Text-to-Speech Alignment of Long Recordings Using Universal Phone Models

Sarah Hoffmann, Beat Pfister

Speech Processing Group, ETH Zurich, Switzerland
{hoffmann,pfister}@tik.ee.ethz.ch

Abstract

Large recordings like radio broadcasts or audio books are an interesting additional resource for speech research but often have nothing but an orthographic transcription available in terms of annotation. The alignment between text and speech is an important first step for further processing. Conventionally, this is done by using automatic speech recognition (ASR) on the speech corpus and then aligning recognition result and transcription. This has the drawback that an ASR system needs to be available for the target language. In this paper, we introduce an approach based on forced alignment with hidden Markov models (HMM) normally applied only to shorter utterances. We show that by using a set of generalized phone models computed over phonetic groups, forced alignment is able to reliably align text and speech while being robust against transcription errors. In contrast to ASR methods, the alignment models can be used in a language-independent way.

Index Terms: speech corpora annotation, sentence segmentation, large corpora processing

1. Introduction

Research in speech processing has for a long time concentrated on a few widely spoken languages for the simple reason that only for those languages sufficient speech material was available for statistical processing. The omni-presence of the Internet has changed this situation significantly. A wide range of recordings from TV shows over radio broadcasts to audio books has become available. When compared with previously recorded special-purpose corpora, these resources are larger and cover a much wider set of phenomena and languages. The disadvantage for the speech researcher is that these corpora are insufficiently annotated for further processing. Transcripts are often available but normally without any relation to the speech corpus. An alignment between text and speech is therefore a necessary first step in order to build a corpus usable for further speech processing.

The most common approach to this text-to-speech alignment is to make use of a standard commercial speech recognizer to produce a time-aligned transcription of the recording which is then aligned with the original text transcript on a graphemic level. The idea was introduced by Robert-Ribes in [1] and further refined in recent years ([2], [3], [4]). The alignment in the text domain is much less computationally expensive and allows large deviations between text and speech. Its main disadvantage is the dependency on a high-quality ASR system, which may not be available for the target language. Haubold et al [5] presented a variation that uses a phone recognizer to only detect well-recognizable phones which are then used as landmarks in the subsequent alignment process. The principle is similar to the HMM aligner presented below but in contrast to it the alignment still takes place in the text domain. Although their goal was mainly to align very imprecise transcriptions, it may also be suitable for under-resourced languages. However, they report a much higher error rate than solutions that use a commercial ASR system. Another possible solution to the problem of alignment of data in previously unseen languages are cross-lingual ASR systems, e.g. [6], [7]. Sufficient recognition results are reported only for systems that are adapted to a certain degree to the target language which again poses the problem that additional training material needs to be available.

A rather different method for text-to-speech alignment is the use of hidden Markov models (HMMs) in forced-alignment modus [8]. This has mainly been used in the context of phone segmentation for shorter utterances. The advantage of HMM segmentation is that the models can be trained directly on the data to be segmented, thus no external training material is necessary. However, HMM segmentation works only efficiently if a precise phonetic transcription is available. In the presence of transcription errors, it has been reported to introduce gross misalignments, especially if pruning is applied, which is necessary for large corpora. Prahallad [9] presented a solution that circumvents the pruning by aligning large texts with a modified Viterbi algorithm that allows the speech signal to be longer than the text to align against. Thus the text is aligned in small pieces one at a time.

Here we present a different approach to a purely HMM-based forced alignment of large corpora. We show that a small, generalized set of HMMs can be used as a universal model to align a large corpus in an arbitrary language with sentence granularity. Once this sentence alignment is done, the speech can be split into sentence-like parts and then further segmented using standard HMM segmentation methods. If a flat-start initialization is used then the only external dependency of the entire alignment process is the training data for the sentence alignment HMM set. We will show that these HMMs do not need to be trained in the target language. Therefore, the method is especially interesting for boot-strapping development of speech resources in new languages but is also useful in multi-lingual processing.

This paper focuses on the sentence segmentation of large speech recordings using Viterbi forced alignment. The following section introduces the idea of segmentation with generalized HMMs in more detail. Then the training of a language-independent HMM set is discussed and finally we present experimental results for the sentence segmentation for texts in five different languages.

