Proceedings of the EURALI Workshop @LREC2020, pages 17–21
Marseille, 20 June 2022

Semiautomatic Speech Alignment for Under-Resourced Languages

Juho Leinonen[1], Niko Partanen[2], Sami Virpioja[2], Mikko Kurimo[1]
[1]Aalto University, [2]University of Helsinki
{juho.leinonen, mikko.kurimo}@aalto.fi
{niko.partanen, sami.virpioja}@helsinki.fi

Abstract

Cross-language forced alignment is a solution for linguists who create speech corpora for very low-resource languages. However, cross-language is an additional challenge making a complex task, forced alignment, even more difficult. We study how linguists can impart domain expertise to the tasks to increase the performance of automatic forced aligners while keeping the time effort still lower than with manual forced alignment. First, we show that speech recognizers have a clear bias in starting the word later than a human annotator, which results in micro-pauses in the results that do not exist in manual alignments, and study which is the best way to automatically remove these silences. Second, we ask the linguists to simplify the task by splitting long interview audios into shorter lengths by providing some manually aligned segments and evaluating the results of this process. Finally, we study how correlated source language performance is to target language performance, since often it is an easier task to find a better source model than to adapt to the target language.

Keywords: speech recognition, cross-language forced alignment, low-resource

1. Introduction

When collecting new speech corpora, a valuable type of metadata to add is timestamps for the words in the audio. These are useful for other researchers for checking the context of the spoken word and for speech recognition and synthesis research. This matching of text to speech is called forced alignment, and it is necessary for many linguists' work. Both word and utterance-level alignments have various uses and clear benefits when compared to transcriptions that have no timestamp information available. Creating word-level alignments from utterance-level annotations can also be considered a special context of forced alignment.

Currently, many tools have automated the alignment process with speech recognizers. However, this automatic forced alignment is limited to languages with capable speech recognizers. While there are ready-made recipes to train a speech recognizer when given data, one should not underestimate the domain expertise required to accomplish this, especially if an error occurs and needs fixing. Another significant issue is the data. A large corpus can have a recognizer trained on it, and afterward, it can align the data. Nevertheless, there are languages and domains with insufficient data to train a recognizer. Especially in case of seriously underdocumented languages there are no possibilities to have larger amounts of training data of any type. Here, cross-language forced alignment can help researchers quickly create good alignments with significantly less effort than manual aligning. However, cross-language recognition adds complexity to an already challenging task, so we examine what researchers might do to make this task easier.

Automatic forced alignment is not a new concept. Many of the first automatic speech recognition (ASR) systems have been used for generating alignments since it is a natural part of speech recognition workflow; speech recognition frameworks generate forced alignments of the audio and then use these as training examples for the machine learning method underneath. Forced aligners only need the first part of the process since they assume the model already exists. FAVE (Rosenfelder et al., 2011) and Munich AUtomatic Segmentation system (MAUS) (Kisler et al., 2017) work like this.

In contrast, Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) allows the researcher to train a new recognizer for a new language or adapt an existing one for a new corpus. However, this does require data, recommendations being at least one hour in the optimal case (Johnson et al., 2018). In addition, if the researcher is working with a language with no speech recognizer, there usually is no pronunciation lexicon either. This is an issue since most aligners work with the conventional speech recognition framework of using a lexicon to combine the orthography with the acoustic models performing the aligning. A recent approach gaining popularity is using end-to-end framework to both recognize (Chan et al., 2015) and align speech (Li et al., 2022) without the use of lexicons. The other benefit of this approach is the ability to jointly train the whole model, which has shown great promise with very large datasets.

Cross-language forced alignment (CLFA) solves these issues because it uses a well-resourced source language to do the alignments without the researcher having to train a new acoustic model with insufficient data. However, this requires a model for a language that is related to the target language. The process is still challenging, so it would benefit from all advantages it can get.

In this paper, we examine how different types of linguistic knowledge affect cross-language forced alignment. We demonstrate the concepts first with a larger high quality Finnish corpus, and then with a real use

case and try to apply them to a current Komi documentation project. We focus on word-level alignment, as it is sufficient resolution for the corpus documentation task we are trying to improve, and much easier to produce for evaluation purposes than phone-level alignment.

