HMM-Based Speech Segmentation: Improvements of Fully Automatic Approaches

Sandrine Brognaux and Thomas Drugman

***Abstract*—Speech segmentation refers to the problem of determining the phoneme boundaries from an acoustic recording of an utterance together with its orthographic transcription. This paper focuses on a particular case of hidden Markov model (HMM)-based forced alignment in which the models are directly trained on the corpus to align. The obvious advantage of this technique is that it is applicable to any language or speaking style and does not require manually aligned data. Through a systematic step-by-step study, the role played by various training parameters (e.g. models configuration, number of training iterations) on the alignment accuracy is assessed, with corpora varying in speaking style and language. Based on a detailed analysis of the errors commonly made by this technique, we also investigate the use of additional fully automatic strategies to improve the alignment. Beside the use of supplementary acoustic features, we explore two novel approaches: an initialization of the silence models based on a voice activity detection (VAD) algorithm and the consideration of the forced alignment of the time-reversed sound. The evaluation is carried out on 12 corpora of different sizes, languages (some being under-resourced) and speaking styles. It aims at providing a comprehensive study of the alignment accuracy achieved by the different versions of the speech segmentation algorithm depending on corpus-related specificities. While the baseline method is shown to reach good alignment rates with corpora as small as 2 minutes, we also emphasize the benefit of using a few seconds of bootstrapping data. Regarding improvement methods, our results show that the insertion of additional features outperforms both other strategies. The performance of *VAD*, however, is shown to be notably striking on very small corpora, correcting more than 60% of the errors superior to 40 ms. Finally, the combination of the three improvement methods is also pointed out as providing the highest alignment rates, with very low variability across the corpora, regardless of their size. This combined technique is shown to outperform available speaker-independent models, improving the alignment rate by 8 to 10% absolute.**

***Index Terms*—Corpora annotation, hidden Markov models, phonetic alignment, speech segmentation.**

Manuscript received October 06, 2014; revised March 06, 2015; accepted June 02, 2015. Date of publication July 14, 2015; date of current version November 09, 2015. The work of S. Brognaux and T. Drugman was supported by the FNRS. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yang Liu.

S. Brognaux is with the Cental Laboratory, Catholic University of Louvain, 1348 Louvain-la-Neuve, Belgium, and also with the Circuit Theory and Signal Processing Laboratory, University of Mons, 7000 Mons, Belgium (e-mail: sandrine.brognaux@uclouvain.be).

T. Drugman was with the Circuit Theory and Signal Processing Lab, University of Mons, 7000 Mons, Belgium. He is now with the Amazon Development Center Germany GmbH, 52068 Aachen, Germany (e-mail: drugman@amazon.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2456421

I. Introduction

LARGE speech corpora play a major role both in linguistic research and speech technologies. A particularity of such data is that the sound is rarely studied alone. Orthographic and phonetic transcriptions are usually required. Phonemes, in particular, should be time-aligned with the sound. A precise correspondence between the sound and specific speech segments is essential to allow for prosodic or phonetic analyses. The alignment precision of corpora used to train speech synthesizers or recognizers also determines the quality of the resulting systems [1]–[3]. Many alignment tools offer the possibility to define various transcription levels (see e.g. WaveSurfer [4], Praat [5], ELan [6]). These can then be manually aligned with the sound. However, such a task exhibits two major drawbacks. First it is time-consuming, requiring 130 [7] to 800 [8] times the sound duration. For corpora of several hours as used in speech technologies, this is clearly prohibitive. A second issue lies in the consistency of the alignment, especially if several annotators work on a same corpus. Even with trained phoneticians, high consistency is rarely achieved when several annotators collaborate [9].

To alleviate these issues, automatic alignment tools have been developed (e.g. EasyAlign [10], SPPAS [11]). They offer a consistent and reproducible alignment at reduced cost. The task they perform is known as 'linguistically constrained segmentation' or 'forced alignment'. The phonetization of the text (possibly with phonetic variants) is supposed to be known and only the time boundaries of the phonemes have to be determined. For this purpose, acoustic modeling based on Hidden Markov Models (HMMs), relying on speech recognition techniques, has been shown to achieve the best results [12], [13]. Most existing alignment tools provide the user with pre-trained speaker-independent HMMs of several languages, trained on large databases. The set of provided models being limited, only a reduced set of languages can be aligned. Furthermore, the models highly depend upon the training corpus. When trained on neutral speech, these models tend to produce low-quality alignment of expressive corpora [14]. Some phonemes may also be improperly aligned if they are under-represented in the training corpus (as in [15]).

A possible way to solve these issues is to train the models directly on the corpus to align, thereby providing a better agreement between the training and alignment stages. This technique exhibits the advantage of applying to any language or speaking style without the need for manually-aligned data. It also addresses the widespread criticism stating that HMM-based

alignment tends to be language specific and requires a high amount of training data [16], [17]. This work builds upon our previous studies [14], [18] which indicated that such methods achieved high alignment rates on corpora in French and English, even for small-sized databases. They also showed the significant improvement reached with few bootstrapping data. The first aim of this paper is to investigate whether these results can be generalized, applying the method to a wide range of corpora, with a large diversity of languages, speaking styles and sizes. It focuses on the evaluation of the alignment when modifying various training settings: context-dependent models, number of training iterations, size of the corpus, etc. It then further analyzes typical errors found in the automatic alignment and proposes some refinements of the conventional HMM-based alignment. The proposed improvement methods were developed so that they do not require training on manually-aligned data. Our study offers the advantage to be based on a large database, made of 15 corpora in French, English and rather under-resourced languages such as Faroe or Gaelic. The evaluation is performed on clean, mostly read speech, with manually-verified phonetic transcriptions, as used for speech synthesis purposes. The results are compared to alignment rates obtained with publicly available models of the languages, as used by most existing tools.

The paper is structured as follows. Section II presents existing alignment techniques and focuses on improvement strategies of HMM-based alignment proposed in the literature. Our development protocol is presented in Section III. The baseline method and the typical errors it produces are studied in Sections IV and V. Improvement methods proposed to alleviate these issues are further described in Section VI. The proposed method is then evaluated in Section VII and compared with the use of available speaker-independent models. Finally, Section VIII concludes the paper.

II. Existing Techniques

Several techniques have been proposed to automatically provide the segmentation of speech files. Among these, we can essentially distinguish between methods based on Dynamic Time Warping (DTW) algorithms [19], [20] and those using HMMs [21]–[23]. Additional methods also include the use of acoustic rate of change or discontinuity of the signal [17], [24]. These latter techniques are said to be unsupervised as they rely on the sound only and do not require any phonetic transcription. They automatically identify potential phoneme boundaries based on acoustic features. Their main drawback is that they usually either over or under-detect the number of phonemes, which makes it hard, in a second stage, to map the segments with linguistic units. While DTW algorithms were shown to provide acceptable results [20], HMM-based acoustic modeling has been pointed out as being the most reliable technique for automatic phonetic alignment [12], [13], [25]. It is currently the most widely-used technique for forced alignment.

HMM-based forced alignment uses methods derived from speech recognition. Its specificity is that it is 'linguistically constrained,' which means that the transcription of the sound files is required as input. This transcription may contain phonetic variants, the most likely being selected during the alignment stage [11], [26]. The training stage allows for the building of context-(in)dependent HMMs of each phoneme. Two different initialization stages are proposed, depending upon the provided phonetic transcription. If the latter is not aligned with the speech signal, the phonetic transcription is first uniformly aligned with the sound, with a so-called 'flat-start initialization'. On the other hand, if some part or the whole transcription is time-aligned (referred to as *bootstrap*), the models are directly trained on the corresponding segments for the initialization. The Baum-Welch algorithm allows for the training of the model parameters. In the alignment stage, the models are used to align a corpus (possibly identical to the training corpus) with the Viterbi algorithm, which provides the best path among the network of possible transitions.

