Automatic Phone Alignment

A Comparison between Speaker-Independent Models and Models Trained on the Corpus to Align

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

[1] ICTEAM - Université catholique de Louvain, Belgium
[2] CENTAL - Université catholique de Louvain, Belgium
[3] TCTS Lab - Université de Mons, Belgium
[4] Nuance Communications, Inc, Belgium*
sandrine.brognaux@uclouvain.be, sophie.roekhaut@uclouvain.be,
thomas.drugman@umons.be, richard.beaufort@nuance.com

Abstract. Several automatic phonetic alignment tools have been proposed in the literature. They generally use speaker-independent acoustic models of the language to align new corpora. The problem is that the range of provided models is limited. It does not cover all languages and speaking styles (spontaneous, expressive, etc.). This study investigates the possibility of directly training the statistical model on the corpus to align. The main advantage is that it is applicable to any language and speaking style. Moreover, comparisons indicate that it provides as good or better results than using speaker-independent models of the language. It shows that about 2 % are gained, with a 20 ms threshold, by using our method. Experiments were carried out on neutral and expressive corpora in French and English. The study also points out that even a small neutral corpus of a few minutes can be exploited to train a model that will provide high-quality alignment.

Keywords: Phonetics, HMM, Alignment, Corpora, Annotation

1 Introduction

Large speech corpora are required both in linguistic research and speech technologies. A characteristic of these corpora is that the sound cannot be studied alone. Most of the time, an orthographic and a phonetic transcription of the audio files are needed. The phonemes, in particular, should be synchronized with the sound. Indeed, the analysis of intonation, pronunciation, etc. requires to know the precise position of the phonetic temporal boundaries. Similarly, unit-selection and HMM-based speech synthesis rely on the segmentation of the sound in phones or diphones. The quality of the generated voice strongly depends on the alignment accuracy. The phonetic alignment, also called forced alignment, can be done manually. However, this process has two serious drawbacks. First, it is time-consuming: from 130 [1] to 800 times real-time [2]. For large

* The study was carried out while Richard Beaufort was still working at the CENTAL (Université catholique de Louvain, Belgium).

corpora of several hours, as used for speech synthesis or speech recognition, the resulting manual annotation time would become prohibitive, which is economically impracticable. Secondly, a language expert is required for the task. The alignment process is not trivial and needs to be done as consistently as possible. This is even more problematic if different human annotators are working on a same corpus.

To overcome these problems, automatic alignment tools such as EasyAlign [3], SPPAS [4] or P2FA [5] have been developed. They allow the alignment to be both consistent and reproducible at a very low cost. Most of these tools rely on the acoustic modeling of the language with Hidden Markov Models (HMM). During the training, the acoustic model of each phoneme or group of phonemes of the language is built. During the alignment phase, these models are used to align an audio file with its phonetic transcription. The process is very similar to speech recognition techniques except that the phonetic transcription is known.

Generally, the acoustic models are trained on large corpora with several speakers. They account for an overall realization of the language which is not specific to one speaker or speaking style. Most automatic alignment tools provide the user with such speaker-independent models which can be used to align new corpora. This method has several disadvantages. First, the number of languages covered by the provided models is limited. Therefore, some corpora cannot be aligned. Secondly, the performance of the model strongly depends on the agreement between the training corpus and the corpus to align. If they are too different, the alignment quality may be low.

A way to alleviate these two issues is to train the model directly on the corpus to align. It offers the advantage of applying to any language and any speaking style. Besides, training the models on the speaker to align was proven to be highly profitable in speech recognition [6].

The aim of this paper is to investigate the quality of the alignment produced with a model trained on the corpus to align. This is done in comparison with the use of available speaker-independent models provided by recent alignment tools. The paper is organized as follows. Section 2 proposes an overview of the state of the art. Section 3 provides a detailed description of our method based on a training on the target corpus to align. The experimental protocol designed to evaluate our results is stated in Section 4. Results of our experiments are then shown in Section 5, providing an assessment of the proposed approach as well as a comparative evaluation with state-of-the-art techniques based on speaker-independent models. Finally Section 6 concludes and discusses further works.

