TRAIN&ALIGN: A NEW ONLINE TOOL FOR AUTOMATIC PHONETIC ALIGNMENT

Sandrine Brognaux[1,2], Sophie Roekhaut[1], Thomas Drugman[3], Richard Beaufort[4]

[1] CENTAL - Université catholique de Louvain, Belgium
[2] ICTEAM - Université catholique de Louvain, Belgium
[3] TCTS Lab - Université de Mons, Belgium
[4] Nuance Communications, Inc.*

ABSTRACT

Several automatic phonetic alignment tools have been proposed in the literature. They usually rely on pre-trained speaker-independent models to align new corpora. Their drawback is that they cover a very limited number of languages and might not perform properly for different speaking styles. This paper presents a new tool for automatic phonetic alignment available online. Its specificity is that it trains the model directly on the corpus to align, which makes it applicable to any language and speaking style. Experiments on three corpora show that it provides results comparable to other existing tools. It also allows the tuning of some training parameters. The use of tied-state triphones, for example, shows further improvement of about 1.5% for a 20 ms threshold. A manually-aligned part of the corpus can also be used as bootstrap to improve the model quality. Alignment rates were found to significantly increase, up to 20%, using only 30 seconds of bootstrapping data.

Index Terms: Phonetic Alignment, HMM, Annotation, Corpus

1. INTRODUCTION

Large speech corpora play a key role in both linguistic research and speech technologies. A particularity of these corpora is that the sound cannot be studied alone. Its orthographic and phonetic transcriptions are generally required. The phones, in particular, must be time-aligned with the sound. Annotation tools like Praat [1], WaveSurfer [2], etc. allow defining several levels of annotation that can be determined by the user. They offer the possibility to manually align the sound with these annotations. However, manually aligning large corpora has two major drawbacks. First, this process is time-consuming: from 130 [3] to 800 [4] times real-time. For corpora of several hours as used for speech synthesis and speech recognition, this is economically unfeasible. Secondly, it requires trained language experts. Indeed, consistency should be as high as possible. This is even more difficult if several annotators align the same corpus. For example, it is shown in [5] that vowel-to-vowel boundaries or liquid-to-vowel boundaries are especially prone to alignment divergences between human annotators. This can be explained by the fact that the boundary is rather gradual for these phone sequences.

To solve these issues, automatic alignment tools have been developed. These tools rely on HMM-based alignment methods which are similar to speech recognition approaches. Acoustic models of each phoneme or group of phonemes are trained on a corpus. These models are then used to align a corpus with its phonetic transcription.

Two types of tools can be distinguished regarding the simplicity of their interface and the potential control of some parameters they offer to the user:

User-friendly tools with a graphical interface (e.g. EasyAlign [6] and SPPAS [7]) or several general methods that can be called in command lines (e.g. P2FA [8]). These tools present three major flaws. First, they provide the user with speaker-independent models. However, these models are supplied for a very limited number of languages only. Table 1 shows the languages covered by the various available tools. Many widespread languages like Russian or German are not covered, not to mention most African and Asian languages. Besides, they grant access to the alignment stage only. It is impossible for the user to train new models, e.g. for a non-covered language. The issue also applies to different language varieties. For example, aligning Belgian or Canadian French on the basis of a model trained on another variety, e.g. standard French, is unlikely to produce an accurate alignment. A second issue is that training parameters cannot be tuned. A pre-existent model is provided to the user and cannot be modified, with respect to the corpus to align. The third issue concerns the generalization abilities of the provided models. Ideally, they should be generic enough to produce high-quality alignment of different speaking styles: neutral speech, expressive speech, spontaneous speech, etc. Corpora with various emotions or phonostyles (sports comments, political speech, etc.) can have a significantly different acoustic variation that will result in a low-quality alignment. Indeed, the model provided to the user is strongly related to the corpus used for the training. It also means that, if some phonemes were rare or mis-represented in the training corpus, they will be prone to alignment errors. This issue was pointed out in [9] where semi-vowels were found to be badly aligned because of a possible under-representation of these phonemes in the training corpus.