2. Alignment with Broadly-Trained Models

Forced alignment means that a fixed sequence of HMMs representing the phonetic transcription is time-aligned against the speech signal using the Viterbi algorithm. If the phonetic transcription corresponds exactly to the speech, then the overall

alignment can be expected to be correct, the detailedness of the HMM set will only influence the accuracy of the boundary placement between phones. The more precise the HMMs are the better the accuracy of boundaries.

Precise transcriptions are rarely available for large corpora that have not been annotated specifically for speech processing. If only the text transcription is available, it needs to be converted into a phonetic transcription first, a process which in itself is a source of errors, either because of pronunciation variations introduced by the speaker or because of conversion errors due to ambiguities. In addition, the text itself may contain errors which can reach from simple word confusion to entire sentences being omitted.

If the transcription deviates from the spoken utterance, the alignment is no longer guaranteed to be correct because the erroneous parts of the transcriptions may align with wrong parts of the signal with a high probability. Due to the forced phone sequence, the overall error is normally locally confined if the alignment is computed completely. This, however, is not possible for large speech corpora because the search space of the Viterbi algorithm increases quadratically with the length of the utterance to be aligned. The standard optimization is to employ pruning so that only solutions within a certain distance of the best solution are kept (beam search). Deviating transcriptions may have a very low likelihood locally, thus the right solution is more likely to be pruned prematurely. Very precise HMM sets, such as those used in phone segmentation, make the problem worse because the difference in likelihood between correctly and incorrectly aligned HMMs is much larger. For a more error-resistant alignment a generalized set of HMMs is needed. If the models are more broadly trained, they are more likely to be aligned with partly diverging speech sections without being pruned too early.

Such broadly trained models come at the cost of accuracy and are therefore not usable for precise phone segmentation. They are, however, precise enough to align text and speech on a sentence level. Finding sentence boundaries is a much simpler task because they are normally accompanied by long silences that can be detected relatively reliably. Furthermore, the required precision is much lower as it is sufficient to mark any point within the silence as the boundary. Once the sentence boundaries have been found, the alignment can once more be reduced to a classic phone segmentation problem for which many algorithms for precise boundary placement have been proposed.

Such an HMM alignment is guided by the transcription. As a first step, potential sentence boundaries need to be marked in the text transcript. Then the text is converted into a phonetic transcription which includes the boundaries in form of silences. Finally, the speech is aligned to this phone sequence. Thus, the process presented here takes a somewhat reverse approach with respect to ASR-based alignment methods where the speech is first split on suitable silences and the text is then aligned against the speech segments.

3. Universal HMMs for Alignment

The universal HMM set required for alignment is different from those used in multilingual speech recognition because we do not seek to obtain a precise description of a language. On the contrary, the goal is to construct a minimal HMM set that still allows text and speech to be aligned correctly. The HMMs only have a landmarking function used to identify the characteristic signature of each sentence. Two aspects need to be considered: the choice of the phoneme set and the training material.

3.1. Phoneme Set

While a purely acoustically motivated HMM set would be possible for the alignment task, a model set based on grouping of phones has the advantage that mappings from language-specific phonetic transcriptions to the universal set are straightforward. The choice of phoneme groups has an important effect on the performance of the alignment. On the one side, the phonemes must not be too specific to avoid that the HMMs are trained too narrowly. On the other side, there must be enough differentiation for the phonemes to have the desired landmarking effect. In addition, the clustering should remain language-independent.

In order to estimate the influence of different phoneme classes, we tested phoneme clusterings on the audio book test corpora described in the next section. A set of two classes, vowels and consonants, was considered the minimal configuration, then both classes were split until the alignment was sufficiently exact. Table 1 shows for selected clusterings the maximum deviation from the manually annotated boundaries, which is a good indicator of how well Viterbi is able to keep the alignment. The results indicate that a more fine-grained distinction is more important for vowels than for consonants. This is most likely because a consistent clustering across languages is much easier to find for vowels. Nonetheless, a more fine-grained clustering for consonants seemed to help robustness in the cross-lingual case. For the final models the following set is used:

- vowels: open-front, open-back, neutral, closed, nasal, semi-vowel
- consonants: plosive, aspirant, approximant, 4 fricatives
- pre-plosive pause
- silence (sentence boundary)

3.2. HMM Training

To guarantee the generalizability of the HMMs, they need to be trained on sufficiently diverse training material, both in terms of speaker and language. For the experiments below, we used the recordings from the TIMIT database [10] to ensure speaker-independence together with single speaker corpora in English, German and French (each approx. 1200 sentences by a female professional speaker) for better generalization over languages.