2 State of the Art

HMM-based phonetic alignment has been pointed at as the most reliable technique for automatic phonetic alignment [7, 8]. It relies on speech recognition paradigms. Most existing alignment tools and studies [3, 5, 4, 1, 9] are based on the HTK toolkit [10] or similar toolkits like Julius [11]. HTK offers an implementation of HMM and meth-

ods for speech recognition and forced alignment. Both the training and the alignment developed for the experiments in this paper make use of HTK.

With such toolkits, the number of models to train can generally be defined by the user. Usually, each phoneme is linked to one model, called a monophone model (Table 1 (1)). Three to five states represent the different stages of its realization: the transition and the stabilization phases. However, the models can also be associated with phonemes in context, regarding the phonemes on the left and on the right (Table 1 (2)). They are called triphones and allow modeling the coarticulation phase. Their use, however, can be problematic: a lot of data is required to offer a good representation of each triphone. Besides, the augmentation in the number of models also increases the processing time. A solution is the use of tied-state triphones: the phonetic context of each phone is no longer modeled in terms of phonemes but in terms of classes (Table 1 (3)). Classes are generally articulatory characteristics that should be defined beforehand.

Beside the phoneme models, two specific models can be added (see Fig. 1). A silence model ('sil') represents silent pauses. Conversely to other phoneme models, it allows a direct transition from the second to the fourth state and a backward skip from the fourth to the second. Silences can be indicated in the phonetic transcription. The second specific model is a short-pause model ('sp'), which is a one-state model. Its emitting state is tied to the centre state of the silence model. 'sp' models are automatically inserted between the words. It allows to detect a silence that was not mentioned in the phonetic transcription.

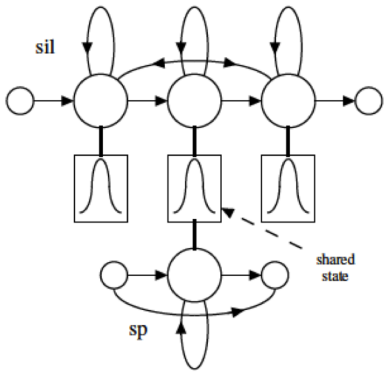

Fig. 1. Silence model and short-pause model [10]

Most research assessing the performance of HMM-based alignment use speaker-independent models [3, 4, 12, 13]. [13] offers an insightful comparison of the alignment rates when using various available speaker-independent models. The existing alignment tools (EasyAlign, SPPAS, P2FA, etc.) also provide the user with speaker-independent models, for several languages. These models can be used to align new corpora. The user has no access to the training stage and cannot train new models.

Table 1. 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]

A first drawback of such methods is that the number of available language models is limited. This means that many languages are not modeled and that the corresponding corpora cannot be aligned. The languages covered by some of the most widely-used tools (which will be presented in Section 4 and evaluated in Section 5.3) are shown in Table 2. It is worth noting that widespread languages, like Russian or German, are not covered by these tools.

A second flaw regards the quality of the provided models. They should be generic enough to produce high-quality alignment of different speech varieties or various speaking styles: neutral speech, spontaneous speech, expressive speech, etc. However, the model is strongly related to the corpus used for the training. Besides, if some phonemes were rare or mis-represented in the training corpus, these phonemes will be prone to alignment errors.

Table 2. Languages covered by various existing tools for automatic phonetic alignment

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

To alleviate these problems, we propose to train the model directly on the corpus to align. Few studies have evaluated the alignment quality obtained with such models. The performance of speaker-dependent models was analyzed in [1]. In this study, however, the model is trained on *aligned data* of one speaker and used to align some *other part* of a corpus of the same speaker. This improves the quality of the alignment because of a better agreement between the model and the corpus to align. However, it requires some part of the corpus to be manually-aligned. This is time-consuming, and hence costly.

The potential of training the model on the corpus to align was evaluated in [14] and [8]. It showed promising results, notably for its use for under-resourced languages [8]. However, it was not compared with the results obtained when aligning the same corpus with existing speaker-independent models. In [14], it was only tested on a corpus of 100 utterances and no claim was made about the minimum size required for the corpus to train a model. Results in [8] on African languages suggest that a small corpus of about 20 sentences would be enough. This remains to be proved on other Indo-European languages.