2. Related Research

Forced aligners are based on popular speech recognition frameworks, such as HTK (Young et al., 2002) for FAVE and Prosodylab-aligner (Gorman et al., 2011). Some of the first research on CLFA is using these frameworks for Yoloxóchitl Mixtec, an Oto-Manguean language (DiCanio et al., 2013). Currently, the most popular forced alignment toolkit is the MFA (McAuliffe et al., 2017). Based on Kaldi (Povey et al., 2011), it uses its popular speech recognition pipeline from input features to model choice using Gaussian mixture models (GMM) to model the phonemes of the speech. It provides useful features such as data validation, speaker adaptive training (SAT), and fine-tuning on new data. Tang and Bennett (2019) use it to align cross-language speech by pooling the source and target data together before training the model. Another tool based on Kaldi is Gentle[1], and in comparison to MFA, it uses deep neural networks (DNN) instead of GMMs to model phonemes. Munich AUtomatic Segmentation system (MAUS) (Kisler et al., 2017) is another tool capable of forced alignment. Its framework is based on statistical expert systems and it has been used in forced alignment and cross-language forced alignment. Strunk et al. (2014) use a language-independent version of MAUS to align many under-resourced languages with good results. Jones et al. (2019) use MAUS to align English based Kriol, comparing Italian source language to the language-independent MAUS. Surprisingly they found that Italian performed better than the independent model, showing promise for large related language source models. Promising results have also been achieved by applying Connectionist temporal classification (CTC) algorithm to align speech (Kürzinger et al., 2020).

3. Experiments

We will evaluate the effects of two methods for utilizing expert knowledge to improve the accuracy of automatic alignments. First, we eliminate artificial micropauses of different durations from end-alignments. Second, we simplify the data before alignment by segmenting it. The performance metric we use is the percentage of aligner-created word-boundaries 10ms, 25ms, 50ms, and 100ms length from the correct gold label boundary. A better model will have a higher percentage in lower millisecond ranges. Our code and models are publicly available.[2]

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

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

Lang	Dataset	Length	Tokens
fin	Finnish	1h7m27s	6464
kpv	Recording 1 (R1)	2m45s	179
	Recording 2 (R2)	4m11s	259
	Recording 3 (R3)	4m45s	446
	Recording 4 (R4)	4m32s	344

Table 1: The length of the speech data and number of tokens, each adding two datapoints (start and end boundary). Datasets in Komi (kpv) represent recordings from 1950 representing four subvarieties of Ižma dialect.

3.1. Datasets

Komi materials used in this experiment have been archived into the Institute for the Languages of Finland (Kotus). They were initially recorded in the 1950s and transcribed in the 1960s (Itkonen, 1958; Stipa, 1962). The whole collection represents different Komi dialects, and in this sample, four localities of the Ižma dialect were included. The recordings are available for research purposes in the Tape Archive of the Finnish Language maintained by Kotus, and the whole collection will be published when ready in the Language Bank of Finland. The manual word-level alignment was created primarily to test different forced alignment systems, and the annotations also include more extensive utterance-level segmentation. In practice, the intended goal is to align transcriptions and audio recordings at a coarser level, which would make them comparable to different contemporary language documentation corpora. Since the Komi dataset is so small, we also experiment with a Finnish dataset (Vainio, 2001; Raitio et al., 2008) of read-speech from one speaker created for speech synthesis purposes. Details of the data can be seen in Table 1.

3.2. Models

We take the Finnish Kaldi ASR model created in (Mansikkaniemi et al., 2017) and used in (Leinonen et al., 2021) as a baseline (Base). In addition, we experiment with two other Kaldi-based Finnish ASR models: Donate Speech (DS) model from a Finnish crowdsourcing project called Lahjoita Puhetta that collected over 3600 hours of speech (Moisio et al., 2022), and Conversational (Conv) model from (Moisio, 2021). For the former, a 100h manually transcribed subset of the whole data was used for training. The sizes of the acoustic models are 36.5, 16.6 and 16.5 million parameters for Base, DS and Conv. While the baseline is considerable larger, the latter two have more modern speech recognition architecture utilizing Kaldi's most recent updates and trained with a larger variety of speech data.