Table 1. Languages covered by various existing automatic phonetic alignment tools based on speaker-independent models

Tool	Language
EasyAlign	French, Spanish, Portuguese, Taiwan Min
SPPAS	French, English, Italian, Chinese
P2FA	American English

HMM-based recognition toolkits like HTK [10] or Julius [11]. These toolkits are used by the previously-mentioned tools. Their advantage is that they offer methods to train new models. However, they also have several drawbacks. First, their use requires programming skills which most linguists do not have. The methods

* The tool was developed while Richard Beaufort was still working at the CENTAL (Université catholique de Louvain, Belgium)

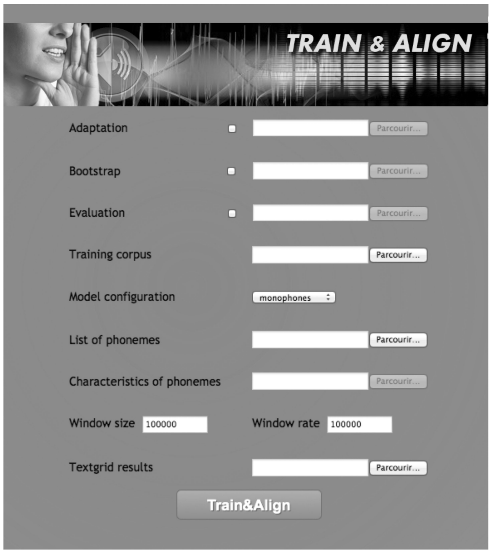
proposed by these toolkits should be integrated in a broader script that provides the expected input files. Besides, mastering such toolkits is time-consuming. As a result, most users stick to the basic version of the various methods and do not attempt to tune the training and alignment parameters. Not exploiting the full capacity of the methods obviously induces a lower quality of the resulting alignments.

The proposed Train&Align tool alleviates the aforementioned issues. Like EasyAlign, it offers a user-friendly graphical interface implementing HTK methods. Conversely to EasyAlign which only works on Windows O.S., our tool is proposed as a web service that can be accessed from any platform. It also offers the possibility to train new models. Its specificity is that the acoustic models are trained directly on the corpus to align, no pre-existent model being required. This means that it can be used to align any new language or speaking style. Finally, our tool also implements various training options that allow improving the quality of the alignment. The configuration of the models can be modified to take into account the phonetic context. With the bootstrapping option, a manually-aligned part of the corpus can also be exploited to improve the quality of the model.

The objective of this paper is to present Train&Align, a new online automatic phonetic alignment tool which can be used for any language and speaking style. It also shows a comparison of its performance with other available tools. The paper is organized as follows. Section 2 gives a description of Train&Align. Section 3 presents the protocol established to compare the performance of Train&Align with what is obtained using other available tools. This comparison is given in Section 4. Finally Section 5 concludes and discusses further improvements.

2. OVERVIEW OF TRAIN&ALIGN

Train&Align is a new online automatic phonetic alignment tool. Its particularity is that it does not rely on pre-existent models of each language. The models are directly trained on the corpus to align. As a result, it can be used to align any language or speaking style. Besides, it will be shown in Section 4 that it performs comparably or better than most existing tools. Finally, it should be noted that it is completely compatible with Praat formats, as it both accepts TextGrids as input and generates the aligned TextGrids as output. The Train&Align graphical interface is shown in Figure 1.



Adaptation □ Parcourir...

Bootstrap □ Parcourir...

Evaluation □ Parcourir...

Training corpus Parcourir...

Model configuration monophones

List of phonemes Parcourir...

Characteristics of phonemes Parcourir...

Window size 100000 Window rate 100000

Textgrid results Parcourir...

Fig. 1. Train&Align graphical interface

In a first stage, the entire (unaligned) corpus to align and its phonetic transcription are used to train a new language model. The speech is sampled at 16Khz. The parameters of the models are 12 Mel Frequency Cepstral Coefficients (MFCC) and their first and second derivatives. The phonemic models are five-state monophones. It also implements a silence model. The acoustic models are initialized with the so-called 'flat start' function. This means that, initially, the sound is aligned uniformly with the phonetic transcription. Then, the parameters are iteratively adapted with the Baum-Welsh algorithm [10]. In a second stage, these models are used to align the training corpus itself. HTK is used for both the training and the alignment.