3 Our method: Train&Align

Our method works as follows. In a first stage, the entire (unaligned) corpus to align is used to train a new language model. Acoustic parameters are extracted from the sound files and modeled [5]. The phonemic models are five-state monophones. It implements both silence and short-pause models. In the second stage, these models are used to align the training corpus itself. For that matter, it makes a specific use of HTK methods for both the training and the alignment. The advantage of the method is that it can apply to any language or speaking style, as no pre-existing model is needed. Another benefit is that the training parameters can be modified. This method will be referred to as Train&Align-mono (**T&A-mono**) in the remainder of this paper.

The method is proposed with monophones but also with triphones (**T&A-tri**) and tied-state triphones (**T&A-tied**). In this latter approach, the phonetic context is defined in terms of classes. The list of the characteristics exploited in the method is shown in Table 3.

Table 3. Classes used to determine the context with tied-state triphones

Classes	Values
Type	Vowel/Consonant/Semi-vowel
Articulation place	Bilabial/Labiodental/Alveolar/Palatal/...
Articulation mode	Plosive/Constrictive/Fricative/Liquidly/...
Voicing	Voiced/Unvoiced

Considering the phonetic context is an advantage in our method. Indeed, the use of pre-existing speaker-independent models makes it harder to use triphones.

In pre-existing models, all the triphones of the language should be present. If the corpus to align contains new triphones, the alignment process fails. However, the phonetic context coverage of the training corpus usually differs from the coverage of the target corpus, even if the training corpus is rather large. Obviously, this problem does not arise when the model is trained on the corpus to align. For pre-existing models, a particularly large corpus would be required to model every triphone. A solution might be to assign average values to non-existing triphones. However, that could harm the quality of the model and hence, the alignment.

4 Experimental Protocol

For the experimentation, Train&Align is used to align two corpora :

1. A neutral French-speaking corpus used in the LiONS unit-selection synthesis [15]. It consists of 510 speech files that are phonetized and manually aligned. The total duration is around 110 minutes.

[5] The acoustic parameters are 12 Mel Frequency Cepstral Coefficients (MFCC) and their first and second derivatives.

2. The Woggle corpus [16], a corpus of American English. It contains expressive speech related to five emotions (angry, sad, happy, fear and neutral) uttered by five female speakers. It consists of 1,068 files for a total duration of 51 minutes. It was phonetized and manually-aligned by the first author of this paper. Its particularity is its high degree of variability.

The automatic alignment is evaluated in comparison with the manual alignment. The performance is measured as the percentage of boundaries that are similar in both alignments, with a certain tolerance threshold. In other words, accuracy metrics used in this work consider the proportion of alignment boundaries for which the timing error is lower than a threshold varying from 10 to 40 ms.

To allow an insightful interpretation of the performance, a few benchmarks should be considered. Large discrepancies are noticed between human-made alignments. Usually, 20 ms constitutes a limit above which the agreement rate is fairly high. Using this 20 ms threshold, [3] obtains agreement rates of about 81 % and 79 % for the alignment of a French and of an English corpus, respectively. On an Italian corpus, [17] find rates between 88 % and 95 %. It is also insightful to know which rate is sufficient for a speech corpus to be used for speech synthesis. In [7], it is 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.

In a further experiment, Train&Align is compared to the use of existing speaker-independent models. The models used for the comparison come from VoxForge [18] and from recent alignment tools (EasyAlign [3], SPPAS 1.4 [4] and P2FA [5]).

1. **EasyAlign** provides a model for French but not for English. Its French model was trained on "30 minutes of unaligned multi-speaker speech for which a verified phonetic transcription was provided" [3]. The model consists of monophones.
2. **VoxForge** only provides a model for English. We used the latest version (June 15, 2012). It was trained on nearly 100 hours of read speech that were automatically phonetically transcribed but not aligned. The model consists of tied-state triphones.
3. **SPPAS** provides models for both French and English. SPPAS French model was trained on 8 hours of phonetically transcribed but not aligned speech from the CID and the AixOxCorpus. CID contains conversational speech while AixOxCorpus is made of read speech. SPPAS English model is the model of July 2011 provided by VoxForge. It contains about 85 hours of multi-speaker read speech. The corpus was automatically phonetically transcribed but not aligned. The models consists of triphones.
4. **P2FA** only provides an English model. It was trained on 25.5 hours of manually word-aligned speech from the Scotus corpus. This corpus consists of oral arguments from the Supreme Court of the United States. The model is made of monophones.