However, all of them are fundamentally DNN-based acoustic models trained with the lattice-free maximum mutual information (LF-MMI) criterion. They

all also use the same acoustic features, 39 dimension Mel-frequency cepstral coefficients (MFCCs) and Cepstral mean and variance normalization (CMVN). For speaker adaption they employ Kaldi's i-vectors. Since these are conventional automatic speech recognition models, they need pronunciation dictionaries. We create them by assuming a direct grapheme-to-phoneme (G2P) mapping. For Finnish this is straightforward as the writing system has a clear phoneme-to-grapheme correspondence. For Komi this direct G2P assumption is also sensible so we use domain expertise to match a Komi letter to the closest Finnish phonetic equivalent. In cases where a single phoneme would be insufficient, we combine multiple Finnish phonemes to represent a single Komi phoneme.

We also experiment with a Wav2Vec2 (W2V2) model (Baevski et al., 2020) from the Lahjoita Puhetta project. The pre-trained model is based on VoxPopuli corpus (Wang et al., 2021), fine-tuned with the same subset as the DS model. We modify the code in (Hira, 2021), which is using the CTC segmentation method presented in (Kürzinger et al., 2020) to create the alignments. Wav2Vec2 maps the speech directly to text so no pronunciation dictionary is needed.

3.3. Removing Silence

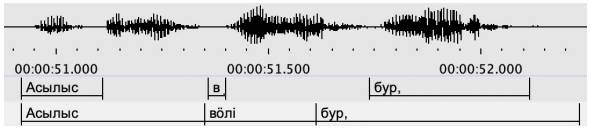

Figure 1: Difference in occurrences of small silences in automatic (upper) and manual (lower) alignments.

Figure 1 shows that automatic forced aligners leave short silences between words, while human-produced annotations for continuous speech, as in this example, have word boundaries with no silences. Current speech aligners' original purpose is to produce training material for speech recognition training. This material is optimized to contain all necessary information to differentiate between the tokens used, not to be optimal for corpus documentation in linguistic settings. This optimization can create micro-pauses as a side effect. However, to be a valuable tool for linguists, the results should serve their needs.

Therefore we experiment on the optimal way to remove the silences, testing where to split the silence before merging it to the surrounding words, extending the word boundary markings to cover the silence. We try three different values for a duration below which we consider a silence to be a splittable micro-pause. For convenience we use the same values as for the cutoffs used for errors, excluding 10ms, as that is our model's minimum resolution and therefore splitting it would be ineffective. We also compare three types of splitting: a start split merges the silence to the next word, middle divides it evenly between the words, and end split extends the word boundary of the first word to contain the silence.

3.4. Segmenting Audio

Another possible solution to make the alignment task simpler is to use segmented audio. Longer segments are challenging since errors from the beginning can propagate. Here we compare aligning the four Komi audios in full to shorter sentence-long segments of the same audio.

4. Results

Model	<10	<25	<50	<100
Base	0.21	0.55	0.84	0.98
DS	0.22	0.62	0.94	1.00
Conv	0.29	0.67	0.93	1.00
W2V2	0.12	0.29	0.59	0.91

Table 2: Accuracies of different Finnish models with the Finnish data. Percentages describe the amount of alignment deviations below the 10, 25, 50 and 100 millisecond cutoff values.

As we can see from Table 2, there was room for improvement in Finnish forced alignment. While the DS model is marginally better than Base in shorter than 10ms errors, it has 7 percentage points more of its deviations below the 25ms cutoff, and over 94% of its errors are below 50ms. Depending on the purpose of the automatic alignments, it might not need any manual correction. The Conv model is even better, with almost a third of the mistakes being less than 10ms, the resolution of Kaldi's alignments. It is also 5 percentage points better than the DS in 25ms range, however slightly worse in 50ms. This even though the data of the DS model, one person speaking uninterrupted, might be a closer match to the domain of the test data. The Wav2vec2 approach performs the poorest, being basically one error range behind the other models.