Two main options are also offered by Train&Align. They make it possible to modify some training parameters. First, the **model configuration** can be modified. Each phoneme can be represented by one model, called a monophone (Table 2 (1)). However, the models can also be associated with the contextual phonemes on the left and on the right (Table 2 (2)). They are called triphones and model the coarticulation phase. Their use, however, can be problematic: a large corpus is necessary to properly model each triphone (whose number results from all possible associations of 3 consecutive phonemes). Besides, the augmentation in the number of models also results in an increase in processing time. To overcome this problem, a solution is offered to the user, i.e. the use tied-state triphones. The phonetic context of each phone is no longer modeled in terms of phonemes but in terms of classes (Table 2 (3)). Classes are generally articulatory characteristics that should be determined beforehand.

Table 2. Different model configurations

	Models	Examples
(1)	Monophones	[a]; [u]
(2)	Triphones	[b-a+f]; [p-a+v]; [j-u+z]; [w-u+Z]
(3)	Tied-State Triphones	[occlusive-a+fricative]; [semi-vowel-u+fricative]

The second option is the **bootstrapping**. If a (even small) portion of the corpus has been manually aligned, it can be used to improve the quality of the model, and hence the alignment. Our tool exploits bootstrapping methods offered by HTK [10]. Instead of a 'flat start' initialization, the manually-aligned corpus helps to improve the initialization stage. An iterative procedure determines the value of the parameters. In a first stage, the training data is uniformly segmented. Each model is matched with the corresponding data segments and the means and variances are estimated. In the second and successive stages, this segmentation is replaced by a Viterbi alignment [10]. The initial parameter values computed by Viterbi are then further re-estimated by a Baum-Welch procedure.

An additional option is currently under development. It allows the user to handle with adaptation techniques. This is particularly helpful when the corpus to align is made of various speakers or speaking styles. The models are then trained on the entire corpus and adapted to each speaker or speaking style to align that specific part of the corpus. The parameters of the models can then be modified so as to better match idiosyncrasies. Ongoing research on a corpus with different geo-linguistic varieties of French shows promising results.

It should be noted that the various options can be combined, e.g. the use of tied-state triphones with bootstrap.

Practically speaking, the use of Train&Align is rather straightforward. Audio files (in *.wav format) and their corresponding phonetic transcription should be provided to the tool (*Training corpus* in Figure 1). They will be used to train the models which will align the corpus itself. The aligned files (in a TextGrid format) will be automatically saved in a directory chosen by the user (*TextGrid results*). A comprehensive list of the phonemes used in the transcriptions should also be provided (*List of phonemes*). When using tied-state triphones, a file containing the articulatory characteristics of each phoneme must be given (*Characteristics of phonemes*).

Another advantage of Train&Align is that it makes it possible to evaluate the quality of the alignment produced. If a manually-aligned part of the corpus is provided, it can be compared to the automatic alignment to compute alignment rates (*Evaluation*), as described in Section 3. This file should be provided in a TextGrid format. The bootstrapping option comes with a specific evaluation process. Because manually-aligned data is required for both the bootstrap and the evaluation, a five-fold cross-validation is performed. This means that 4/5 of the bootstrapping dataset is used for the bootstrap and 1/5 is exploited for the evaluation, this operation being repeated 5 times such that all the annotated data is used for the evaluation. Train&Align then provides the average alignment rates. This specific evaluation option is automatically applied if an evaluation of Train&Align is asked in addition to the bootstrapping option.

