10.3390/s23249650

Ljubešić, N., & Dobrovoljc, K. (2019, August). What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th workshop on balto-slavic natural language processing* (pp. 29–34). Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-3704 doi: 10.18653/v1/W19-3704

Lubbers, M., & Torreira, F. (2013-2018). *Praatalign: an interactive praat plug-in for performing phonetic forced alignment.* https://github.com/dopefishh/praatalign. (Version 2.0)

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. interspeech 2017* (pp. 498–502). doi: 10.21437/Interspeech.2017-1386

Rastorgueva, E., Lavrukhin, V., & Ginsburg, B. (2023). NeMo Forced Aligner and its application to word alignment for subtitle generation. In *Proc. interspeech 2023* (pp. 5257–5258).

Sun, L. (2023). Unsupervised forced alignment on syllable and phoneme with universal phonetics transcriptions. In *2023 ieee international conference on signal processing, communications and computing (icspcc)* (p. 1-5). doi: 10.1109/

ICSPCC59353.2023.10400287

Verdonik, D., Bizjak, A., Žgank, A., Bernjak, M., Antloga, Š., Majhenič, S., ... Bordon, D. (2023). *ASR database ARTUR 1.0 (audio).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1776

Verdonik, D., Dobrovoljc, K., Erjavec, T., & Ljubesic, N. (2024). Gos 2: A new reference corpus of spoken slovenian. In *International conference on language resources and evaluation.* https://api.semanticscholar.org/CorpusID:269804654

Verdonik, D., Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., ... Rupnik, P. (2023). *Spoken corpus gos 2.1 (transcriptions).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/ 11356/1863

Wu, H., Yun, J., Li, X., Huang, H., & Liu, C. (2023, 07). Using a forced aligner for prosody research. *Humanities and Social Sciences Communications*, *10*. doi: 10.1057/s41599-023-01931-4

Young, N. J., & McGarrah, M. (2023). Forced alignment for nordic languages: Rapidly constructing a high-quality prototype. *Nordic Journal of Linguistics*, *46*(1), 105–131. doi: 10.1017/S033258652100024X

Yuan, J., Lai, W., Cieri, C., & Liberman, M. (2023). Using forced alignment for phonetics research. In C.-R. Huang, S.-K. Hsieh, & P. Jin (Eds.), *Chinese language resources: Data collection, linguistic analysis, annotation and language processing* (pp. 289–301). CSpringer International Publishing. doi: 10.1007/ 978-3-031-38913-9_17

UPORABA VSILJENE PORAVNAVE ZA FONETIČNO ANALIZO SLOVENSKEGA GOVORA

V članku proučimo uporabo knjižnice MFA (ang. Montreal Forced Aligner) za vsiljeno poravnavo posnetkov slovenskega govora. Udejanjimo izračun dodatnih jezikovnih in akustične meritve, ki omogočajo poglobljeno analizo slovenskega govora. Natančnost poravnave ovrednotimo na korpusu GOS2.1, pri čemer z analizo različnih podkorpusov ocenimo učinkovitost MFA pri različnih vrstah govora. Rezultati raziskave kažejo, da MFA dosega visoko natančnost pri poravnavi, kar potrjuje njegovo uporabnost za nadaljnje fonetične raziskave in aplikacije v jezikovnih tehnologijah.

Ključne besede: vsiljena poravnava, govorni korpusi za slovenščino, akustična analiza