In Section 5.3, the SPPAS and EasyAlign tools were used to align the corpus as the end user would have done. For VoxForge and P2FA, which do not provide a user-friendly graphical interface, the models were used with HTK. The models were all provided with the correct phonetic transcription.

5 Experiments

Experiments are divided into three evaluations. First, we evaluate in Section 5.1 the performance of Train&Align on the French and the English corpus. Secondly, the minimum size of the target corpus to use is investigated in Section 5.2. Finally, Section 5.3 provides a comparative evaluation between Train&Align, and the five state-of-the-art speaker-independent models presented in Section 4.

5.1 Assessment of Train&Align

The three versions of Train&Align (mono,tri and tied) were applied on the French and on the English corpus. The alignment rates are shown in Table 4.

Table 4. Alignment accuracy of the French-speaking corpus and the English-speaking corpus with Train&Align-mono, -tri, and -tied

	Correct $<$10 ms	Correct $<$20 ms	Correct $<$30 ms	Correct $<$40 ms
French-speaking corpus				
T&A-mono	58.25 %	82.56 %	91.75 %	95.91 %
T&A-tri	60.74 %	84.23 %	91.9 %	96.27 %
T&A-tied	61.58 %	84.59 %	92.22 %	96.43 %
English-speaking corpus				
T&A-mono	42.84 %	63.22 %	77.97 %	86.8 %
T&A-tri	42.26 %	62.8 %	78.43 %	87.92 %
T&A-tied	42.44 %	62.84 %	78.6 %	87.99 %

The results on the French-speaking corpus exceed 80% for a 20 ms threshold. It is rather close to the inter-annotator agreement rates reported in [3]. However, it yields some major errors ($>$30 ms tolerance) which should be manually corrected. We can assume that only a quick manual check should be enough to produce high-quality alignment. This would largely reduce the required processing time.

For the English-speaking corpus, the correct alignment rates are significantly lower. This is due to the high variability of the corpus which contains several speakers and emotions. Section 5.3 examines whether low results are also found when the alignment is performed with speaker-independent models of English.

The results indicate that considering a larger phonetic context helps in modeling the language. For both corpora, an increase in the alignment rates with a threshold of 30 ms or more is observed. For the neutral French-speaking corpus, the use of triphones should clearly be recommended as it improves the overall quality of the alignment. For the English-speaking corpus, however, the alignment rates for smaller tolerance thresholds decrease. This could be due to the high variability of the 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.

The results on the French-speaking corpus tend to be in line with [12]. They point out that the use of context-dependent models like triphones improves the alignment for small tolerance thresholds. In Table 4, the increase for 10 ms is clearly higher than for 40 ms, which indicates a clear increase in the precision of the alignment. Contrary to their study, we do not notice any degradation of the model with a tolerance of 20 ms. This might be due to the enormous size of their corpus, consisting of 1,037 speakers. The corpus may be big enough to ensure a very precise modeling of the phonemes that is partially damaged with the use of the phonetic context.

It is worth wondering whether these rather high alignment rates for the French-speaking corpus might be due to the size of the corpus to align. More than 100 minutes of speech are used to train the models. This provides a fair amount of occurrences for each phoneme. That question is now addressed in the Section 5.2.

5.2 Influence of the Size of the Corpus

The corpora that need to be aligned can be of a rather small size. This section investigates the minimal size of the corpus so as to build a high-quality model. This was studied on both test corpora. The total size of the corpus was gradually decreased and the alignment performance with Train&Align was assessed. The results for the French-speaking corpus are displayed in Fig. 2 (1). They show that the quality remains rather stable up to a two-minute corpus, beyond which the alignment performance rapidly degrades. This short duration is due to the low variability of the corpus consisting of read neutral speech. A similar test on the English-speaking corpus displayed a sooner decrease, between 30 and 15 minutes (see Fig. 2 (2)). On the whole, a few minutes of neutral speech seem to be enough to train and align a new corpus. This confirms the findings of [8] on African languages.