3. EXPERIMENTAL PROTOCOL

To assess the performance of Train&Align, three corpora were used:

- A neutral read corpus in French used in the LiONS unit-selection synthesis [12]. It consists of 510 speech files for a total duration of 110 minutes. The corpus is phonetized and phonetically manually aligned.
- The Woggle corpus (US English) [13]. It is made of expressive read speech with 5 emotions (sad, happy, angry, afraid and neutral), uttered by five female speakers. The corpus contains 1068 files for a total duration of 51 minutes. The entire corpus has been phonetically manually aligned. Its particularity is its high degree of variability.
- A corpus in Kirundi[1] developed at the Valibel research center in the Université catholique de Louvain. It consists of 16 files for a total duration of 16 minutes. It is made of different speaking styles (read speech, guided conversation and narration) uttered by four different male speakers. Only one file of about 30 seconds was manually aligned and used for the assessment of the Train&Align. This corpus is also characterized by a rather high degree of variability.

In the following sections, Train&Align is compared to three available automatic alignment tools: P2FA (for English), SPPAS (for French and English) and EasyAlign (for French). The models that these tools provide have been trained on different corpora, varying in style, speaking style, and annotation. They also make use of various model configurations (monophones or tied-state triphones). A brief description of the models is given in Table 3.

It is worth noting that:

- Conversely to EasyAlign, P2FA and Train&Align, SPPAS uses Julius instead of HTK, for license matters. However, Julius skips silences at alignment time. Therefore, SPPAS processes silences in a previous stage that is independent from the alignment and based on the signal only. The segments between the silences are then aligned separately, with their supposedly corresponding transcription. However, the silence detection is sometimes erroneous. The alignment phase tries then to align a sound with a transcription which does not correspond to it. To avoid such errors, only files in which the silences were detected at the right position were kept for comparison. It does not mean that the length of the silences was correct but only that they were found at the right position.
- P2FA English model takes the level of stress into consideration, for the vocalic phonemes. It distinguishes between three levels: primary, secondary and no stress. This means that each vowel is associated with three different models. To exploit the full potential of the tool, the transcriptions were stress-annotated for the comparison with P2FA.

Table 3. Description of the models provided by available alignment tools

Model	Training corpus	Model Configuration
EasyAlign Fr	35 minutes of not aligned multi-speaker speech	monophones
SPPAS Fr	8 hours of not aligned conversational and read speech	tied-state triphones
SPPAS En	85 hours of not aligned read speech	tied-state triphones
P2FA En	25.5 hours of manually word-aligned speech from the Supreme Court of the US	monophones

In our experiments, all tools were provided with the correct phonetic transcription. To evaluate their performance, the automatic alignment was compared with the manual alignment. As a standard metrics of alignment accuracy, we used the ‘boundary-based’ measure [9]. The performance is measured as the percentage of boundaries which are similar in both alignments, with a certain tolerance threshold. In other words, we consider the proportion of alignment boundaries for which the timing error is lower than a threshold varying from 10 to 40 ms, by steps of 10 ms.

For an insightful interpretation of the results, a few benchmarks should be considered. It is well-known that alignments between various human annotators are often diverging. Above a 20 ms threshold, however, the agreement becomes fairly high. Using this threshold, the study in [6] reported inter-annotator rates of about 81% and 79% for the alignment of a French and of an English corpus, respectively. Rates between 88% and 95 % were obtained in [14] on an Italian corpus. It is also worth wondering how the alignment rates affect the quality of speech synthesis. [15] investigated this question for HMM-based alignment. It was shown that unit-selection based synthetic speech produced from a corpus aligned with a 92% rate with a 20 ms threshold was perceived as nearly as good as speech based on a manually-aligned corpus.

[1] Kirundi is a Bantu language with about 8 million speakers. It is a tonal language essentially spoken in Burundi.

4. EVALUATION OF TRAIN&ALIGN IN COMPARISON WITH OTHER AVAILABLE TOOLS

This section provides a comprehensive assessment of the performance of Train&Align in comparison with the other available user-friendly alignment tools presented in Section 3. Results obtained with the basic version of Train&Align, *i.e.* by using a model trained on the corpus to align, are described in Section 4.1. Section 4.2 investigates the improvement achieved by considering the phonetic context. It evaluates the performance of the alignment with monophones, triphones and tied-state triphones. Finally, Section 4.3 shows the alignment rates obtained when using some manually-aligned part of the corpus to improve the model, with the bootstrap option. The impact of the bootstrap size is also studied.