Most studies in the literature have investigated the alignment obtained when training the models on a large database and using these language-dependent models to align other corpora [26], [27]. Depending upon the study, the training stage relies on aligned or non-aligned data. While training directly on the corpus to align allows alleviating the need for large transcribed corpora of the language, the performance of this method has been the topic of very few studies. In [13], its interest for under-resourced languages was investigated. This method was also shown to provide interesting results for English corpora [9] but no evaluation was made of the obvious role played by the size of the corpus. Moreover, the alignment rate was not compared to results achieved when using available models of the language trained on large amounts of data.

Several studies have tried to improve conventional HMM-based forced alignment. Different post-processing methods have been developed to refine the produced alignment, using various statistical techniques. However, they usually depend on the training of a second model (e.g. a regression tree [12], Gaussian Mixture Models (GMMs), HMMs [28] or support vector machines (SVM) [29]) which requires a manually-annotated corpus. Heuristic rules can also be used to modify the boundaries as a post-process, making use of the average deviation for each pair of phonemes [30]. In a similar way, it is proposed in [31] to train, on a manually-aligned corpus, boundary models depending on the right and left context of each phoneme to improve the alignment. They also suggest to make use of an automatic detection of discontinuities in the signal, based on [24]. It was proposed in [21] to make use of a Spectral Variation Function (SVF) as post-treatment to move boundaries to the nearest junction. This, however, did not yield alignment rate improvement.

The combination of various HMM-based alignments trained with different parameters has also been investigated [32]. Here again, manually-aligned data is required for the training. For approximate transcriptions, a slightly modified version of HMM-based alignment was also proposed in [33], proceeding in multiple iterations and combining recognition and alignment methods.

Most of the aforementioned techniques are supervised and therefore require manually-aligned data for training. The annotation process can be tedious when aligning rare languages for which few data are available. Furthermore, manual alignment is known to be very time-consuming and consequently costly

[7], [8]. Contrary to the methods described above, the approach considered in this work is unsupervised and does not require any labeled data. A flat-start initialization is first considered on the target corpus, and the alignments are expected to converge across the training iterations. This technique is used as a baseline in the remainder of this paper. In [14], [18], however, this method was shown to achieve rather poor results on some corpora, especially for highly expressive speech.

An analysis of the typical errors made by a standard HMM-based alignment is provided in Section V. Based on this analysis, we propose in Section VI three improvement techniques specifically designed to alleviate these alignment flaws. Conversely to most existing improvements of HMM-based forced alignments, these methods exhibit the advantage of not requiring the use of manually-aligned data. A first widely-known improvement lies in the augmentation of the feature vector. The addition of new acoustic characteristics to the feature vector has already been investigated in [21], [34]–[36]. In [21], for instance, features related to the spectral variation have been added and led to a slight increase in the alignment rates. Besides, we also propose two novel strategies: a better initialization of the silence model based on a VAD algorithm and the use of the alignment of the time-reversed sound to provide smoothed boundary estimations.

These refinement techniques are detailed in Section VI.

III. Development Protocol

A. Development Database

The proposed techniques are developed on clean and read speech, as used for speech synthesis purposes. Our development set is made of 3 corpora, kindly provided by Acapela Group SA, and sampled at 22050 Hz. Each corpus has a total duration (silences included) ranging between 11 and 14 minutes, with 131 to 200 speech files. The aim is to try to cover various features of speech: different types of expressivity and speaking style (sad, happy and little creature, a monster-like voice), two languages (English and French) and three different speakers (of both genders). The phonetic transcriptions were checked and manually aligned with the sound. The French and English corpora are respectively annotated with 39 and 56 phonetic symbols.

B. Evaluation Metrics

The evaluation of the methods is based on the comparison between the automatic and the manual segmentation. Two metrics are used throughout our experiments. The first measure, known as the boundary-based measure [15], is used in most studies. It computes the percentage of boundaries that are correct, with a certain tolerance threshold ranging from 10 to 40 ms. This measure will be referred to as the *correct alignment rate* in the remainder of the paper.

It should be noted that several studies have pointed out the high degree of inter-rater disagreement, with sometimes large discrepancies between human-made alignments [37]. Generally, results are provided within a certain tolerance threshold on the timing error. Usually, 20 ms constitutes a limit above which the inter-annotator agreement rate is fairly high. Using this threshold, [10] 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 90% were obtained on an Italian corpus in [38]. Similarly, 93.5% agreement were obtained on the TIMIT corpus [35]. In [39], the average distance between boundaries of human annotators was shown to be 16 ms. A value of 20 ms is also considered as an acceptable limit for speech synthesis purposes [12]. Throughout our experiments, we provide the performance measures using a threshold of 20 and 40 ms (errors exceeding 40 ms can be seen as gross errors).

The second metric used to assess the performance of the proposed method is the relative improvement over our baseline: the Train&Align algorithm described in Section IV. The so-called *relative improvement* is here defined as the relative reduction of the alignment error rate at a certain tolerance threshold. Positive values therefore indicate an effective improvement, while negative values indicate a degradation of the alignment performance.

IV. HMM-Based Forced Alignment Baseline: Train&Align

Our first experiments make use of the standard HMM-based alignment as described in Section II. The only specificity of our technique is that the models are directly trained on the corpus to align. This method, called Train&Align (T&A), was presented in [14], [18], and was further developed into a tool [18] freely available online[1]. Its specific implementation, along with the tuning of its parameters, is further described in the next paragraphs. *T&A* will also serve as baseline for the evaluation of the proposed improvement methods.

In a first stage, the entire (unaligned) corpus to align and its phonetic transcription are used to train a new language model. The correct phonetization of the sound is provided, with no variant. Acoustic vectors are made of 39 features, i.e. 12 Mel-Frequency Cesptral Coefficients (MFCCs) and the log-energy, along with their delta and acceleration coefficients. The models are left-to-right 5 state monophones with no skip and three emitting states. They are initialized with a so-called 'flat start,' i.e. a uniform segmentation of the speech signal. For silences, a standard widely-used configuration is applied. Two specific models are added: a silence model ('sil') represents silent pauses and allows for a direct transition back and from second to fourth state to better model duration variations. A short-pause tee model ('sp') is also implemented and automatically inserted between words. It allows for automatic detection of non-annotated silences. It has one emitting state tied to the center state of the silence model. *T&A* computes a commonly-used two-pass training: three iterations of the Baum-Welch algorithm are applied before the introduction of the 'sp' model and five iterations after, as proposed in [34]. Training and alignment are performed with the HTK toolkit [34].

Based on that standard alignment method, the tuning of several training parameters was investigated. Specifically, the size

[1]http://cental.fltr.ucl.ac.be/train_and_align/

TABLE I
ALIGNMENT RATES FOR MONOPHONE ALIGNMENT AND RELATIVE IMPROVEMENT (IN %) WITH TRIPHONES AND TIED-STATE TRIPHONES

Tolerance	20 ms			40 ms		
	Mono	Tri	Tied	Mono	Tri	Tied
Will (Creature)	63.13	-2.32	3.12	84.27	0.84	1.67
Antoine (Sad)	75.40	1.64	3.39	86.71	3.46	7.36
Margaux (Happy)	83.71	2.92	2.47	95.08	-0.37	-3.35
Average		0.75	2.99		1.31	1.89
Standard Deviation		2.73	0.47		1.96	5.36

and overlap of the speech frames, the use of context-dependent models, and the number of training iterations were further analyzed.