The signals are converted to 16 kHz and segmented using the automatic phone segmentation described in [11]. The resulting HMMs use a feature set of the usual 12 MFCC features and their first derivatives together with energy estimated over a window of 24 ms. The phone models are linear HMMs with 3 states. Only the silence model is handled slightly differently to enforce a minimal length of the sentence boundary and avoid confusion with pre-plosive pauses: it consists of 7 states that

Vowels/Cons.	1/1	2/1	2/4	5/4	5/5	5/8
ger	30.5	30.3	30.3	1.3	1.3	1.3
eng	6.3	6.3	5.9	0.4	0.4	0.4
fra	3.5	3.4	0.8	0.9	0.9	0.3
fin	30.2	30.0	30.0	1.7	1.7	0.3
bul	27.9	23.0	23.9	16.4	16.4	1.4

Table 1: *Maximum deviation (in s) of sentence boundaries from their manually annotated position for different phoneme clusterings. The headings give the number of vowel and consonant groups in the phoneme set.*

text	lang	m/f	parts	boundaries silent	boundaries non-sil.
Hauff Maerchen	ger	f	6	1051	32
Emma	eng	f	3	870	52
Les Miserables	fra	m	3	922	15
Helsinkiin	fin	f	3	776	21
Staroplan. Legendi	bul	m	3	851	9
total			18	4665	129

Table 2: *Audio books from Librivox used for testing alignment: name of book, language, gender of speaker, number of chapters, number of sentence boundaries with and without silence.*

language	boundary errors silent	boundary errors non-sil.	deviation (in ms) silent	deviation (in ms) non-sil.
ger	4 (0.4%)	11 (34%)	1081	489
eng	1 (0.1%)	18 (35%)	354	193
fra	5 (0.5%)	7 (47%)	167	421
fin	1 (0.1%)	4 (19%)	261	76
bul	30 (3.5%)	4 (44%)	277	185
total	41 (0.9%)	44 (34%)	344	292

Table 3: *Results of sentence alignment using unadapted models trained in three languages. Number of wrongly placed boundaries and average deviation in case of error are shown.*

are all tied to one GMM. The HMM training and subsequent alignment were done using the HTK toolkit [12].

4. Evaluation

We tested the sentence segmentation on audio books from the Librivox project[1] in five different languages. The following subsection describes these corpora in more detail before segmentation results are presented in the subsequent section.

4.1. Test Corpora and Phonetic Transcriptions

The Librivox recordings are high-quality, noise-free readings of audio books by single speakers. The texts used by the speakers for reading are made available together with the recording. Therefore, there is very little deviation between text and speech. An analysis by Braunschweiler in [3] showed a word error rate (WER) of 0.61%. We chose five different languages: German, English, French, Bulgarian and Finnish. The first three languages are already contained in the training material. Bulgarian and Finnish are used to test how well the universal phone set performs on languages unseen in the training data. The properties of the texts are listed in table 2. The French and Bulgarian texts are read by male speakers, the remaining texts by female speakers. The recordings are divided in chapters of a length between 7 and 41 min, each chapter is aligned separately.

The forced-alignment algorithm also requires a phonetic transcription for each text, which we produced as follows: for each language a grapheme-to-phoneme (g2p) model was trained using pronunciations extracted from the English version of Wiktionary[2]. These models were then used to convert the text transcripts into phonetic transcriptions using the statistical g2p conversion tool Sequitur G2P [13]. This method has the advantage that it is easily applicable to all five investigated languages. The disadvantage is that the resulting phonetic transcription has a much higher phone error rate (PER) than transcriptions obtained with a dictionary-based approach. When analysing the quality of Wiktionary pronunciations in the context of speech recognition in [14], Schlippe et al found a PER between 3 % and 30 % depending on the language. Thus, the transcription used in the experiments below can be considered a lower baseline on transcription quality. A more elaborate conversion process is bound to reduce the PER and improve the quality of alignments.

Finally, the potential sentence boundaries needed to be determined, for which we used a naive method, placing sentences boundaries where there are final punctuations marks (period, exclamation mark, question mark) in the text. As punctuation is used rather liberally in the texts, this resulted in a number of assumed boundaries that were not actually realised as a pause by the speaker. These boundaries provide a much more difficult case for the alignment algorithm because the annotated silence becomes essentially a phone error in the transcription which, in order to be placed correctly, needs to be aligned exactly with the word boundary. A more elaborate text segmentation might be able to reduce the number of these errors but they cannot be avoided completely when the segmentation is text-driven. We therefore decided to not remove them and evaluate separately how the alignment algorithm performs with respect to these non-silent boundaries.

