Grapheme-Based Cross-Language Forced Alignment: Results with Uralic Languages

Juho Leinonen
Aalto University
juho.leinonen
@aalto.fi

Sami Virpioja
Helsinki University
sami.virpioja
@helsinki.fi

Mikko Kurimo
Aalto University
mikko.kurimo
@aalto.fi

Abstract

Forced alignment is an effective process to speed up linguistic research. However, most forced aligners are language-dependent, and under-resourced languages rarely have enough resources to train an acoustic model for an aligner. We present a new Finnish grapheme-based forced aligner and demonstrate its performance by aligning multiple Uralic languages and English as an unrelated language. We show that even a simple non-expert created grapheme-to-phoneme mapping can result in useful word alignments.

1 Introduction

Matching speech signal and its orthographic transcription is a necessary first step for many research questions in linguistics (Yuan et al., 2018; Olsen et al., 2017; DiCanio et al., 2013). For well-resourced languages, manually aligned corpora exist, providing an easy starting point for linguistic research. For under-resourced languages such corpora are rare, and for all languages new corpora are continuously studied. In these situations, the researcher needs to complete this task before any actual research can begin. Forced alignment, i.e., automatically matching text to speech using automatic speech recognition (ASR), is widely used, and tools that can accomplish this automatically exist, such as FAVE (Rosenfelder et al., 2011), Prosodylab-aligner (Gorman et al., 2011), MAUS (Kisler et al., 2017), and Montreal Forced aligner (MFA) (McAuliffe et al., 2017).

If the researcher is studying a language that is supported by an existing tool for forced alignment, learning to use it will be beneficial, since manual segmentation is much more arduous than transcription (Jarifi et al., 2008). However, the effort for this necessary, but often uninteresting step increases tremendously if no suitable model exits. The reason may be that the target data is out-of-domain of what the acoustic model was trained with, or the target language is under-resourced and there is no model available at all. Some aligning tools do not support retraining models. For others, such as FAVE and Prosodylab, the model has been trained with a known ASR framework, here HTK (Young et al., 2002), and the researcher could use the framework to train their own models. However, at this point it would be more straightforward to use the ASR framework itself. In addition to all of this, the technical knowledge required to train an acoustic model with minimal or difficult data is formidable.

MFA provides ample documentation, and has a user friendly wrapper over Kaldi (Povey et al., 2011), a popular speech recognition framework. It gives users the option to retrain the model to fit their own data, and add new languages. Gonzalez et al. (2018) used MFA to experiment on iterative forced alignment, and how it compared to the traditional linear method. Even though they used a ready-made tool, the effort to try two alignment methods on an under-resourced language was enough to qualify as a research paper on its own right. For a linguist, who might not have technical expertise on ASR, this may be intimidating as the first step.

An alternative solution to the task of training new models is cross-language forced alignment, in which an aligner trained with a different language than the speech and transcriptions to be aligned, is used. In this paper we introduce a new word-level forced alignment tool based on Kaldi. We show that this very simple command line tool can align closely related languages, is robust against speaker variability without any fine-tuning, and can even adequately align linguistically very dissimilar languages. This paper shows the first results for cross-language forced alignment involving Finnish. In addition, using the tool we forcealigned a Northern Sámi corpus without proper word alignments with very little expert knowledge of the language.

2 Related research

2.1 Forced aligners

In their paper (McAuliffe et al., 2017), the designers of MFA compared their tool to FAVE and Prosodylab. The latter tools are based on monophone models, while MFA utilizes triphones, and adds speaker adaptation to the process. A central underlying difference is that, similar to us, MFA uses Kaldi as the speech recognition framework. However, MFA uses Gaussian mixture models (GMM), popular in speech recognition before deep neural networks (DNN), while our tool uses the modern machine learning methods trained with Kaldi's lattice-free maximum mutual information cost function (Hadian et al., 2018). Another Kaldi-based tool is Gentle [1], which also uses DNNs. Munich AUtomatic Segmentation system (MAUS) is a popular aligner based on its own speech recognition framework, utilizing a statistical expert system of pronunciation.

2.2 Cross-language forced alignment

Forced alignment has also been successfully used across languages, e.g., when the target language does not have enough transcribed data. This task is called cross-language or cross-linguistic forced alignment (CLFA), sometimes untrained forced alignment. Kempton et al. (2011) used their own phonetic distance metric to evaluate the accuracy of three phoneme recognizers on isolated words from under-resourced language, and again in (Kempton, 2017) to a different target language. In another early experiment (DiCanio et al., 2013), tools trained on English were used to align isolated words from Yoloxóchitl Mixtec. Free conversations were aligned in (Kurtic et al., 2012), where authors tested multiple phoneme recognizers on Bosnian Serbo-Croatian.