Regarding the speech frames, six options were considered with overlapping and non-overlapping frames varying from 10 to 30 ms. Our results show that frames of 10 ms with no overlap provide the best alignments, with an absolute improvement of the boundary-based alignment rate of about 15%, with a 20 ms threshold compared to frames of 30 ms with a shift of 10 ms, as commonly used in automatic speech recognition (ASR). While this may look contradictory to ASR standards, it can be explained by the fact that narrower windows with no overlap avoid mixing data belonging to different phonemes. Conversely, ASR usually makes use of highly context-dependent models, mixing data helping in the recognition process as they are not concerned with the precise location of the phoneme boundaries.

The use of context-dependent models was also investigated, replacing monophones by triphones (sequences of three phonemes) or tied-state triphones, in which the left and right contexts are defined by phonetic classes. These classes are here determined based on the following phonetic criteria: kind (e.g. vowel, semi-vowel, etc.), voicing, location, articulation type and lip rounding. Results in Table I show that the effect of context-dependent models is highly dependent on the corpus. While they improve the alignment of Antoine (Sad), triphones tend to degrade the alignment of Will (Creature) and tied-state triphones have a negative effect on the alignment Margaux (Happy). Interestingly, Margaux is also the best-aligned corpus with monophones. We may therefore wonder whether the use of tied-state triphones would be more useful only when the monophone initial alignment is relatively low, but this would need to be confirmed on more data. Interestingly, there seems to be a tradeoff, tied-state triphones carrying out, for all three corpora, improvement for finer resolutions, while monophones could be more suited at coarser resolution (here with the exception of Antoine (Sad)).

Earlier studies have indicated that the use of context-dependent models may not be suited for forced alignment [7], [40], [41]. A possible justification was that the context-dependent models are always trained within a specific context, which implies that they may not learn to correctly discriminate between the phoneme itself and its context. Our results partly confirm this hypothesis. Besides, the use of such context-dependent models slows down the training process. For these reasons, the remainder of this paper will focus on the use of monophone models only.

TABLE II
FOR SPECIFIC PHONEME CLASSES, PERCENTAGE GLOBAL AND SPECIFIC ERRORS ($>$ 40 ms)

Phoneme class	Global errors (%)	Specific errors (%)
Silences	61.61 %	49.79 %
Vowel-vowel	25.13 %	38.31 %
Approximants	13.89 %	7.61 %
Plosives	6.69 %	2.17 %

As previously mentioned, the standard alignment technique typically includes 3 iterations for the first training stage and 5 for the second, i.e. after the insertion of the short-pause tee models. However, it is encouraged in [34] to optimize that number of training iterations. We investigate here, on the development dataset, the role played by the number of iterations on the alignment rates. A first round of experiments showed that only the number of iterations of the second training stage should be optimized. All three corpora were then aligned with a varying number of iterations in that stage, ranging from 2 to 40. This indicated high variability across the corpora, the best alignment rates being reached after 17 to 31 iterations. To account for this influence of the corpus, we investigated the possibility of dynamically optimizing the number of embedded training iterations based on the log-likelihood per frame of the training data. This technique relies on the fact that Pearson's tests revealed strong positive correlation scores (average Rho of 0.93) between the log-likelihood of the model and the alignment rates at 40 ms, consistently across all three corpora: if the log-likelihood stagnates, it can be interpreted as an indication that the training should be stopped.

Based on experimental tests, the minimum increase in log-likelihood was set to 0.001, as the alignment performance curve seems to flatten past that level. The results also drove us to set the maximum number of iterations to 35. This dynamic optimization of the number of training iterations provides consistent improvement across all development corpora and for all tolerance thresholds. Average relative improvement reaches 7.08% at 20 ms and 11.50% at 40 ms.

V. ANALYSIS OF ERRORS MADE BY T&A

In order to develop methods improving HMM-based phonetic alignment, an in-depth analysis of the errors most frequently made by *T&A* was carried out (see Table II). We focused on errors with a timing difference superior to 40 ms which correspond to high incongruencies unlikely to be produced by human annotators. In this analysis, we distinguished between two measures. The *global error rate* regards the percentage of all errors that this specific phoneme class accounts for. The *specific error rate* relates to the percentage of transitions of that class which are erroneous. This distinction is important as it shows, for example, that transitions between vowels are often prone to errors (in 38.31% of the cases) but that they account for a smaller global error rate (i.e. 25.13%), being less frequent than most other transitions in the corpora.

This study drove us to distinguish between 4 typically problematic transitions. The first relates to transitions to and from silences, which account for more than 60% of the total amount of

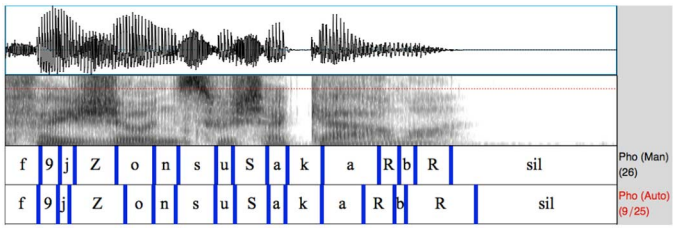

Fig. 1. Example of silence and liquid misalignment in the automatic alignment (Pho (Auto)) of one speech file of the development set.

errors higher than 40 ms. In fact, 49.79% of the silences are erroneously aligned with that threshold. This issue has been pointed out by other studies [26], [31], which noticed low alignment rates for silence boundaries in their corpus. Interestingly, other segmentation algorithms like unsupervised methods based on the acoustic rate of change were also shown to produce similar errors [17]. This misalignment is even more striking provided that our development set contains corpora designed for speech synthesis, i.e. with no background noise. This issue may then be even worse for corpora recorded in noisier conditions.

Other error types that were found to be frequent are, in decreasing order, vowel-vowel transitions, approximants and plosives. For all three categories, pairs containing a silence were excluded, silences inducing a high amount of errors. Approximants include lateral approximants (like [l]), non-lateral approximants (like [ɹ]), and semi-vowels (like [j], [w] or [ɥ]). Semi-vowels were shown to be problematic in [15], and also present issues for human annotators, probably due to their high degree of co-articulation. Vowel-to-vowel transitions were also pointed out as prone to errors in [9], [21], [37]. For approximants, especially the initial boundary tends to be problematic. Conversely, final boundaries of plosives seem to be more prone to errors, the explosion stage being harder to model. Fig. 1 shows an example of typical misalignment.

While the optimization of the number of training iterations was shown to improve the alignment rate, it should be noted that it does not impact a specific type of error. All classes of errors tend to be reduced with better improvement for some phoneme classes, depending on the corpus.

A possible way to reduce the amount of errors consists in using some manually-aligned data as bootstraping data so as to produce a better initialization of the models [21]. This solution is further explored in Section VII, along with the role played by the size of the bootstrapping corpus. It requires, however, some manual intervention. In order to stick to a fully automatic method, the next section presents three improvement strategies driven from the aforementioned typical errors.

VI. Improvement Methods

A. Using Voice Activity Detection Algorithm (VAD)

Seeing that silences are often prone to alignment errors, we propose here to improve their initialization. As no manually-aligned part of the corpus is provided, our standard HMM alignment (*T&A*) relies on a 'flat start' uniform initialization. This means that each phone model is first assigned average values (for means and variances) and uniformly aligned with the sound.

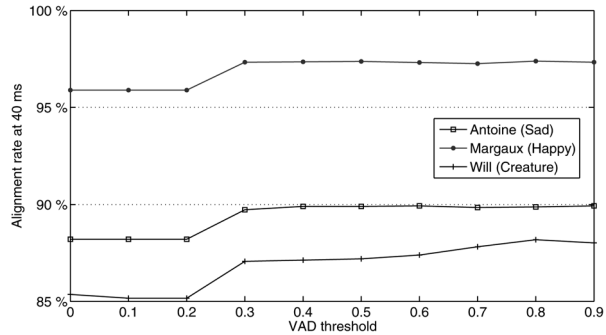

Fig. 2. Correct alignment rates with a varying VAD threshold with a 40 ms tolerance threshold, for the three development corpora.

