ALISA: An automatic lightly supervised speech segmentation and alignment tool

ALISA: An Automatic Lightly Supervised Speech Segmentation and Alignment Tool[☆]

A. Stan[a,1], Y. Mamiya[b], J. Yamagishi[a,c], P. Bell[b], O. Watts[b], R.A.J. Clark[b], S. King[b]

[a] *Communications Department, Technical University of Cluj-Napoca, 26-28 George Baritiu St., Cluj-Napoca, 400027, Romania*
[b] *The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom*
[c] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan*

Abstract

This paper describes the ALISA tool, which implements a lightly supervised method for sentence-level alignment of speech with imperfect transcripts. Its intended use is to enable the creation of new speech corpora from a multitude of resources in a language-independent fashion, thus avoiding the need to record or transcribe speech data. The method is designed so that it requires minimum user intervention and expert knowledge, and it is able to align data in languages which employ alphabetic scripts. It comprises a GMM-based voice activity detector and a highly constrained grapheme-based speech aligner. The method is evaluated objectively against a gold standard segmentation and transcription, as well as subjectively through building and testing speech synthesis systems from the retrieved data. Results show that on average, 70% of the original data is correctly aligned, with a word error rate of less than 0.5%. In one case, subjective listening tests show a statistically significant preference for voices built on the gold transcript, but this is small and in other tests, no statistically significant differences between the systems built from the fully supervised training data and the one which uses the proposed method are found.

Keywords: speech segmentation, speech and text alignment, grapheme acoustic models, lightly supervised system, imperfect transcripts.

[☆] This paper is based on our previous work [1, 2, 3], and presents it into a more coherent and thorough manner. Additional results and discussions are presented for the following key aspects: introducing more relaxed confidence measure conditions; grapheme-level acoustic likelihood scores within the confidence measure; unsupervised state tying for the tri-grapheme models; evaluation for an additional speech resource in a different language; and re-evaluation of our results using subjective listening tests for two languages.

Email address: Adriana.Stan@com.utcluj.ro (A. Stan)
[1] Corresponding author

1. Introduction

Over the past decade, speech-enabled applications have progressed to the point where their presence in human-computer interfaces is almost ubiquitous. However, this is true only for the languages for which a sufficient degree of effort has been invested in creating purposely built tools and resources. Any speech-based application requires a large amount of high-quality data and expert knowledge, which are time consuming and computationally expensive to collect.

In this paper we try to alleviate one of the major problems which occurs when migrating a speech-based solution either from one language to another, or from one set of resources to another: speech data preparation. In both speech recognition and speech synthesis, the performance of the resulting system is highly dependent on the quality and amount of training data. But the most widespread method for gathering speech resources nowadays is either by recording a voice talent in a studio or by manually transcribing existing recorded data. However, both of these methods are tedious and usually deter developers from expanding their language and/or speaker environment portfolio.

Accordingly, we turn our attention towards the theoretically unlimited supply of speech data available on the internet, of which the majority is recorded in professional or semi-professional environments – an essential requirement to begin with. Examples of such data include audiobooks, podcasts, video lectures, video blogs, company presentations, news bulletins, etc. These resources are even more appealing when accompanied by an approximate transcript, even though the precise synchronisation between text and audio is not generally available.

The automatic alignment of speech and text has long been studied, and in Section 2 we present some of the most prominent methods. However, our goal is slightly different than theirs, as we aim to rely on no previous knowledge or prepared data, and try to provide a solution for any language with an alphabetic writing system.[2] This is of course highly error prone, but we will show in the Results section that the errors are minimal and negligible both for speech recognition and speech synthesis tasks.

The paper is structured as follows: in Section 2 we describe the state-of-the-art for speech and text alignment methods. Section 3 gives a brief overview of the proposed method, its individual steps being expanded in Sections 4-8. Objective and subjective evaluations of the tool are presented in Sections 9 and 10 respectively. Section 11 provides discussion and concludes the paper.

2. Related Work

Approaches to the task of speech and text alignment can be divided into two major categories: ones where accurate orthographic or phonetic transcripts are

[2]Other types of writing system might be viable, but within this work we do not investigate them.

available, and ones where errors and omissions occur in the transcripts.[3]

Within the first category, the task is simply to determine a direct correspondence between the acoustic units and the text symbols, whether they be sentences, words, letters or phones. This type of method is based either on well-trained pre-existing acoustic and language models, or on dynamic time warping algorithms [4]. One of the challenges presented by this type of approach is the alignment of long audio segments, for which an acoustic model based Viterbi decoder would require a large amount of computational resources. In [5] the authors propose a phonetic alignment for English and French, under Praat using good acoustic models, utterance segmentation, grapheme to phoneme conversion, and manual intervention. [6] proposes a GUI for speech alignment with integrated user feedback for manual checking and tuning, again relying on pre-existing acoustic models and speech segmentation [7] determines a set of anchors within the speech data, and uses a recursive, gradually restrictive language model to recognise the text between two consecutive anchors.