Most of the tools introduced at the start of this section have also been tried for CLFA. The authors of MAUS experimented a language-independent 'sampa' version on a multitude of under-resourced languages by comparing word start and end boundaries (Strunk et al., 2014). Later Jones et al. (2019) compared MAUS' language-independent and Italian versions for conversational speech in Kriol, finding that the Italian version surpassed the language-independent one.

A unifying method was presented by Tang and Bennett (2019), who combined a larger source language and the target language with MFA to train the aligner. Finally Johnson et al. (2018) reviewed previous CLFA research and experimented on the minimum amount of data necessary for language dependent forced alignment, achieving good results with an hour of transcribed speech.

3 Experiments

We evaluate our Kaldi-based aligner on related and unrelated languages, with a small amount of expert knowledge added to grapheme-to-phoneme mapping. We also experiment on speaker variation. This is the first time either has been done in CLFA literature. The code and tool used in this paper are publicly available. [2]

3.1 Kaldi pipeline

Our method uses Kaldi to force-align transcibed audio. As is customary in Kaldi when aligning speech with neural networks, we employ 39 dimension Mel-frequency cepstral coefficients (MFCCs) and Cepstral mean and variance normalization (CMVN). Kaldi's i-vectors are used for speaker adaptation. The original Finnish acoustic model and i-vector exctractor are the same as in (Mansikkaniemi et al., 2017). After the feature generation we create a dataset-specific dictionary from all the words in the transcription. The orthography is assumed to be phonetic, so the words in the lexicon are composed of their graphemes, which are mapped to closest Finnish match manually by non-experts. Smit et al. (2021) show that with DNN-based acoustic models, the assumption of phonetic orthography works reasonably well even for a language like English. As a final preparation for alignment Kaldi uses the lexicon, acoustic model and transcripts to create dataset-specific finite state transducers.

3.2 Datasets

We first evaluate the model on Finnish data using manually annotated Finnish read speech from one male speaker (Vainio, 2001; Raitio et al., 2008). We use Pympi (Lubbers and Torreira, 2013-2015)

[1] https://github.com/lowerquality/gentle

[2] https://github.com/aalto-speech/finnish-forced-alignment

to prepare the data. Here the grapheme-to-phoneme mapping is one to one due to Finnish being a phonetic language. For experimenting on speaker variability and CLFA, we align nine Estonian speakers with data gathered from the corpus of lecture speeches introduced in (Meister et al., 2012). For each speaker we have little over 15 minutes of speech, much less than the recommended hour by Johnson et al. (2018). We create a rough mapping between Estonian graphemes and Finnish phonemes, which is a straightforward task as the languages are closely related. We also evaluate our model on Northern Sámi, by force-aligning the Giellagas corpus (Kielipankki, 2014-2017). Since there are no accurate word boundaries for the dataset, we use ELAN (Wittenburg et al., 2006) to manually annotate roughly 20 seconds of speech from 11 native speakers to compare to our automatically generated boundaries. The annotations should be considered only approximative, as the recorded speech has poor quality and the annotator did not know the Sámi language. For Northern Sámi, we use the grapheme-to-phoneme mapping introduced by Leinonen (2015). While most of CLFA papers use closely related or otherwise similar languages, we also try to align English speech with our Finnish model using the clean test sets from Librispeech corpus (Panayotov et al., 2015). For the lexicon we map the graphemes e, and y to Finnish i, and a to ä, otherwise assuming one-to-one mapping.

For all datasets, we follow McAuliffe et al. (2017), and compare what percentage of absolute differences in word start and end boundaries are inside the ranges 10, 25, 50 and 100 milliseconds, when comparing the aligner's results to the gold standard boundaries. Since we do not have manual alignments for the English and Estonian datasets, we align the audio with language-dependent acoustic models and use the predicted boundaries as gold standards. For Estonian this is done with a dockerized Estonian aligner[3]. The Librispeech datasets were aligned with an acoustic model trained with Kaldi Librispeech recipe[4]. We use the final GMM-based model called tri6b to create the word boundaries. We also experiment with other triphone models trained with the Librispeech recipe, varying in the amounts of training data, and model complexity, to test what improvements the advances in triphone models bring, and how well our Finnish model compares to language dependent models. Table 1 summarizes the sizes of studied datasets.

Lang	Dataset	lenght	tokens
fin	Finnish	1h7m27s	6464
est	al	16m41s	1910
	ao	16m45s	2199
	hv	16m40s	1697
	jp	16m46s	1953
	mk	16m41s	2602
	ms	16m48s	1523
	mj	16m48s	1394
	mr	16m42s	2025
	th	16m48s	1344
eng	dev-clean	5h23m16s	54402
	test-clean	5h24m12s	52576
smi	Giellagas	3min19s	384

Table 1: Speech and text data used for evaluations, with initials of the participant names for Estonian data as they were in the corpus.