The proposed refinement aims at modifying this initialization stage. Sohn's voice activity detection (VAD) algorithm [42] is first applied to the sound files and allows detecting non-speech segments. These segments are then used to initialize silence models only. All other phonemes are initialized with the 'flat start' strategy, exactly as it was used in *T&A*. This particular use of the VAD exhibits the advantage not to fix the silence boundaries, which would be problematic as VAD algorithms tend to over- or under-detect silences notably in plosion stages or noisy pauses. For that reason, these VAD-detected boundaries cannot be directly used to allow for a better uniform segmentation of the other phonemes. It allows, however, the silence models to be better initialized and helps to iteratively converge to better silence boundaries. Besides, our method provides an initialization of the silence model specific to the corpus and its recording conditions, conversely to the use of a generic silence model. This may also help for corpora recorded in noisier conditions.

Sohn's VAD algorithm generates trajectories of posterior probabilities about the presence or not of speech activity. In order to draw a binary decision, a threshold has to be applied on these trajectories, below which the sound is assumed not to contain speech. With a null threshold, all frames are regarded as voice and no signal section is exploited to train the silence models. The alignment is then that of the original *T&A*. Experiments on our development database showed that the performance of the alignment reaches a plateau from a threshold of 0.3 (see Fig. 2). For the remainder of our experiments, we set the threshold to 0.8 which provides the best alignment rates on average. This achieves an average relative improvement of 23.29% on the development set at a 40 ms threshold.

Interestingly, the alignment rates achieved by *VAD* 0.8 are very similar, regardless of the optimization of the number of training iterations. When discarding this optimization from the baseline, the relative improvement of *VAD* reaches 37% on average with a 40 ms threshold. This indicates that the *VAD* also helps to compensate for a lack of training iterations: if the initialization is better performed, less embedded training iterations are required. Another interesting observation is that *VAD* especially reduces the amount of gross errors (i.e. errors superior to 30 ms) while it has less effect on finer errors.

B. Augmentation of the Feature Vector (AddFeat)

It is widely acknowledged that part of the errors made by automatic phonetic alignment techniques comes from the fact

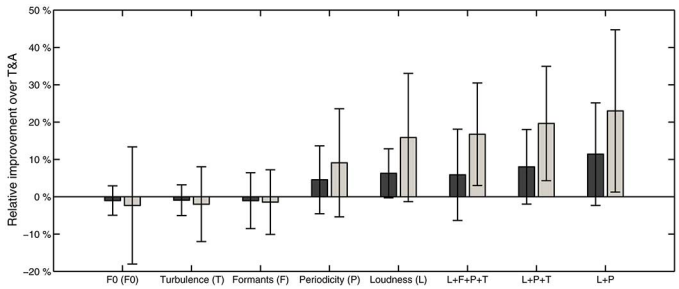

Fig. 3. Average relative improvement over *T&A* with additional acoustic features with 20 and 40 ms tolerance thresholds, on the development set, along with their 95% confidence interval.

that humans make use of additional cues, which are not represented in the MFCC coefficients, to decide the precise location of the boundaries [30]. This may especially be the case for vowel-vowel transitions or transitions to and from semi-vowels which are very hard to model. For that reason, we considered the possibility to add up to 9 supplementary acoustic features on top of the MFCC coefficients.

These additional features are: *i)* the total perceptual loudness [43], *ii)* the frequencies of the first five formants estimated by the Differential-Phase Peak Tracking (DPPT [44]) technique, *iii)* the fundamental frequency F0 and a measure of periodicity based on the Summation of the Residual Harmonic (SRH [45]) algorithm, and *iv)* a measure of turbulence using the Chirp Group Delay (CGD) function which was shown in [46] to highlight irregularities of phonation.

For each of the proposed additional features, we have calculated its coefficient of correlation with the MFCCs (static and dynamic values) using the canonical correlation method. This analysis was based on the speech segments of the development dataset. The results are as follows: perceptual loudness (0.93), F1 (0.63), F2 (0.48), F3 (0.33), F4 (0.29), F5 (0.17), F0 (0.57), periodicity (0.74), turbulence (0.47). For some features, this correlation is quite high. That is the case of the perceptual loudness (which was expected) and surprisingly of the periodicity measurement. Note however that even if a feature is redundant with an existing set, it might still convey relevant complementary information. As further described, that is the case of the loudness, which despite its high correlation is shown to carry out an interesting improvement.

The contribution of each feature, individually, is shown in Fig. 3. High variability is observed across the corpora and only the addition of loudness significantly improves the alignment of all datasets. Various feature combinations have also been investigated. While the addition of both loudness and periodicity is shown to provide the best relative improvement on average, their combination with turbulence leads to a reduced variability across the corpora. With that configuration, the initial alignment is improved by 10 to 35%, with a 40 ms tolerance threshold, for all three development corpora. This combination will be exploited in the remainder of the paper and referred to as *AddFeat*. This amounts to a total of 42 coefficients.

Here again, the interaction between the addition of new features and the dynamic optimization of the number of training iterations is worth discussing. With a standard fixed number of iterations, all individual features are shown to provide relative improvement over the baseline. The minor impact played by features such as formants or F0 seems however to disappear with an optimized number of training iterations.

Seeing that the initial set of parameters contains delta and acceleration coefficients, the insertion of these derived coefficients for the 9 additional features was also considered. However, the analysis of the results on our corpora shows that the consideration of these derivatives does not bring any relevant complementary information when included in addition to their static version. The 9 supplementary features are therefore added to the feature vector without their derivatives.

TABLE III
EXAMPLE OF SYSTEMATIC ERRORS MADE BY *T&A*

Boundary	Percentage of cases	Average deviation
Late boundaries		
Initial boundary of silences	74.6 %	47.80 ms
Final boundary of [j]	70.85 %	7.62 ms
Early boundaries		
Final boundary of silences	82.60 %	28.90 ms
Initial boundary of [t]	69.19 %	3.84 ms

C. Exploitation of the Time-Reversed Sound (Reverse)

An in-depth analysis of the alignment errors with *T&A* revealed systematic errors, some phonemes being often misaligned with a too early or too late boundary. Examples of such errors can be observed in Table III. It shows, for example, that final [j] boundaries tend to be predicted too late while final silence boundaries tend to be too early. As can be observed, several aforementioned classes of errors are here concerned (e.g. plosives, semivowels, silences). This type of errors occurring in the same direction for specific transitions has been pointed out by [12]. In [41], schwas tended to be left-shifted in their corpus. On the whole, we notice that a majority of boundaries are predicted too early.

To reduce such errors, we propose to time-reverse the sound and its phonetic transcription. The initial feature vector of 39 MFCC-based features (and possibly additional features) is extracted on this time-reversed copy of the sound. *T&A* is then used to align the reversed and the original corpora. Both alignments are then exploited to compute average boundary locations. This sound reversal essentially modifies transition probabilities of the HMM. The segmentation of the sound also begins from the end of the file, which might modify the content of the respective frames of signal. The assumption behind this method is that, for highly probable boundaries, the alignment of both corpora should provide similar results which will not influence the resulting alignment. Conversely, for uncertain boundaries, computing the average between both alignments should provide smoothed estimations, thereby reducing errors with a high tolerance threshold. To the best of our knowledge, this technique has never been investigated in previous studies.