4.2. Segmentation Results

The alignment for all corpora was done with the same universal HMM set as described in the previous section. The Librivox books have the additional difficulty that each recording contains a preamble and epilogue section with title and copyright information for which no transcription is available. The exact wording varies between books and languages. While it would be feasible to remove these section manually, this would still be time consuming. Therefore, we added an additional skip model to the HMM set, which models speech in general. It was constructed as a 3-state HMM with one shared state whose parameters are computed over all available training material. This model was added at the beginning and end of each chapter, to allow skipping over preamble and epilogue. However, as these sections are relatively long, the pruning needs to be reduced in order to avoid misalignments. We found that a factor of ten is required to correctly skip preambles. Even with the reduced pruning, the chapters could be segmented in well under two minutes on standard PC hardware.

For all recordings, the assumed boundaries were manually annotated as a baseline with the boundary region being marked from the end of the final word in the preceding sentence to the beginning of the first word in the following sentence. Boundary placement by forced alignment was considered correct if the automatically determined boundary fell somewhere within this region. Non-silent boundaries were evaluated in the same way.

Since the main purpose of the alignment is the splitting of the large corpus into sentence-like segments, the most important evaluation measure is the correct placement of the boundaries. Misplaced boundaries will mean that transcription and audio signal are not synchronised correctly, introducing additional errors in the subsequent phone segmentation step. Next to the error rate, we also considered the deviation in case of an error which gives an idea how severe the additional error is. The results for the five test corpora are shown in table 3. For the intra-lingual case, the error rate for the silent boundaries is

[1] http://www.librivox.org

[2] http://www.wiktionary.org, a wiki-based open-content dictionary

between 0.1 and 0.5 %. All boundaries are correct within 1.4 s of the manually annotated boundary. Manual inspection confirmed that the errors are confined to a deviation of one or two words. Thus, the additional errors introduced by the sentences segmentation are well below the WER to be expected from transcription deviations. The error rate for non-silent boundaries is much higher, with an average of 37 % of the boundaries being shifted away from the true boundary. However, the deviation is equally low. That indicates that the errors are only due to the imprecise nature of the HMMs which do perform worse in annotating precise word boundaries, the approximate position is still correctly found. If necessary, the non-silent boundaries can be removed after the alignment process by examining the length of the annotated pauses. If they have the minimal length, it can be assumed that there was no silence present and the boundary can be deleted. For the three intra-lingually aligned texts, 38 boundaries can be eliminated in this way decreasing the total error rate from 1.6 % to 0.3 %.

The two cross-lingual test corpora behaved very differently. The Finnish text showed the best results of all text corpora. Finnish is a very regular language in terms of orthography. Therefore, the PER introduced by the statistical g2p converter are significantly lower than for the other texts, which in turn increases the performance of the HMM forced alignment. The cross-lingual use of the HMMs has a much smaller influence on the quality in this case.

The Bulgarian text performed significantly worse for two reasons. First of all, the corpora by male speakers did worse in general, possibly due to the fact that the training material was unbalanced with respect to gender. More importantly, Bulgarian is phonetically much less related to the training languages than Finnish. It features a rich set of palatalized consonants that are not present in either training language.

In order to compensate for the different phonetics, the HMMs can be further adapted to the specific speech recording. Using the segmentation obtained with the original alignment HMM set, a new model set is computed from the input corpus alone. Then the forced-alignment process is repeated with this new set.

The alignment results for the so adapted models are shown in table 4. With this further step, the forced-alignment performs for Bulgarian as well as for the other languages. The overall error rate over all languages including non-silent boundaries is thus 1.1 %. If non-silent boundary candidates are removed as described above, the error rate is further reduced to 0.3 %. Note, however, that the deviation in case of an error has slightly increased. The adaption of the models not only moved the phone models closer to the language but also increased their precision. This immediately leads to a lower error-tolerance and consequently to a higher likelihood that the alignment process will fail. A possible strategy to counter-act this effect is to include some of the original training material when adapting the models or using a different phoneme clustering better adapted to the language in question.