4 Results

The Finnish alignment results in Table 2 are quite comparable to what McAuliffe et al. (2017) achieved using MFA for the English Buckeye corpus (Pitt et al., 2005). This seems reasonable since both are using Kaldi. The different amounts of smaller boundary errors might be due to audio quality, speaking style or method of annotation. For instance the Finnish dataset was more focused on phoneme labels than word boundaries.

Model	Dataset	<10	<25	<50	<100
Finnish	Finnish	0.21	0.55	0.84	0.98
MFA	Buckeye	0.33	0.68	0.88	0.97

Table 2: Differences in word boundary accuracy between language-dependent forced alignment. MFA results from (McAuliffe et al., 2017) using the English Buckeye corpus.

When analysing the Estonian results in Table 3, they look comparable to Finnish. Aside from the last 100ms range, they are very similar to MFA's results for Buckeye. And for smaller ranges are actually better than Finnish alignments. This can be due to similarities in how the speech recognizers generally align speech. Speaker variation is small,

[3] https://github.com/alumae/kaldi-align-server

[4] https://github.com/kaldi-asr/kaldi/tree/master/egs/librispeech/s5

Speaker	<10	<25	<50	<100
al	0.32	0.65	0.82	0.90
ao	0.36	0.72	0.89	0.94
hv	0.32	0.64	0.81	0.88
jp	0.37	0.67	0.83	0.90
mk	0.29	0.59	0.77	0.88
ms	0.33	0.64	0.82	0.89
mj	0.38	0.70	0.86	0.92
mr	0.30	0.62	0.84	0.93
th	0.34	0.64	0.81	0.89
Median	0.33	0.64	0.82	0.90
Std	0.027	0.038	0.033	0.02

Table 3: Cross-language forced alignment for Estonian: results of word boundary accuracy for speaker-wise alignments with median and standard deviation.

with standard deviation being 0.02-0.038. Overall, compared to how well MFA aligned English speech, this is a more fat-tailed distribution, with 10% of boundary errors being larger than 100ms.

Dataset	<10	<25	<50	<100
Giellagas	0.12	0.26	0.45	0.62

Table 4: Cross-language forced alignment for Northern Sámi: word boundary accuracy using a part of the Giellagas corpus.

The results for Northern Sámi in Table 4 are not as good as for Estonian, with some of the possible reasons listed in Section 3.2. With closer inspection of the differences between manual and forced alignment, it could be argued that the automatic method is more accurate. It is definitely much faster, being seconds instead of taking hours.

Dataset	<10	<25	<50	<100
dev-clean	0.12	0.30	0.51	0.68
test-clean	0.12	0.30	0.51	0.67

Table 5: Cross-language forced alignment for English: word boundary accuracy using Librispeech datasets.

The results for English in Table 5 are weaker than for any other target language, with the largest 100ms range having the same results as 25ms range for Estonian. While any researcher who needs to align English speech naturally has language-dependent models, this demonstrates the worst case scenario for CLFA, with multiple wrong assumptions including rough grapheme-to-phoneme mapping, and even using phonetic orthography. If there is very little target speech, using an unrelated source language might be more cost effective than trying to train a new model or manual alignment.

Model	<10	<25	<50	<100
tri1	0.55	0.87	0.97	1.00
tri2b	0.65	0.93	0.98	1.00
tri3b	0.72	0.95	0.99	1.00
tri4b	0.80	0.97	0.99	1.00
tri5b	0.88	0.99	1.00	1.00

Table 6: Librispeech word boundary accuracy with different English HMM-GMM models trained with Librispeech recipe. Dataset is dev-clean, using tri6b as a gold standard.

The authors of MFA hypothesize the effects of using different phone models, speaker adaptive training and other methods in (McAuliffe et al., 2017). Also to give context to the Finnish-English results, we experimented on how simpler ASR models might perform at the task. Table 6 show that improving the basic model underneath does improve the results for the smallest ranges, and that a much simpler language-dependent model is much better than results with cross-language alignment.

5 Future work

Most of the papers in related research use some tool to automatically generate a phoneme-based lexicon for the target language. These lexicons do contain errors, so we have evaluated our results with word boundaries, since the words can be extracted as is from the transcription. However, automatic phoneme mapping would be an interesting next step, and allow better comparison with previous research effort in this multidisciplinary field.