VII. EVALUATION

This section shows the results obtained by our alignment methods on the evaluation database (presented in

TABLE IV
EVALUATION DATA SET FEATURES

Corpus	Language	Speaking style	Duration (minutes)
Will (Bad guy)[2]	English	Read/Expressive	12
Will (Neutral)[2]		Read/Neutral	14
Woggle [47]		Read/Expressive	51
Antoine (Happy)[2]	French	Read/Expressive	12
Margaux (Sad)[2]		Read/Expressive	12
Marie		Read/Neutral	108
Sportic [48]		Spontaneous/Expressive (sports commentaries)	15
Faroe [2]		Read/Neutral	21
Gaelic [3]		Read/Neutral	8
Afrikaans [13]		Read/Neutral	22
Setswana [13]		Read/Neutral	47
Isizulu [13]		Read/Neutral	20

[2] kindly provided by Acapela Group SA
[3] kindly provided by the Phonetics and Speech Laboratory, Trinity College, Dublin

Section VII-A). The baseline *T&A* alignment is evaluated in Section VII-B along with the role played by the size of the corpus (Section VII-C) and the size of the bootstrapping corpus when one is exploited (Section VII-D). Our improvement methods are then evaluated in Section VII-E and the impact of the size of the corpus is investigated in Section VII-F. Finally, Section VII-G shows the comparison between our best alignment method and alignment rates obtained by state-of-the-art available models of the corresponding languages.

A. Speech Material

To assess the performance of the various techniques, 12 corpora are used (see Table IV). They vary in terms of language, size, and speaking style. Most are read speech, with high recording quality, used for speech synthesis. The use of Sportic allows for an analysis of the results on spontaneous speech. Each corpus contains one speaker, male or female, except for Woggle which consists of recordings from 5 female speakers. This corpus is characterized by a high level of variability, containing also 5 different emotional states (e.g. happy, sad, angry). All corpora are classified as neutral or expressive. The expressive tag contains different kinds of expressivity: emotions with different valences (happy, sad, angry and afraid) and specific attitudes/speaking styles (bad guy and sports commentaries). The advantage of our basic *T&A* technique being to apply to any language, the methods are also tested on under-resourced languages like Faroe and Gaelic. All corpora were manually phone-aligned by experts.

B. Results with the Baseline: Train&Align

Our baseline approach with training on the corpus to align (*T&A*, as described in Section IV) is first applied to all corpora. The resulting correct alignment rates are shown in Table V. Rates around 80% with a 20 ms threshold are reached for all of our corpora in French and in some under-resourced languages. This is comparable to observed inter-annotator agreement rates in [10]. Low alignment rates for Woggle, can be explained by its high degree of expressivity and diversity, as it contains 5 speakers and 5 different emotions. Conversely to the other models, the one trained on Woggle does not capture a specific speaking style of a speaker, which might explain the lower alignment rates.

TABLE V
CORRECT ALIGNMENT RATES (IN %) OF *T&A* ON OUR EVALUATION CORPORA IN ENGLISH (EN), FRENCH (FR) OR UNDER-RESOURCED LANGUAGES (O)

Tolerance threshold		10 ms	20 ms	30 ms	40 ms
En	Will (Bad guy)	47.04	72.14	83.08	88.19
	Will (Neutral)	50.86	78.86	89.72	94.64
	Woggle	44.95	65.23	80.07	88.82
Fr	Antoine (Happy)	59.08	82.70	89.14	92.59
	Margaux (Sad)	59.38	81.12	88.99	92.26
	Marie	60.06	84.42	92.93	96.63
	Sportic	65.01	82.53	90.31	94.22
O	Faroe	48.45	75.18	87.95	93.74
	Gaelic	77.34	89.98	94.25	96.05
	Afrikaans	45.66	73.50	88.38	93.11
	Isizulu	46.78	72.18	85.22	90.94
	Setswana	47.29	70.51	84.71	90.31

The optimization of the number of training iterations was shown in Section VII-B to improve the alignment and was therefore included in the baseline. While its application allows for an average relative improvement of 8% (40 ms threshold) of the alignment of the evaluation dataset, it performs much better on highly expressive corpora like Woggle (+16%) and Sportic (+45%), which was shown to contain highly excited parts [48]. Conversely, it achieves lower performances on Marie (−3%), which contains read neutral speech.

The acoustic models being here directly trained on the corpus to align, the size of the corpus plays an obvious role in the model quality. It can therefore be wondered whether the rather high alignment rates obtained for Marie can be explained by the size of the corpus. More than 100 minutes of speech are used to train the models. This provides a fair amount of occurrences for each phoneme, which is not possible for all databases. This question is now addressed in Section VII-C.

C. Influence of the Corpus Size

The influence of the corpus size on the alignment performance was analyzed by varying the size of the various corpora. This evaluation was performed on the three largest corpora of our evaluation dataset which have each a duration superior to 30 minutes (initial and final silences excluded). The resulting correct alignment rates are shown in Fig. 4. It shows that the alignment rate plateaus for corpora larger than 5 minutes. A similar curve was observed with a 40 ms tolerance threshold. For Setswana, 2 minutes are enough to reach the correct alignment maximum. For corpora smaller than 2 minutes, the alignment performance rapidly degrades. It should be noted that the alignment of the 2-minute Woggle corpus is, comparatively to the two other corpora, of rather low quality. This may be explained by the fact that properly learning the higher variability of the corpus (various emotions and speakers) requires larger datasets. On the whole, a few minutes of neutral speech seem to be sufficient to train and align a new corpus, which confirms findings in

TABLE VI
RELATIVE IMPROVEMENT (IN %) OVER *T&A* WITH 20 AND 40 MS TOLERANCE THRESHOLDS FOR *VAD*, *ADDFEAT*, *REVERSE* AND THE COMBINATION OF THE THREE METHODS (*T&A2*), FOR ALL EVALUATION CORPORA, ALONG WITH THE AVERAGE AND STANDARD DEVIATION OF THE IMPROVEMENT

Tolerance threshold	20 ms				40 ms			
Improvement method	VAD	Reverse	AddFeat	T&A2	VAD	Reverse	AddFeat	T&A2
Will (Bad guy)	0.09	2.59	20.03	19.69	34.62	4.28	43.99	40.73
Will (Neutral)	0.95	-0.19	5.15	5.15	6.77	-33.83	14.29	13.53
Woggle	30.26	24.60	18.66	37.83	30.60	20.72	21.23	36.94
Antoine (Happy)	0.48	0.16	5.94	20.22	12.36	-6.37	4.87	45.32
Margaux (Sad)	25.61	3.51	17.43	40.31	50.59	-2.38	25.42	63.18
Marie	15.95	13.61	19.68	26.25	17.53	31.25	18.93	23.92
Sportic	2.06	13.57	0.98	0.65	-11.15	9.51	11.15	9.51
Faroe	-18.00	7.19	5.63	14.42	-10.00	12.57	10.27	22.70
Gaelic	-73.62	-15.96	2.61	-85.02	5.79	1.65	10.74	12.40
Afrikaans	3.79	-7.79	9.10	22.04	-11.10	6.44	3.82	1.67
Isizulu	36.47	8.21	0.17	38.39	26.51	39.37	-2.89	46.33
Setswana	12.77	-6.91	14.36	11.54	11.50	1.53	10.81	22.75
Average	3.07	3.55	9.98	12.62	13.67	7.06	14.39	28.25
Standard deviation	28.55	10.97	7.60	33.28	19.48	18.74	12.13	18.28

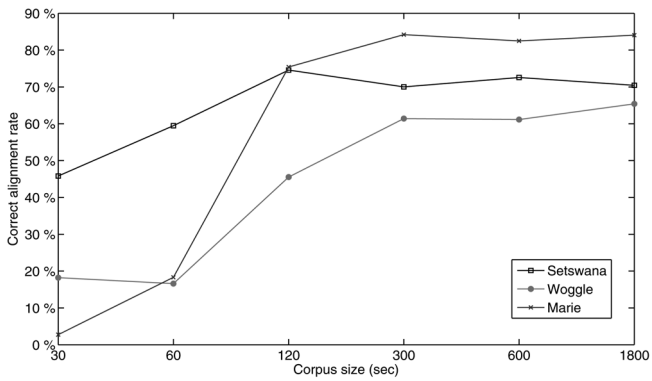

Fig. 4. Correct alignment rates with a 20 ms tolerance threshold on the evaluation set, with varying corpus sizes.