Table 3 shows the results of the silence tests. When comparing results to those shown in Table 2 we see that every conventional model gains minor improvements from the start split, with other types either not changing the results or worsening them. The most significant gains are achieved for the DS model, with 2-3% in absolute terms and 9-5% relative. As mentioned earlier, these improvements are related to that the speech recognition models start the words later than human annotators. While splitting does not change the order of the models in terms of performance, it is a simple algorithm with a consistent improvement with start split and 50ms or less micro-pause duration.

With the Wav2Vec2 algorithm, we see even larger gains with splitting silences, with absolute increases of 3-14 percentage points and 24-31% relative. Here the best improvements are again with the start split; however,

Model	Type Duration	<10			<25			<50			<100		
		25ms	50ms	100ms	25ms	50ms	100ms	25ms	50ms	100ms	25ms	50ms	100ms
Base	start	0.22	0.22	0.22	0.56	0.56	0.56	0.86	0.86	0.86	0.98	0.98	0.98
	middle	0.21	0.21	0.21	0.55	0.54	0.54	0.84	0.84	0.84	0.98	0.98	0.98
	end	0.20	0.20	0.20	0.53	0.53	0.53	0.83	0.82	0.82	0.98	0.98	0.97
DS	start	0.24	0.24	0.24	0.64	0.65	0.65	0.95	0.95	0.95	1.00	1.00	0.99
	middle	0.23	0.22	0.22	0.62	0.62	0.61	0.95	0.95	0.94	1.00	1.00	1.00
	end	0.21	0.21	0.20	0.59	0.58	0.58	0.94	0.93	0.92	1.00	1.00	0.99
Conv	start	0.30	0.30	0.30	0.68	0.68	0.68	0.93	0.94	0.93	1.00	1.00	0.99
	middle	0.29	0.29	0.29	0.67	0.67	0.67	0.93	0.93	0.93	1.00	1.00	0.99
	end	0.28	0.28	0.28	0.67	0.66	0.66	0.93	0.93	0.92	1.00	1.00	0.99
W2V2	start	0.12	0.14	0.15	0.30	0.38	0.38	0.60	0.72	0.73	0.91	0.93	0.93
	middle	0.11	0.11	0.11	0.29	0.28	0.29	0.59	0.60	0.62	0.91	0.92	0.92
	end	0.11	0.09	0.08	0.28	0.20	0.19	0.58	0.47	0.45	0.91	0.89	0.88

Table 3: Different Finnish models with resulting alignments post-processed with different silence removal methods using the Finnish dataset. Type describes the split used, Duration tells below what length is considered a micro-pause.

the optimal micro-pause duration is longer than in conventional models, being the maximum distance measured in errors, 100ms.

When we combine the previous results with those of segmenting the Komi audio, we can see mixed results from Table 4. If the model recognizes the full audio at all, it gets better results than with the segmented pieces, but the segments allow every audio to be aligned: only failed cases are with complete audios. The splitting of silences less than 50ms helps only marginally. It may be the case that when the results are poor, a longer micro-pause duration needs to be considered, as with the Wav2Vec2 model. Interestingly, the superior model for Finnish performs worse here overall. The only cases where it surpasses the baseline are for the segmented Recordings 1 and 4, in both below 100ms errors. The baseline model seems more robust against cross-language speech recognition, something that cannot be tested on Finnish data. Unfortunately, this does not allow easy comparison of models before choosing the right one for a CLFA task.

5. Conclusion

In this paper, we evaluated expert knowledge on cross-language forced alignment in the form of segmenting audio and splitting inaccurate micro-pauses in resulting alignments.

We began experimenting with the splitting of silences due to feedback from linguists. We found that silence splitting performed well on language-specific alignment on excellent audio quality but less on cross-language tasks with poorer audio quality. Instead of finding a global parameter for this, it might be best to find ways to describe the results and allow users to set these values themselves while trying to generate metrics to warn of unsafe values.

As for segmenting audio, low-quality recordings can be aligned with segmented audio, resulting in poorer quality alignments, while the full length has a chance of failing. Segmentation helps in the case of challenging recordings, but the correct places to cut audio are not obvious. However, the speed of automatic forced alignment allows an iterative process to experiment with both methods to achieve the best results possible.