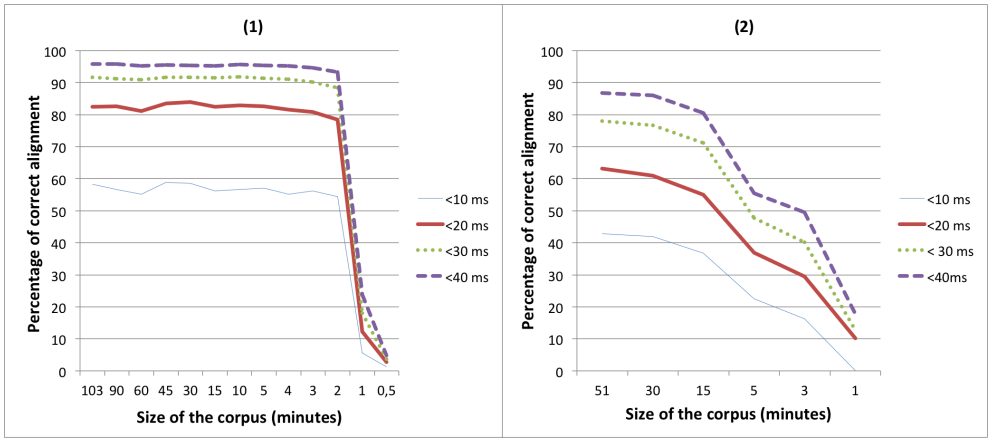

Fig. 2. Alignment rates with Train&Align-mono as a function of the size of the French-speaking corpus (1) and of the English-speaking corpus (2)

5.3 Comparison with Speaker-Independent Models

Train&Align is now compared to the five speaker-independent models presented in Section 4. It should be noted that :

1. Due to the conditions of distribution based on a GPL license, SPPAS uses Julius [11] and not HTK [10] to align the corpus. A disadvantage is that silences and short-pauses are skipped during the alignment stage. Silences are then processed separately, with the inter-pausal units (IPUs) segmentation tool. An orthographic transcription must be provided to the tool with a specific label for silences. Those silences are detected on the basis of the signal only, in a phase that is independent from the alignment. The segments between the silences are aligned separately, with their supposedly corresponding transcription. However, silences are sometimes erroneously assigned to the signal. This penalizes the quality of the alignment as the system tries to align a signal with a transcription that does not correspond to it. To avoid such errors, only sentences for which the position of the silences was correctly detected were kept for the evaluation with SPPAS. It does not mean that the detected length of the silences was correct, but only that they were found at the right position. Both P2FA and EasyAlign use HTK and provide a silence model. A silence model is also implemented in Train&Align.
2. P2FA model depends on the lexical stress level of the phoneme. Three levels are considered: no stress, primary and secondary stress. Each vocalic phoneme is associated with three models. To exploit the full capacity of the system, all the phonetic transcripts were stress-annotated when aligning with P2FA model.

All the alignment tools used for the comparison do not provide models for both English and French. Table 5 shows the alignment performance on the French-speaking corpus with SPPAS model, EasyAlign model and Train&Align. Interestingly, Train&Align is observed to clearly outperform SPPAS and EasyAlign models across all measures. The gain compared to SPPAS goes up to nearly 15 % with a 20 ms tolerance threshold. Besides, we know from Section 5.2 that the quality of the alignment remains stable up to a 2-minute corpus. It is striking to notice that training on 2 minutes of speech specific to the corpus to align provides better results than the use of a model trained on more than 8 hours of multi-speaker speech.