4.1. Basic Version

This section compares the basic version of Train&Align (T&A) with EasyAlign, SPPAS and P2FA. All tools do not provide models for both English and French. The alignment of the corpus in French with Train&Align is compared with SPPAS and EasyAlign. The alignment of the corpus in English is assessed with Train&Align, SPPAS and P2FA. As previously mentioned, no models are proposed by the available tools for African languages. We align therefore the corpus in Kirundi only with Train&Align. The alignment rates obtained by the various tools are shown in Table 4. It should be noted that the corpora in French and in English have both been entirely manually aligned. Therefore, we train here the model on the entire (unaligned corpus) and evaluate its performance by aligning the entire training corpus. For the corpus in Kirundi, however, only 30 seconds were manually-annotated. The models are thus trained on the entire corpus and evaluated on these 30 seconds only.

Table 4. Alignment accuracy obtained with various alignment tools on the three speech corpora

Model	Correct <10 ms	Correct <20 ms	Correct <30 ms	Correct <40 ms
Corpus in French				
SPPAS	45.73%	69.65%	81.02%	88.07%
EasyAlign	54.28%	80.83%	90.45%	94.45%
T&A	60.67%	84.55%	92.88%	96.69%
Corpus in English				
SPPAS	10.88%	27.09%	50.84%	72.34%
T&A	43.44%	63.87%	78.28%	87.17%
P2FA	46.86%	69.6%	81.66%	87.98%
Corpus in Kirundi				
T&A	35.62%	65.57%	79.04%	88.32%

For the corpus in French, interestingly, Train&Align is observed to clearly outperform SPPAS and EasyAlign models across all measures. The gain compared to SPPAS goes up to 15% with a 20 ms tolerance threshold. The overall rate is fairly high. It is comparable to inter-annotator rates reported in [6].

For the corpus in English, only P2FA provides results better than Train&Align. The explanation for this may be two-fold. First, the English model used by P2FA was trained on 25.5 hours of manually-word aligned speech. To provide such results for every language, it should be repeated on a very large number of corpora. This is economically impracticable, especially to cover all (even rare) languages. However, it is bound to produce better results. It is striking to notice that Train&Align offers slightly inferior but still comparable results while using no preexistent model trained on aligned data. The second explanation for the higher alignment rate of P2FA results from the use of acoustic models associated with various levels of stress. The Woggle corpus being an expressive corpus, it contains many emphatic stresses. It is well-known that these stresses usually fall on the same position as lexical stresses. This could also explain the better achievements of P2FA. The overall low alignment rates for the corpus in English are clearly due to its high acoustic variability.

Finally, it can be noticed that alignment rates obtained on the corpus in Kirundi are rather close to those of the corpus in English. This seems logical since the corpus in Kirundi is also characterized by a high degree of variability. The results, however, should be further confirmed by evaluating the alignment on a larger manually-aligned corpus. As previously mentioned, it is tested here on 30 seconds only, uttered by a single male speaker who could possibly display specific idiosyncrasies.

On the whole, it should be noted that Train&Align provides alignment rates better or comparable to most other alignment tools. In addition, it has the advantage of exhibiting the ability to be applied to any language, as shown here for Kirundi.

A particularity of Train&Align is that the acoustic models are directly trained on the corpus to align. The size of the corpus clearly plays a role in the model quality. It can therefore be wondered whether the rather high alignment rates obtained for the corpus in French are due to the size of the corpus, *i.e.* more than 110 minutes of speech. Indeed, a large number of occurrences for each phoneme is here used during the training stage, which is not possible for all databases. To answer this question, the influence of the corpus size on the alignment performance is displayed in Figure 2. It can be observed that the quality remains rather stable up to a two-minute corpus, beyond which the alignment performance rapidly degrades. This rather low limit is due to the low degree of variability of the corpus, made of neutral speech uttered by a single speaker. Experiments on the expressive corpus in English showed a comparable decrease, occurring between 30 and 15 minutes.