	boundary errors		deviation (in ms)	
language	silent	non-sil.	silent	non-sil.
ger	2 (0.2%)	10 (31%)	1102	389
eng	2 (0.2%)	15 (29%)	248	260
fra	3 (0.3%)	7 (47%)	340	353
fin	1 (0.1%)	4 (19%)	231	75
bul	3 (0.4%)	5 (56%)	892	345
total	11 (0.2%)	41 (32%)	602	250

Table 4: *Errors in sentence alignment and average deviation in case of error using models adapted for the corpus.*

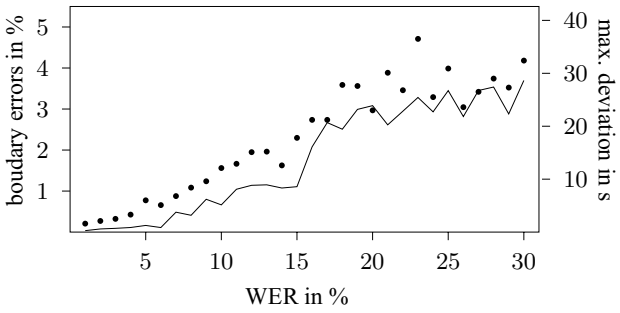

Figure 1: *Stability of HMM alignment with respect to WER. Percentage of misaligned boundaries (dots) and maximum deviation from true boundary (line).*

While the transcriptions used in this paper were very close to the audio recordings, such precise transcriptions may not always be available. To further test the robustness of the alignment, we artificially introduced word errors by adding and deleting words randomly. This corresponds to single word errors made by the speaker. We used the Finnish text for this experiment because it has performed best in the previous tests. Figure 1 shows the numbers of errors and the maximum deviation for silent boundaries for increasing WER. The algorithm is able to produce stable alignments up to approximately 5 % WER. At higher rates the deviation of misplaced boundaries increases significantly causing more than simple word errors when the corpus is segmented. At a WER of 15 %, the HMMs no longer can compensate for errors and alignment is partially lost.

5. Conclusion

In this paper, we showed that HMM forced-alignment is a viable option for the alignment of text and speech for large corpora. In order to be able to handle transcription errors, the alignment needs to be done with phone models that have a much larger coverage than conventionally used phone models. The main advantage of such an approach is that the models can be trained and used in a language-independent manner. While the resulting segmentation is not precise enough to obtain exact word or phoneme boundaries, it can be used to reliably segment the corpus into sentences. Once sentence boundaries are correctly determined, any phone segmentation algorithm can be employed to obtain precise phone boundaries. In [11], we introduced a phonetic segmentation algorithm that relies solely on the audio resource without the need of external training material. Together with the cross-language approach to sentence segmentation it implements a powerful hierarchical approach to precise phone segmentation in many languages, even allowing to automatically boot-strap training material for new languages.

6. Acknowledgements

This work was supported by the Swiss Innovation Promotion Agency CTI.

7. References

[1] J. Robert-Ribes and R. Mukhtar, "Automatic generation of hyperlinks between audio and transcript," in *Eurospeech*, 1997.

[2] P. J. Moreno, C. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *International Conference on Spoken Language Processing*, vol. 8, 1998.

[3] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Interspeech*, 2010, pp. 2222–2225.

[4] O. Boeffard, L. Charonnat, S. L. Maguer, D. Lolive, and G. Vidal, "Towards fully automatic annotation of audiobooks for TTS," in *International Conference on Language Resources and Evaluation*, 2012.

[5] A. Haubold and J. Kender, "Alignment of speech to highly imperfect text transcriptions," in *IEEE International Conference on Multimedia and Expo*, 2007, pp. 224–227.

[6] H. Lin, L. Deng, D. Yu, Y.-f. Gong, A. Acero, and C.-H. Lee, "A study on multilingual acoustic modeling for large vocabulary ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4333–4336.

[7] L. Burget *et al.*, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4334–4337.

[8] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[9] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1444–1449, 2011.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium, Philadelphia*, 1993.

[11] S. Hoffmann and B. Pfister, "Fully automatic segmentation for prosodic speech corpora," in *Interspeech*, 2010, pp. 1389–1392.

[12] S. Young, "The HTK Hidden Markov Model Toolkit: Design and philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, 1994.

[13] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, pp. 434–451, 2008.

[14] T. Schlippe, S. Ochs, and T. Schulz, "Grapheme-to-phoneme model generation for Indo-European languages," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4801–4804.