Table 5. Alignment accuracy of the French-speaking corpus with various models

Model	Correct <10 ms	Correct <20 ms	Correct <30 ms	Correct <40 ms
SPPAS	43.78 %	67.68 %	79.7 %	87.44 %
EasyAlign	54.18 %	80.7 %	90.27 %	94.28 %
T&A-mono	58.25 %	82.56 %	91.75 %	95.91 %
T&A-tied	61.58 %	84.59 %	92.22 %	96.43 %

Table 6 shows the alignment performance on the English-speaking corpus with SPPAS model, VoxForge model, P2FA model and Train&Align. A first observation that

can be highlighted is that VoxForge significantly outperforms SPPAS that uses an earlier VoxForge model. This can be explained by the fact that a silence model is included in VoxForge model and processed by HTK. Errors made by SPPAS IPUs segmentation are thus cancelled. A problem of VoxForge is that the model consists of triphones. The resulting flaw is that all triphones are not modeled and that some files cannot be aligned. Only 400 speech files out of the 1,068 files could be aligned. This problem was solved by SPPAS by adding unobserved triphones.

The improvement of the alignment quality with Train&Align compared to SPPAS and VoxForge is, here again, rather clear. The gain compared to VoxForge goes up to nearly 15 % with a 20 ms tolerance threshold. However, it turns out that P2FA gives the best results, in particular for thresholds lower than 30 ms. This is probably due to the corpus used for the training, *i.e.* several hours of word-aligned speech. This is bound to improve the quality of the model. P2FA also takes different levels of stresses into account. This might improve the alignment as expressive speech displays more emphatic stresses. These stresses usually fall on the same position as primary stresses. It is again striking to notice that training on the (unaligned) corpus to align produces results that are comparable or slightly inferior to those provided by a model trained on more than 25 hours of word-aligned speech. The overall low alignment rate is clearly due to the high acoustic variability of the Woggle corpus.

On the whole, it is worth noting that Train&Align offers nearly as good or even better alignment of the corpus than existing tools used for comparison. This shows evidence that the alignment does not need to rely on existing speaker-independent models. This means that unseen languages or speaking styles could be automatically aligned.

Table 6. Alignment accuracy of the English-speaking corpus with various models

Model	Correct <10 ms	Correct <20 ms	Correct <30 ms	Correct <40 ms
SPPAS	11.04 %	26.25 %	49.39 %	70.6 %
Voxforge	23.78 %	48.56 %	70.85 %	84.82 %
T&A-mono	42.84 %	63.22 %	77.97 %	86.8 %
T&A-tied	42.44 %	62.84 %	78.6 %	87.99 %
P2FA	44.92 %	68.11 %	79.78 %	86.35 %

6 Conclusion

To align speech sound files with their phonetic transcription, HMM-based alignment methods have been developed. For the alignment of new corpora, pre-existent speaker-independent models, as provided by EasyAlign, SPPAS or P2FA can be used. However these models are only available for a very limited number of languages. Furthermore, they may produce low-quality alignments when used to align a corpus that strongly differs from the corpus used for the training (neutral vs. expressive, read vs. spontaneous, etc.). A solution offered by this article is to use the target corpus, which needs to be

aligned, to train the acoustic model.

Several experiments showed that using a model trained on the target corpus yields nearly as good or even better results than using available speaker-independent models of the language. These available models were those provided by recent alignment tools: EasyAlign, SPPAS and P2FA, and the VoxForge English model. The improvement of our method was observed for neutral and expressive speech, as well as on both French and English corpora. Improvements in the alignment quality of about 2 % can be observed with our method with monophones (with a threshold of 20 ms). This can be explained by the fact that the model better captures the specificity of the target corpus. On the English-speaking expressive corpus, only P2FA outperforms our method, by about 5 % for 20 ms but only 2 % for 30 ms. This is due to their training corpus that consists of more than 25 hours of word-aligned speech. The minimum size of the corpus to use to obtain high-quality alignment was also investigated. On a neutral speech corpus, it was found that only 2 minutes were sufficient to train the model properly. However, expressive speech is more variable and around 15 minutes of speech are needed. On the whole, this study points out that even small-sized corpora can be aligned without the need for pre-existing models of the language. The advantage is that this implies that any corpus in any language could be aligned autonomously, without alignment quality loss. The method also allows modifying training parameters like the model configuration. It was shown that the use of triphones instead of monophones further increases the alignment rates by about 2 % for large neutral corpora.

Other modifications of the training, left unexplored in this study, might also be applied. If a portion of the corpus is manually aligned, it could be used to improve the quality of the model, with bootstrapping methods. Ongoing tests show very promising results, especially on corpora for which low initial alignment rates were obtained, e.g. on the expressive English-speaking corpus.