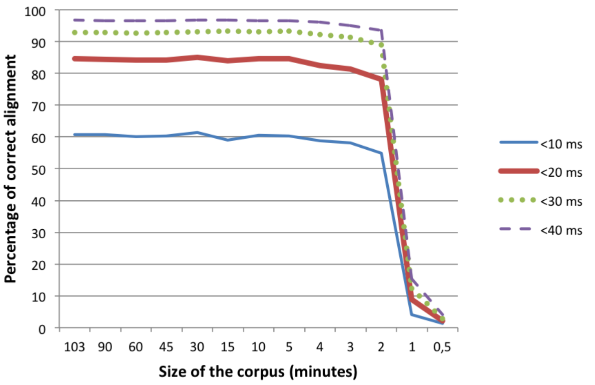


Fig. 2. Quality of the alignment with monophones as a function of the size of the corpus in French

A last question to investigate is the size of the speech files provided to Train&Align. Most existing alignment tools rely on an initial segmentation stage to prevent large audio file from being aligned [6, 7]. To better alleviate this issue, Train&Align makes use of pruning methods to reduce the processing time for longer files. Experiments with Train&Align showed that files of several minutes can be used without needing a segmentation beforehand. Large files of more than 10 minutes, however, should be cut to avoid very long processing time.

4.2. Consideration of the Phonetic Context

As previously mentioned, Train&Align offers several options. A first option proposes to consider a larger phonetic context. Triphones (possibly tied-state) can be used instead of monophones. The alignment rates with the three model configurations for the corpus in French are shown in Table 5.

Table 5. Alignment accuracy of the corpus in French with Train&Align using monophones, triphones and tied-state triphones

	Correct <10 ms	Correct <20 ms	Correct <30 ms	Correct <40 ms
T&A-mono	60.67%	**84.55%**	92.88%	96.69%
T&A-tri	62.69%	**85.44%**	92.55%	96.53%
T&A-tied	63.63%	**86.04%**	92.91%	96.62%

The results indicate that considering a larger phonetic context helps in modeling the language. For the neutral corpus in French, the use of tied-state triphones should clearly be recommended as it improves the overall quality of the alignment. Further experiments showed that, as expected, the processing time decreases with the use of tied-state triphones instead of triphones, while providing as good or slightly better results. A similar experiment on the corpus in English also showed increasing rates for thresholds higher than 30 ms. However, the results for lower thresholds slightly decreased. This can be explained by the higher variability of this latter corpus. In that respect, the phonetic context might not be the most relevant feature to take into account. The acted emotion or the position of the emphatic stresses could play a more significant role in the acoustic variation. Therefore, the option should be carefully chosen according to the corpus to align. If a small part of the corpus has been manually-aligned, it can be used to assess the quality of the models with monophones, triphones and tied-state triphones. This can be done thanks to the evaluation process proposed by Train&Align.

4.3. Bootstrapping

Another option offered by Train&Align is the bootstrapping. A manually-aligned part of the corpus can be exploited to improve the quality of the acoustic models, and hence of the alignment. For the experiments presented in this section, the specific evaluation process proposed by Train&Align, i.e. the five-fold cross-validation, was not used as such because the size of the evaluation corpus would clearly have influenced the results. When bootstrapping on 30 seconds, the model would have been tested on 6 seconds only, which is not representative of the quality of the alignment. Therefore, the alignment quality was always evaluated when aligning the same 20 minutes. Obviously, the files used for the evaluation were not present in the bootstrapping dataset.

Experiments (with monophone models) were first carried out on the corpus in English which was badly aligned with the basic version of Train&Align (in Section 4.1). As it is obvious that the size of the bootstrapping dataset influences the quality of the model, we investigated its impact. The objective is to reach a tradeoff between the size of the bootstrap and the alignment quality. Therefore, we assessed the performance of Train&Align with various sizes of bootstrap. The performance with different bootstrap sizes is shown in Figure 3. A similar curve was observed on the corpus in French.