Data & Model		<10		<25		<50		<100	
		-	sil	-	sil	-	sil	-	sil
Base	R1#	0.16	0.16	0.29	0.30	0.49	0.49	0.62	0.62
	R1	0.27	0.27	0.37	0.37	0.51	0.51	0.62	0.62
	R2#	0.17	0.17	0.28	0.28	0.41	0.41	0.53	0.53
	R2	0.34	0.34	0.43	0.43	0.50	0.50	0.62	0.62
	R3#	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.03
	R3	-	-	-	-	-		-	-
	R4#	0.19	0.19	0.32	0.32	0.45	0.46	0.61	0.61
	R4	-	-	-	-	-	-	-	-
Conv	R1#	0.13	0.13	0.28	0.28	0.46	0.46	0.70	0.70
	R1	-	-	-	-	-	-	-	-
	R2#	0.14	0.14	0.29	0.30	0.46	0.46	0.62	0.62
	R2	-	-	-	-	-	-	-	-
	R3#	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.02
	R3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R4#	0.11	0.11	0.29	0.29	0.48	0.49	0.70	0.70
	R4	-	-	-	-	-	-	-	-

Table 4: Results with full Komi audios compared to segmented audios (#). Without splitting (-), or 50ms (sil). Dash (-) in results represents failed alignment run.

6. Acknowledgements

We acknowledge the computational resources provided by both the Aalto Science-IT project and CSC – IT Center for Science, Finland. 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).

7. References

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations.

Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O. (2015). Listen, attend and spell.

DiCanio, C., Nam, H., Whalen, D. H., Timothy Bunnell, H., Amith, J. D., and García, R. C. (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.

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

Hira, M. (2021). Forced alignment with wav2vec2. `https://github.com/pytorch/audio/blob/main/examples/tutorials/forced_alignment_tutorial.py`.

Itkonen, E. (1958). Komin tasavallan kielitieteeseen tutustumassa. *Virittäjä*, 62(1):66–66.

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

Jones, C., Li, W., Almeida, A., and German, A. (2019). Evaluating cross-linguistic forced alignment of conversational data in north australian kriol, an underresourced language. *Language Documentation and Conservation*, pages 281–299.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326 – 347.

Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T., and Rigoll, G. (2020). CTC-segmentation of large corpora for german end-to-end speech recognition. In *Speech and Computer*, pages 267–278. Springer International Publishing.

Leinonen, J., Virpioja, S., and Kurimo, M. (2021). Grapheme-based cross-language forced alignment: Results with uralic languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 345–350.

Li, J., Meng, Y., Wu, Z., Meng, H., Tian, Q., Wang, Y., and Wang, Y. (2022). Neufa: Neural network based end-to-end forced alignment with bidirectional attention mechanism.

Mansikkaniemi, A., Smit, P., Kurimo, M., et al. (2017). Automatic construction of the Finnish parliament speech corpus. In *INTERSPEECH 2017–18th Annual Conference of the International Speech Communication Association*.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Moisio, A., Porjazovski, D., Rouhe, A., Getman, Y., Virkkunen, A., Grósz, T., Lindén, K., and Kurimo, M. (2022). Lahjoita puhetta – a large-scale corpus of spoken finnish with some benchmarks.

Moisio, A. (2021). Speech recognition for conversational finnish. Master's thesis, Aalto University School of Electrical Engineering, Espoo.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., and Alku, P. (2008). Hmm-based finnish text-to-speech system utilizing glottal inverse filtering. In *Ninth Annual Conference of the International Speech Communication Association*.

Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). Fave (forced alignment and vowel extraction) program suite. http://fave. ling. upenn. edu.

Stipa, G. J. (1962). Käynti syrjäänien tieteen tyyssijassa. *Virittäjä*, 66(1):61–68.

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

Tang, K. and Bennett, R. (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.

Vainio, M. (2001). Artificial neural network based prosody models for finnish text-to-speech synthesis.

Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. (2021). Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2002). The htk book. *Cambridge university engineering department*, 3(175):12.