If the target corpus includes several speakers or speaking styles, adaptation methods could also be applied. The models would be trained on the entire corpus and then adapted to each speaker or speaking style to align that specific part of the corpus. Obviously, these adaptation techniques could also be applied to the speaker-independent models offered by SPPAS, EasyAlign, etc. to improve the agreement with the corpus to align. This, however, requires the use of existing models that are not available for every language. Conversely, the objective of this study was to show that a corpus could be aligned autonomously, without damaging the quality of the alignment.

The results shown in this paper should be further confirmed by tests on other languages and speaking styles. A study is in progress on the alignment of Kirundi, an African language, and on a French-speaking corpus with different phonostyles (radio, sports, etc.). As previously mentioned, most user-friendly automatic alignment tools (SPPAS, EasyAlign, etc.) do not grant access to the training phase: it is impossible for the user to train a new model on the corpus to align. The tool we developed allows improving the results by training new models. It also offers a solution for languages

for which no model is provided. This tool should be made available to the research community shortly.

Acknowledgments. Sandrine Brognaux is supported by the "Fonds National de la Recherche Scientifique" (FNRS). 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.

References

1. Kawai, H., Toda, T.: An evaluation of automatic phone segmentation for concatenative speech synthesis. In: Proc. of ICASSP 2004, Montreal (Canada) (2004) 677–680
2. Schiel, F., Draxler, C.: The production of speech corpora. Technical report, Bavarian Archive for Speech Signals (2003)
3. Goldman, J.P.: Easyalign: an automatic phonetic alignment tool under praat. In: Proc. of Interspeech 2011. (2011) 3233–3236
4. Bigi, B., Hirst, D.: Speech phonetization alignment and syllabification (sppas): a tool for the automatic analysis of speech prosody. In: Proc. of Speech Prosody 2012. (2012)
5. Yuan, J., Liberman, M.: Speaker identification on the scotus corpus. In: Proc. of Acoustics '08. (2008) 5687–5690
6. Leggetter, C., Woodland, P.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Computer Speech and Language **9**(2) (1995) 171–185
7. Adell, J., Bonafonte, A., Gomez, J.A., Castro, M.J.: Comparative study of automatic phone segmentation methods for tts. In: Proc. of ICASSP 2005. (2005) 309–312
8. van Niekerk, D., Barnard, E.: Phonetic alignement for speech synthesis in under-resourced languages. In: Proc. of Interspeech 2009, Brighton (2009) 880–883
9. Cangemi, F., Cutugno, F., Ludusan, B., Seppi, D., Van Compernolle, D.: Automatic speech segmentation for italian (assi) : Tools, models, evaluation and applications. In: Proc. of AISV, Lecce (Italy) (2011) 337–344
10. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3). Cambridge University. (1995)
11. Lee, A., Kawahara, T., Shikano, K.: Julius — an open source real-time large vocabulary recognition engine. In: Proc. of Eurospeech 2001. (2001) 1691–1694
12. Toledano, D., Gómez, L.: Hmms for automatic phonetic segmentation. In: Proc. of LREC. (2002)
13. Chen, L., Liu, Y., Harper, M., Maia, E., McRoy, S.: Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus. In: Proc. of LREC 2004. (2004) 759–762
14. Ljolje, A., Hirschberg, J., van Santen, J.: Automatic speech segmentation for concatenative inventory selection. In: Second ESCA/IEEE Workshop on Speech Synthesis. (1994) 93–96
15. Colotte, V., Beaufort, R.: Linguistic features weighting for a text-to-speech system without prosody model. In: Proc. of Interspeech 2005. (2005) 2549–2552
16. Dellaert, F., Polzin, T., Waibel, A.: Recognizing emotion in speech. In: Proc. of ICSLP. (1996) 1970–1973
17. Cosi, P., Falavigna, D., Omologo, M.: A preliminary statistical evaluation of manual and automatic segmentation discrepancies. In: Proc. of Eurospeech 1991. (1991) 693–696
18. MacLean, K.: Voxforge (http://www.voxforge.org) (2006-2012)