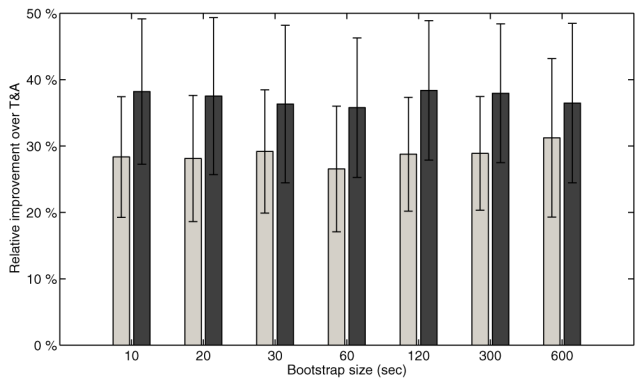

Fig. 5. Average relative improvements over *T&A* with 20 and 40 ms tolerance thresholds, using a bootstrapping corpus varying in size, along with their 95% confidence interval.

[13] but is much lower than the limit of 256 sentences fixed by [21]. This may be explained by the fact that [21] does not make use of speaker-dependent models, the training being performed on a corpus different from the corpus to align.

D. Effect of Bootstrap

Section VII-B showed that some expressive corpora and datasets from low-resourced languages are rather poorly aligned with a standard HMM-based alignment when training on the corpus to align. A possible way to alleviate this issue is by improving the initialization of the models based on some manually-aligned part of the corpus. This section investigates which size of the bootstrapping corpus is required and which improvement may be expected. For this experiment, we gradually increased the size of the bootstrap (from 10 to 600 seconds, initial and final silences excluded) while evaluating on a fixed portion of the corpus (2 minutes) to avoid biases due to a varying evaluation corpus. The evaluation set was never included in the bootstrap data. The monophones are initialized on manually-aligned data only if at least three occurrences of the phoneme are available. If not, average values (for means and variances) are assigned to the model, similarly to a 'flat start' initialization. Fig. 5 interestingly points out that the use of a bootstrap part of the corpus, as small as 10 seconds, leads to a relative improvement of about 25 to 35%. While the use of a larger bootstrap slightly improves the quality, the curve rapidly flattens, the use of a bootstrap corpus larger than 30 seconds being essentially unnecessary. As expected, the use of bootstrap data is especially effective on expressive corpora like Woggle, which was poorly aligned with the initial *T&A* technique. Improvement of about 53% is here observed with only 30 seconds of bootstrap data, for both 20 and 40 ms thresholds. High improvement is also found on some corpora in under-resourced languages: the use of 30 seconds of bootstrap improves the alignment of Setswana, in particular, by 56% with a 20 ms threshold, to reach alignment rates of more than 87%.

While bootstrap methods have been widely acknowledged as playing a crucial role in the improvement of HMM-based alignment, they still require some manual processing, be it very small. The remaining sections of this paper investigate whether the three automatic techniques described in Section VI are also effective in improving the final alignment.

E. Improvement Methods

This section compares the contribution of each refinement method defined in Section VI independently. The relative im-

provements over the baseline (*T&A*) computed on our evaluation database are shown in Table VI.

Interestingly, all techniques are seen to carry out improvement, especially at a 40 ms tolerance threshold. As expected, they help correcting gross errors, unlikely to be produced by human annotators. While both *VAD* and *AddFeat* achieve an average improvement of more than 13% with a 40 ms threshold, *AddFeat* is shown to be much more consistent across the corpora, positively impacting the alignment of all corpora but one. Both *VAD* and *Reverse* display a quite high inter-corpus variability. This variability for *VAD* is mostly due to its low performance on under-resourced languages. When looking at English and French corpora only, the average improvement exceeds 20% at a 40 ms tolerance threshold. Some under-resourced languages seem to be less sensitive to the benefits of this proposed method. In-depth linguistic analyses of these languages may provide greater insight into their reaction to the various improvement methods.

As previously mentioned, the optimization of the number of training iterations in the baseline allows for an average relative improvement of about 8% at a 40 ms tolerance threshold. Experiments on the evaluation set, without this dynamic adaptation of training iterations confirmed the observations made on the development corpora. Here again, *VAD* is shown to compensate for a lack of number of training iterations as it provides similar alignment rates without this training optimization, the relative improvement exceeding then 20% at 40 ms threshold on average, and 30% on French and English corpora only. Conversely, *AddFeat* seems to be rather complementary to the training optimization, their combination providing higher alignment rates.

We also investigated the improvement achieved by combining the methods. It is worth noting that the combination of all three improvement methods, i.e. *T&A2*, provides the best results. The improvement over *VAD* or *AddFeat* alone is statistically significant (respectively, $p < 0.001$ and $p < 0.05$ with paired t-tests). This method is also highly consistent, carrying out improvement for all evaluation corpora at 40 ms, regardless of their style and language. At 20 ms, only the alignment of Gaelic is degraded, due to the poor performances of *VAD* on that corpus, which reduces the correct alignment rate at 20 ms by about 8.5%. This can be partly explained by the fact that Gaelic is the best-aligned corpus with *T&A* and that this baseline alignment contains very few erroneous silence boundaries. On average, *T&A2* reaches high improvement rates, with a decrease of nearly 30% of the errors superior to 40 ms. Interestingly, *T&A2* allows reaching relative improvement at 40 ms threshold that is similar to that obtained when using bootstrapping data, at least for some corpora of our evaluation set. It should be noted that other combinations have been tested, especially the combination of *VAD* with *AddFeat* only, as *Reverse* was seen to achieve few improvement alone. Interestingly, *Reverse* is shown to contribute to the improvement of *T&A2*, increasing the relative improvement by 3.9% at 40 ms, on average.

A detailed analysis of the errors shows that, as expected, *VAD* contributes to the reduction of the amount of errors related to silences. While, on average, 29.7% of the silence boundaries of our evaluation set are erroneously aligned with an error superior to 40 ms with *T&A*, this rate reduces to 21.1% when applying the *VAD* refinement. *AddFeat* is shown to reduce both the errors related to silences and to vowels (by respectively 7 and 4% on average).

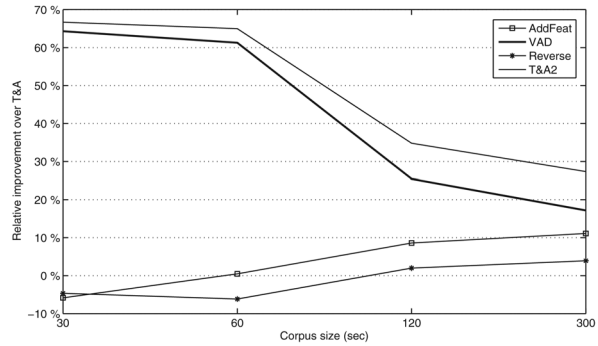

Fig. 6. Relative improvement over *T&A*, with a 40 ms tolerance threshold, for the different improvement methods, with varying corpus sizes.

F. Advantage of T&A2 in Small Corpora

To investigate the role played by the size of the corpus on the performance of the various improvement methods, the size of the corpora was modified to range between 30 and 300 seconds. Section VII-C showed that corpora shorter than 120 to 300 seconds tend to be poorly aligned. This experiment sets out to investigate whether our improvement methods allow alleviating that issue. The results for 40 ms tolerance thresholds are shown in Fig. 6. Similar patterns are found with a 20 ms tolerance threshold.