6 Conclusion

We have demonstrated promising results for cross-language forced alignment using Finnish acoustic model for related and unrelated languages. We have shown that its results for Finnish in language-dependent use are comparable to state-of-the-art aligners for English data. In addition, we present promising results with related and unrelated languages. We also showed the effects of speaker variation in cross-language situations, demonstrating that retraining speaker dependent models is generally not necessary. We share our tool as an easy to use Docker image.

Acknowledgments

We acknowledge the computational resources provided by the Aalto Science-IT project. SV was supported by the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 771113).

References

Christian DiCanio, Hosung Nam, Douglas H Whalen, H Timothy Bunnell, Jonathan D Amith, and Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America*, 134(3):2235–2246.

Simon Gonzalez, Catherine Travis, James Grama, Danielle Barth, and Sunkulp Ananthanarayan. 2018. Recursive forced alignment: A test on a minority language. In *Proceedings of the 17th Australasian International Conference on Speech Science and Technology*, volume 145, page 148.

Kyle Gorman, Jonathan Howell, and Michael Wagner. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193.

Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur. 2018. End-to-end speech recognition using lattice-free mmi. In *Interspeech*, pages 12–16.

Safaa Jarifi, Dominique Pastor, and Olivier Rosec. 2008. A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis. *Speech communication*, 50(1):67–80.

Lisa M Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data. *Language Documentation & Conservation*, 12:80–123.

Caroline Jones, Weicong Li, Andre Almeida, and Amit German. 2019. Evaluating cross-linguistic forced alignment of conversational data in north australian kriol, an under-resourced language. *Language Documentation and Conservation*, pages 281–299.

Timothy Kempton. 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*, pages 165–169, Honolulu. Association for Computational Linguistics.

Timothy Kempton, Roger K Moore, and Thomas Hain. 2011. Cross-language phone recognition when the target language phoneme inventory is not known. In *Twelfth Annual Conference of the International Speech Communication Association*.

Kielipankki. 2014-2017. Pohjoissaamen näytekorpus. Http://urn.fi/urn:nbn:fi:lb-201407302.

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

Emina Kurtic, Bill Wells, Guy J Brown, Timothy Kempton, and Ahmet Aker. 2012. A corpus of spontaneous multi-party conversation in bosnian serbo-croatian and british english. In *LREC*, pages 1323–1327. Citeseer.

Juho Leinonen. 2015. Automatic speech recognition for human-robot interaction using an under-resourced language. Master's thesis, Aalto University School of Electrical Engineering, Espoo.

Mart Lubbers and Francisco Torreira. 2013-2015. pympi-ling: a python module for processing elans eaf and praats textgrid annotation files. `https://pypi.python.org/pypi/pympi-ling`. Version 1.69.

André Mansikkaniemi, Peter Smit, Mikko Kurimo, et al. 2017. Automatic construction of the Finnish parliament speech corpus. In *INTERSPEECH 2017–18th Annual Conference of the International Speech Communication Association*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Einar Meister, Lya Meister, and Rainer Metsvahi. 2012. New speech corpora at IoC. In *XXVII Fonetiikan päivät 2012 — Phonetics Symposium 2012: 17–18 February 2012, Tallinn, Estonia: Proceedings*, pages 30–33. TUT Press.

Rachel M Olsen, Michael L Olsen, Joseph A Stanley, Margaret EL Renwick, and William Kretzschmar. 2017. Methods for transcription and forced alignment of a legacy speech corpus. In *Proceedings of Meetings on Acoustics 173EAA*, volume 30, page 060001. Acoustical Society of America.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Daniel Povey, 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.

Tuomo Raitio, Antti Suni, Hannu Pulakka, Martti Vainio, and Paavo Alku. 2008. Hmm-based finnish text-to-speech system utilizing glottal inverse filtering. In *Ninth Annual Conference of the International Speech Communication Association*.

Ingrid Rosenfelder, Josef Fruehwald, Keelan Evanini, and Jiahong Yuan. 2011. Fave (forced alignment and vowel extraction) program suite. Http://fave. ling. upenn. edu.

Peter Smit, Sami Virpioja, and Mikko Kurimo. 2021. Advances in subword-based hmm-dnn speech recognition across languages. *Computer Speech & Language*, 66:101158.

Jan Strunk, Florian Schiel, Frank Seifart, et al. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using webmaus. In *LREC*, pages 3940–3947.

Kevin Tang 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*, pages 1719–1723.

Martti Vainio. 2001. Artificial neural network based prosody models for finnish text-to-speech synthesis.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

Steve Young, 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):12.

Jiahong Yuan, Wei Lai, Chris Cieri, and Mark Liberman. 2018. Using forced alignment for phonetics research. *Chinese Language Resources and Processing: Text, Speech and Language Technology. Springer*.