In approaches from the second category, as well as determining the boundaries of sentences or words, the presence of transcription errors must also be taken into account. Due to this fact, most of the proposed methods rely on very good acoustic and language models. For example, [8] uses speaker-independent acoustic models previously trained on over 150 hours of speech data in conjunction with a large, smoothed language model biased towards the text being aligned. [9] uses a phone-level acoustic decoder without any language or phonotactic model and then finds the best match within the phonetic transcripts. This approach was a result of the fact that the data to be aligned may contain a mixture of languages. [10] presents a back-tracking method for Viterbi-based forced alignment and uses good acoustic models. It tries to align the data with 125 words, but it does not detail how to deal with long misalignments of missing audio/text segments. [11] introduces the use of a factor automaton as a highly constrained language model trained on the transcripts. It first divides the audio into smaller segments, and using lower perplexity language models, the reference transcript and the recogniser hypothesis are aligned. The regions where the alignment is less confident are processed with the factor automaton. This method can model word deletions and substitutions. Good Hungarian ASR models are also used in [12] in order to select good data for expressive TTS. [13] proposes a dynamic alignment method to align speech at the sentence-level in the presence of imperfect text data, but cannot deal with phrase reordering within the transcripts. The method of [14] detects vowels and fricatives in speech and text and uses dynamic programming for alignment. [15] tries to correct human-generated transcripts by employing an ASR system and a biased language model built from those transcripts. The algorithm finds word sequences common to the ASR and human-generated transcripts and uses them

[3] Take for example a recording of a trial, for which the typist provides a stenotype. This will most commonly include word deletions, substitutions or insertions, and sometimes additional comments.

as anchors for a finite state transducer-based forced alignment. The final result allows human intervention for the parts where the ASR system generated a different hypothesis. A prominent aspect of methods in the second category is that almost all of them rely on advanced acoustic and language models. This restricts their use to languages where this type of resource is available. Switching to an under-resourced language using these methods would require extensive pre-processing and tuning.

3. An Overview of the Proposed Method

Our method belongs to the second category defined in the previous section's review, but in contrast to the previous work described there, seeks to eliminate the use of expert knowledge and thus address the issue of language-dependency. The method is implemented in the ALISA tool which can be downloaded from `http://simple4all.org/product/alisa/`.

ALISA uses a two-step approach for the task of aligning speech with imperfect transcripts: 1) *sentence-level speech segmentation* and 2) *sentence-level speech and text alignment*. Both processes are fully automated and require as little as 10 minutes of manually labelled speech. The required labelling consists of inter-sentence silence segments for the segmentation, and orthographic transcripts of these sentences for the aligner. This labelled data will be referred to as *initial training data*. This manual labelling process can be easily performed even by non-expert users, as it does not require any speech processing knowledge. Indeed, it does not even presuppose advanced knowledge of the target language: by following the available transcript, and having a basic knowledge of the language's letter to sound rules, the user can easily mark the inter-sentence silence segments within the speech data, and also provide its orthographic transcript.

The segmentation step uses two Gaussian Mixture Models (GMM) trained on speech and silence segments, respectively, and determines a threshold to discriminate between short pauses and silence. For the alignment, we use iteratively self-trained acoustic models and highly restricted word networks built from the available text resource in order to drive a grapheme-level Viterbi decoder over the speech data.

Figure 1 shows a block diagram of the steps involved in the alignment. Each of these steps will be detailed and motivated in the following sections. The method can be applied to any language with an alphabetic writing system, given the availability of a speech resource and its corresponding approximate transcript. In this paper we mainly present results obtained with an English audiobook, but the work flow has been successfully applied to a number of other resources in various other languages [16].

4. GMM-based Sentence-level Speech Segmentation

A common format for the speech data used in automatic speech recognition (ASR) and text-to-speech (TTS) systems is an isolated utterance-length audio