Both *Reverse* and *AddFeat* are shown to be ineffective on short corpora. Their positive impact, however, gradually improves with the size of the corpus. Their use for corpora shorter than 2 minutes is highly inadvisable. What should be highlighted, however, is that *VAD*, conversely to the other methods, produces significantly higher alignment rates, especially on very small corpora for which its improvement over the baseline (*T&A*) becomes dramatic. This finding is of utmost importance as it allows achieving fair alignment rates on corpora for which a standard alignment performs very poorly. An example of such improvement can be observed on the 30-second version of the French expressive corpus Antoine (Happy), which shows improvements in terms of alignment rates from 7.2% and 11.7% (for 20 and 40 ms thresholds respectively), to 71.5% and 88.9% with the use of *VAD*. Consistency is found across all small corpora, with a minimum relative improvement of 32.2% and 15.8% for 40 ms errors, with corpora of 30 and 60 seconds respectively. Interestingly, the combined method *T&A2* achieves the best results, even on the smaller corpora. This seems to indicate that, while *AddFeat* and *Reverse* are ineffective when considered separately, their combination to *VAD* still contributes to alignment improvement. It should however be noted that *T&A2* is only shown to significantly outperform *VAD* with corpora of 120 and 300 seconds ($p < 0.05$ with paired t-test).

G. Comparison with State-of-the-Art Models

As previously mentioned, most automatic alignment tools provide the user with pre-trained speaker-independent acoustic

TABLE VII
CORRECT ALIGNMENT RATES (IN %) ACHIEVED BY AVAILABLE SPEAKER-INDEPENDENT MODELS (WHEN AVAILABLE) AND BY OUR COMBINED METHOD (*T&A2*)

Tolerance threshold	10 ms	20 ms	30 ms	40 ms
A French neutral corpus: Marie [50]				
SPPAS	45.76	71.28	83.06	89.84
EasyAlign	52.54	77.54	87.72	92.11
T&A	60.06	84.42	92.93	96.63
T&A2 (proposed)	**65.76**	**88.51**	**94.84**	**97.44**
An English expressive corpus: Woggle [47]				
VoxForge	37.3	65.2	82.07	88.69
T&A	44.95	65.23	80.07	88.82
P2FA	46.68	69.74	81.2	87.46
T&A2 (proposed)	**54.01**	**78.38**	**88.92**	**92.95**
Rare languages: Faroe				
T&A	48.45	75.18	87.95	93.74
T&A2 (proposed)	**52.71**	**78.76**	**90.00**	**95.16**
Rare languages: Isizulu				
T&A	46.78	72.18	85.22	90.94
T&A2 (proposed)	**58.46**	**82.86**	**91.82**	**95.14**

models of the language (e.g. EasyAlign [10], SPPAS [11], P2FA [49]). In [14], we showed that, for medium-size corpora, training the acoustic models directly on the corpus to align achieves comparable alignment rates. We compare here the performance of these speaker-independent models with our proposed combined method: *T&A2*. This comparison is carried out with two available French models (EasyAlign [10] and SPPAS [11]) and two English models (VoxForge used by SPPAS [11] and P2FA [49]). Results are shown in Table VII. Compared to the best speaker-independent model, an absolute increase in the correct alignment rate of more than 10% with a 20 ms threshold is observed for the French neutral corpus. For the English expressive database [47], [14] pointed out that P2FA achieved slightly better results than the *T&A* alignment. A possible cause was that P2FA models were trained on more than 25 hours of word-aligned speech, which required a considerable annotation time. We show here that the combination of the three improvement methods outperforms P2FA by more than 8% absolute with a 20 ms tolerance threshold, while remaining fully automatic. It should also be noted that *T&A2* is also shown to be effective on some rare languages, as shown for Faroe and Isizulu. The absolute improvement over *T&A* reaches in fact more than 10% at a 20 ms threshold for Isizulu.

VIII. CONCLUSION

This paper proposed a systematic, step-by-step study of a particular case of HMM-based speech segmentation, i.e. when training directly on the corpus to align. A development database was first used to tune different parameters (e.g. use of context-dependent models and number of training iterations) to obtain our baseline model, Train&Align (*T&A*). An in-depth analysis was then carried out to reveal typical and systematic errors made by *T&A*. This pointed at the bad alignment of silences, vowel to vowel transitions, semi-vowels and plosives. Based on this analysis, we proposed the integration of three types of refinement techniques: *i)* the use of a VAD to get a better initialization of the silence model, *ii)* the exploitation of the time-reversed sound to reduce systematic errors, and *iii)* the addition of features complementary with the conventional MFCCs. The contribution of each of these components was studied separately on a large dataset of 12 speech corpora, varying in speaking style, size and language. All three methods were shown to improve the alignment and the VAD-based technique turned out to be the most advantageous on small databases. The resulting algorithm, integrating the three refinement methods and called *T&A2*, was shown to achieve the highest results. This algorithm was finally compared to state-of-the-art speaker-independent alignment techniques. Across all our experiments, *T&A2* was observed to achieve an improvement, sometimes by a substantial margin. With a tolerance threshold of 20 ms, the absolute improvement over the best existing approach is 11% on a neutral French corpus, and 8% on an expressive English corpus. The improvement over *T&A* is also appreciable, with notably an absolute improvement ranging from 3.5 to 10.5% on two corpora of under-resourced languages.

Beside these improvements, a second aim of the paper was to investigate the role played by the size of the database or by the use of bootstrap data on the performance of HMM-based phonetic alignment. The conclusions we drew in that respect were that: *i)* the performance seems to plateau beyond 5 minutes of training data, which implies that fairly high alignment capabilities are possible for small databases of only a few minutes, *ii)* manually-annotated bootstrapping data can be used to enhance the initialization of the models. Using only 10 seconds of such data generates a relative improvement of 28% at a tolerance threshold of 20 ms and increasing the amount of bootstrapping data does not seem to significantly improve the alignment.

ACKNOWLEDGMENT

The authors are thankful to C. Wellekens for his insightful advice and J.-P Goldman and B. Bigi for their help and enthusiasm regarding this study. They also warmly thank S. Roekhaut, R. Beaufort, and H. Naets for their help and advice for the implementation of the baseline Train&Align alignment and the development of the online platform.

REFERENCES

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE ICASSP*, 1996, pp. 373–376.
[2] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.
[3] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA, USA: MIT Press, 1997.
[4] K. Sjölander, "Wavesurfer - an open-source speech tool," in *Proc. ICSLP*, 2000, pp. 464–467.
[5] P. Boersma and D. Weenink, Praat: doing phonetics by computer (version 5.1.05) [Computer Program]. May 2009, [Online]. Available: http://www.praat.org
[6] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: A professional framework for multimodality research," in *Proc. LREC*, 2006.
[7] H. Kawai and T. Toda, "An evaluation of automatic phone segmentation for concatenative speech synthesis," in *Proc. IEEE ICASSP*, 2004, pp. 677–680.
[8] F. Schiel, A. Kipp, and H. G. Tillman, "Statistical modeling of pronunciation: It's not the model, it's the data," in *Proc. ISCA Modeling Pronunciat. Variat. for Autom. Speech Recogn.*, 1998.
[9] A. Ljolje, J. Hirschberg, and J. van Santen, "ch. Automatic speech segmentation for concatenative inventory selection," in *Progress in Speech Synthesis*. New York, NY, USA: Springer-Verlag, 1997, pp. 305–311.
[10] J.-P. Goldman, "EasyAlign: An automatic phonetic alignment tool under Praat," in *Proc. Interspeech*, 2011 [Online]. Available: http://www.isca-speech.org/archive/interspeech_2011/i11_3233.html