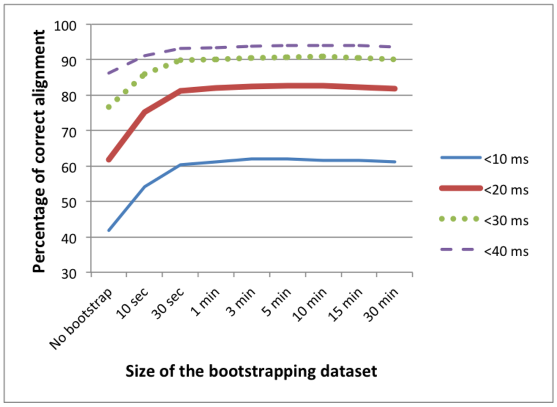


Fig. 3. Quality of the alignment of the corpus in English with various sizes of bootstrapping dataset

Results indicate that a bootstrapping dataset as small as 30 seconds already shows up good results, with alignment rates reaching above 80% for a 20 ms threshold. The rates stagnate with a dataset larger than 3 minutes. As previously mentioned, it is reported in [3] that the manual alignment process takes about 130 times real-time. For 30 seconds, about 1 hour would be required for a human annotator, which seems highly feasible. Besides, the annotator can rely on the initial alignment without bootstrap and only correct the erroneous boundaries. This should reduce the processing time.

We further analyzed the alignment rate gain obtained when using 30 seconds of bootstrap for both corpora in French and English. The results are shown in Table 6.

Table 6. Alignment accuracy of the corpora in French and English with 30 seconds of bootstrapping data

Model	Correct <10 ms	Correct <20 ms	Correct <30 ms	Correct <40 ms
Corpus in French				
T&A No bootstrap	59.73%	**84.67%**	92.93%	96.74%
T&A Bootstrap: 30 sec	63.06%	**87.26%**	94.66%	97.30%
Corpus in English				
T&A No bootstrap	41.81%	**61.75%**	76.61%	86.24%
T&A Bootstrap: 30 sec	60.39%	**81.18%**	89.86%	93.27%

It indicates that the use of a bootstrap significantly increases the alignment quality. On the corpus in French, about 2.5% are gained with a 20 ms threshold. The improvement is even more striking for the alignment of the corpus in English with an increase of nearly 20%. The fairly low results obtained with the version without bootstrap are compensated with 30 seconds of bootstrapping dataset. The alignment rates nearly reach those obtained on a neutral read corpus without bootstrap. This seems to indicate that bootstrapping is particularly advantageous for corpora with a high degree of variability. This remains to be proved on other corpora and will be experimented on the corpus in Kirundi as soon as a larger manually-aligned part of the corpus is available.

We might also wonder to what extent the alignment rates depend on the part of the corpus chosen to perform the bootstrap. The values in Figure 3 for bootstrap of less than 3 minutes are average values computed when randomly selecting 5 different bootstrapping datasets. For bootstrapping sizes as small as 30 seconds for the corpus in English, a standard deviation of only 1% was observed for the 20 ms threshold.

5. PERSPECTIVES AND CONCLUSION

Large speech corpora are required both in linguistic research and speech technologies. The sound generally needs to come along with a time-aligned phonetic transcription. For that purpose, several user-friendly tools have been developed (EasyAlign, SPPAS, P2FA, etc.). However, they do not grant access to the training phase and only provide the user with a set of speaker-independent models to align new corpora. A first issue is that these models are only available for a very limited number of languages. A second drawback is that they are closely related to the corpus used for the training. If the corpus to align is too different, the alignment quality may be low.

This paper presented a new online automatic alignment tool: Train&Align. Its characteristic is that it trains the acoustic models directly on the corpus to align. Therefore, no speaker-independent model of the language is needed. The advantage is that it is applicable to any language and speaking style. As an illustration, it was confirmed to be efficient in the paper on a corpus in Kirundi. Besides we have shown that the results obtained on several corpora (in French and English) are comparable to those of existing tools. The comparison for the French corpus showed improvement of more than 3% (20 ms threshold) with the best aligner, *i.e.* EasyAlign. On the corpus in English, Train&Align largely outperformed SPPAS. However, P2FA provided the best results with an increase of nearly 6% for an accuracy threshold of 20 ms but only 3% for 30 ms compared to Train&Align-tied. Another advantage of Train&Align is that it allows the user to tune several training parameters. The phonetic context can be considered by means of (tied-state) triphones. It was shown that the use of tied-state triphones to train the model to align a French neutral corpus further increases the results by 1.5% for a 20 ms threshold. If some part of the corpus has been manually aligned, it can also be used to improve the quality of the model by bootstrapping. Experiments have shown that as few as 30 seconds of manually-aligned speech can significantly improve the quality of the alignment. A dramatical increase of about 20% with a 20 ms threshold was observed when using 30 seconds of bootstrap for the English expressive corpus.