[11] B. Bigi and D. Hirst, "Speech phonetization alignment and syllabification (SPPAS): A tool for the automatic analysis of speech prosody," in *Proc. Speech Prosody*, 2012.
[12] J. Adell, A. Bonafonte, J. A. Gomez, and M. J. Castro, "Comparative study of automatic phone segmentation methods for TTS," in *Proc. IEEE ICASSP*, 2005, pp. 309–312.
[13] D. van Niekerk and E. Barnard, "Phonetic alignment for speech synthesis in under-resourced languages," in *Proc. Interspeech*, 2009.
[14] S. Brognaux, T. Drugman, and R. Beaufort, "Automatic phone alignment. a comparison between speaker-independent models and models trained on the corpus to align," *Lecture Notes in Comput. Sci.*, vol. 7614, pp. 300–311, 2012.
[15] J.-P. Goldman and S. Schwab, "Easyalign spanish: An (semi-) automatic segmentation tool under praat," in *Proc. 5th Congr. de Fontica Experim.*, 2011.
[16] J. P. H. van Santen and R. W. Sproat, "High-accuracy automatic segmentation," in *Proc. Eurospeech*, 1999.
[17] O. Scharenborg, V. Wan, and M. Ernestus, "Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries," *J. Acoust. Soc. Amer.*, vol. 127, no. 2, pp. 1084–1095, 2010.
[18] S. Brognaux, S. Roekhaut, T. Drugman, and R. Beaufort, "Train&Align: A new online tool for automatic phonetic alignments," in *Proc. IEEE Workshop Spoken Lang. Technol. (SLT)*, 2012 [Online]. Available: http://cental.fltr.ucl.ac.be/train_and_align/
[19] M. Wagner, "Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms," in *Proc. IEEE ICASSP*, 1981, pp. 1156–1159.
[20] F. Malfrère and T. Dutoit, "High-quality speech synthesis for phonetic speech segmentation," in *Proc. Eurospeech*, 1997.
[21] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Commun.*, vol. 12, no. 4, pp. 357–370, 1993.
[22] K. Sjölander, "An HMM-based system for automatic segmentation and alignment of speech," in *Proc. Fonetik*, 2003, pp. 93–96.
[23] D. Toledano and L. Gómez, "HMMs for automatic phonetic segmentation," in *Proc. LREC*, 2002.
[24] A. Brandt, "Detecting and estimating parameters jumps using ladder algorithms and likelihood ratio test," in *Proc. IEEE ICASSP*, 1983, pp. 1017–1020.
[25] S. Paulo and L. C. Oliveira, "Automatic phonetic alignment and its confidence measures," in *Proc. 4th Int. Conf. EsTAL*, 2004.
[26] C. Wightman and T. Talkin, "The aligner: Text-to-speech alignment using Markov models," in *Progress in Speech Synthesis*. New York, NY, USA: Springer-Verlag, 1997, pp. 313–323.
[27] L. Chen, Y. Liu, M. Harper, E. Maia, and S. McRoy, "Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus," in *Proc. LREC*, 2004, pp. 759–762.
[28] A. Sethy and S. Narayanan, "Refined speech segmentation for concatenative speech synthesis," in *Proc. ICSLP*, 2002, pp. 149–152.
[29] H. Lo and H. Wang, "Phonetic boundary refinement using support vector machine," in *Proc. IEEE ICASSP*, 2007, pp. 933–936.
[30] K. Demuynck and T. Laureys, "ch. A comparison of different approaches to automatic speech segmentation," in *Text, Speech and Dialogue*. Berlin/Heidelberg, Germany: Springer , 2002, pp. 277–284.
[31] S. Jarifi, D. Pastor, and O. Rosec, "A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis," *Speech Commun.*, vol. 50, no. 1, pp. 67–80, 2008.
[32] S. S. Park and N. S. Kim, "On using multiple models for automatic speech segmentation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 15, no. 8, pp. 2202–2212, Nov. 2007.
[33] A. Katsamanis, M. P. Black, P. G. Georgiou, L. Goldstein, and S. Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. Workshop New Tools Meth. for Very Large Scale Res. Phon. Sci.*, 2011.
[34] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3)*. Cambridge, U.K.: Cambridge Univ. Press, 1995.
[35] J.-P. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Commun.*, vol. 51, pp. 352–368, 2008.
[36] I. Mporas, T. Ganchev, and N. Fakotakis, "Phonetic segmentation using multiple speech features," *Int. J. Speech Technol.*, vol. 11, pp. 73–85, 2008.
[37] M.-B. Wesenick and A. Kipp, "Estimating the quality of phonetic transcriptions and segmentations of speech signals," in *Proc. ICSLP*, 1996.
[38] P. Cosi, D. Falavigna, and M. Omologo, "A preliminary statistical evaluation of manual and automatic segmentation discrepancies," in *Proc. Eurospeech*, 1991, pp. 693–696.
[39] M. A. Pitt, K. Johnson, E. Hume, K. S. , and W. Raymond, "The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," *Speech Commun.*, vol. 45, pp. 89–95, 2005.
[40] A. Ljolje and M. D. Riley, "Automatic segmentation of speech for TTS," in *Proc. Eurospeech*, 1993.
[41] A. Burki, C. Gendrot, G. Gravier, G. Linars, and C. Fougeron, "Alignement automatique et analyse phonétique: Comparaison de différents systèmes pour l'analyse du schwa," *Traitement Autom. des Langues*, vol. 49, no. 3, pp. 165–197, 2008.
[42] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[43] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the Cuidado Project," Inst. de Recherche et Coordination Acoustique/Musique (IRCAM), Tech. Rep., 2004.
[44] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Improved differential phase spectrum processing for formant tracking," in *Proc. ICSLP*, 2004.
[45] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech*, 2011.
[46] T. Drugman, T. Dubuisson, and T. Dutoit, "Phase-based information for voice pathology detection," in *Proc. IEEE ICASSP*, 2011, pp. 4612–4615.
[47] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. ICSLP*, 1996.
[48] S. Brognaux, B. Picart, and T. Drugman, "A new prosody annotation protocol for live sports commentaries," in *Proc. Interspeech*, 2013.
[49] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," in *Proc. Acoust.*, 2008.
[50] V. Colotte and R. Beaufort, "Linguistic features weighting for a text-to-speech system without prosody model," in *Proc. Interspeech*, 2005.

Sandrine Brognaux holds a Master's degree in computational linguistics (UCL, 2010). Her master's thesis was performed in collaboration with Acapela Group SA and focused on prosody annotation for unit-selection speech synthesis. She is currently a Ph.D. student, with an FNRS grant, at the Université catholique de Louvain and Université de Mons (UMons) in Belgium. Her main interests lie in expressive speech, its analysis and HMM-based speech synthesis. She was an Invited Researcher at Columbia University (May 2014–Sep. 2014) and worked in collaboration with Prof. Julia Hirschberg. She was awarded Best Student Paper by ISCA at the Interspeech 2014 conference for her work on the impact of phonetic variation in speech synthesis of various communicative situations.

Thomas Drugman holds an Electrical and Electronics Engineering degree (UMONS, 2007) and a Ph.D. degree (UMONS, 2011). His master thesis was performed at the Swiss Federal Institute of Technologies (EPFL) in Lausanne and dealt with audio-visual speech recognition. He obtained a FNRS grant for pursuing a Ph.D. thesis about glottal analysis and its usefulness in speech processing. He wrote or co-wrote about 90 international scientific publications. He was awarded a Best Student Paper by ISCA at Interspeech 2009, and was among the finalists at Eusipco 2008. His Ph.D. thesis was awarded by IBM Belgium in 2012. He holds a patent on speech synthesis and coding methods. He was an Invited Researcher at the Izmir Institute of Technology (Jun.–Jul. 09) and at the University of California Los Angeles (Apr.–Jun. 10). He was a Postdoctoral Researcher at the University of Mons from June 2011 to January 2014, working on speech analysis/synthesis and on biomedical engineering. He then spent 6 months at Toshiba Cambridge Research Lab, and joined Amazon in Aachen in September 2014.