The tool is still currently in progress to add new options like adaptation techniques. Ongoing experiments show promising results. In the future, the objective would also be to divide the graphical interface into two separate sections: the training and the alignment. The user could then choose a model for the alignment. This model could be trained on the corpus to align or could be a preexisting model that the user trained on other corpora. Train&Align would then provide a useful platform for comparison of alignment rates between various models. At the long run and with user agreement, models for several languages could be made available on the website. This would facilitate the alignment of corpora that are too small and variable to train reliably a new model. The current version of Train&Align is now tested by several research centers and will be made available online by the time of the publication.

6. ACKNOWLEDGEMENTS

Sandrine Brognaux is supported by the “Fonds National de la Recherche Scientifique” (FNRS). The authors are very grateful to Gelase Nimbona, who recorded and annotated the corpus in Kirundi. The authors would also like to thank Brigitte Bigi and Jean-Philippe Goldman for their help when using their tools and for their enthusiasm regarding this study.

7. REFERENCES

[1] P. Boersma and D. Weenink. (2009, May) Praat: doing phonetics by computer (version 5.1.05) [computer program]. [Online]. Available: http://www.praat.org

[2] K. Sjölander, “Wavesurfer - an open-source speech tool,” in *Proc. of ICSLP*, 2000, pp. 464–467.

[3] H. Kawai and T. Toda, “An evaluation of automatic phone segmentation for concatenative speech synthesis,” in *Proc. of ICASSP 2004*, Montreal (Canada), 2004, pp. 677–680.

[4] F. Schiel and C. Draxler, “The production of speech corpora,” Bavarian Archive for Speech Signals, Tech. Rep., 2003.

[5] A. Ljolje, J. Hirschberg, and J. van Santen, “Automatic speech segmentation for concatenative inventory selection,” in *Second ESCA/IEEE Workshop on Speech Synthesis*, 1994, pp. 93–96.

[6] J.-P. Goldman, “Easyalign: an automatic phonetic alignment tool under Praat,” in *Proc. of Interspeech 2011*, 2011, pp. 3233–3236.

[7] B. Bigi and D. Hirst, “Speech phonetization alignment and syllabification (SPPAS): a tool for the automatic analysis of speech prosody,” in *Proc. of Speech Prosody 2012*, 2012.

[8] J. Yuan and M. Liberman, “Speaker identification on the SCOTUS corpus,” in *Proc. of Acoustics ’08*, 2008, pp. 5687–5690.

[9] J.-P. Goldman and S. Schwab, “Easyalign spanish: An (semi-) automatic segmentation tool under Praat,” in *5th CFE*, 2011.

[10] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3)*, Cambridge University, 1995.

[11] A. Lee, T. Kawahara, and K. Shikano, “Julius — an open source real-time large vocabulary recognition engine,” in *Proc. of Eurospeech 2001*, 2001, pp. 1691–1694.

[12] V. Colotte and R. Beaufort, “Linguistic features weighting for a text-to-speech system without prosody model,” in *Proc. of Interspeech 2005*, 2005, pp. 2549–2552.

[13] F. Dellaert, T. Polzin, and A. Waibel, “Recognizing emotion in speech,” in *Proc. of ICSLP*, 1996, pp. 1970–1973.

[14] P. Cosi, D. Falavigna, and M. Omologo, “A preliminary statistical evaluation of manual and automatic segmentation discrepancies,” in *Proc. of Eurospeech 1991*, 1991, pp. 693–696.

[15] J. Adell, A. Bonafonte, J. A. Gomez, and M. J. Castro, “Comparative study of automatic phone segmentation methods for TTS,” in *Proc. of ICASSP 2005*, 2005, pp. 309–312.