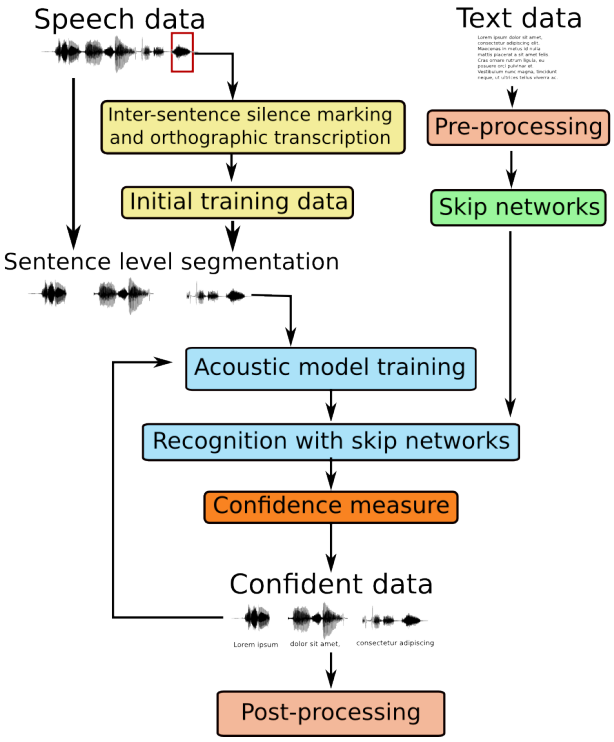


Figure 1: ALISA overview chart

segment and its corresponding transcript, sometimes with additional lower-level labelling. Spoken sentence boundaries can be determined if silent segments of the data can be determined. This is known as end-point detection or Voice Activity Detection (VAD), a well-studied domain [17, 18, 19, 20, 21, 22, 23, 24, 25]. Of the numerous methods and parameters used for such a task, for ALISA we selected an approach based on the Gaussian Mixture Model (GMM). GMMs are a simple statistical model with the necessary power to discriminate between speech and silence. To train the GMMs we use the 10 minutes of initial training data in which inter-sentence silence segments have already been manually marked.[4]. Two 16-component GMMs are then trained, one for speech segments and one for silence segments. The observation vectors consist of energy, 12 MFCCs and their deltas, and the number of zero crossings in a frame. To discriminate between speech and silence, the log likelihood ratio (LLR) of each frame is computed, followed by a moving median filter used for smoothing.

Having marked the silence segments, we then need to discriminate between intra- and inter-sentence silence segments, and discard the former. Short intra-

[4]No distinction is made between inhalation noises, pauses or silence segments. All of them are marked as silence.

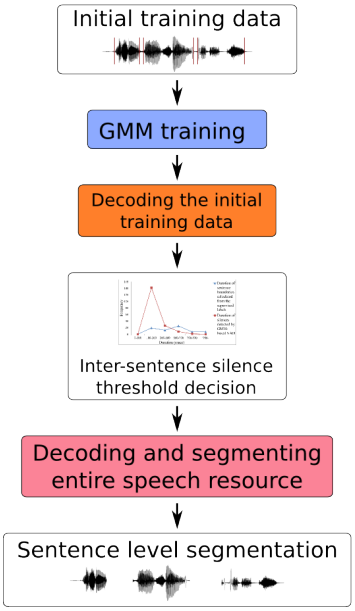


Figure 2: Flow chart of the GMM-based sentence-level segmentation method

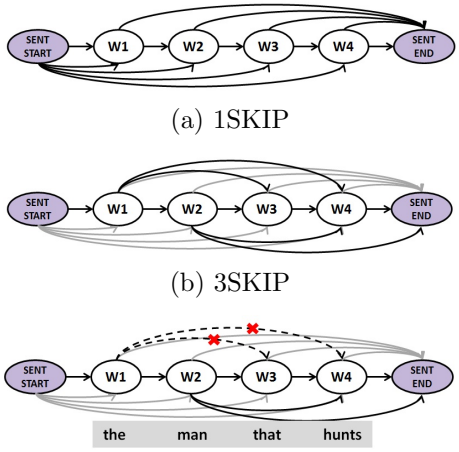
sentence pauses are eliminated by fitting two Gaussian probability distribution functions to histograms of the durations of silence segments annotated in the initial training data: one for intra- and one for inter-sentence silence segments. Their intersection determines the threshold. Using the trained GMMs and this threshold we then go on to segment the entire speech resource. A complete flow chart of the proposed lightly supervised sentence-level segmentation is shown in Figure 2.

It should be noted that although this method is not guaranteed to output fully correct sentence boundaries, the output units are meaningful in terms of spoken language. Individually aligning these units, rather than as a whole sentence, can increase the total duration of the output data. This is due to the fact that most forced-alignment algorithms perform poorly when long silence segments are present within an utterance.

5. Speech Recognition with a Skip Network

In limited vocabulary ASR tasks, or when the approximate transcript of the speech is known, a common practice is to use a biased language model to constrain the decoding process. This has been successfully applied to both TTS [8] and ASR [26] speech database creation, but in both cases the underlying acoustic models were significantly better than the ones we adopt.

We therefore take advantage of the high correlation between speech and text present in the resources, and fit the language model very tightly to the available transcript. As a result, we introduce a highly constrained word network, called a



(a) 1SKIP

(b) 3SKIP

(c) 3SKIP with LM

Figure 3: Word skip network design.

a *skip network*, illustrated in Figure 3. The skip network is used to constrain the Viterbi decoding of each audio segment. The basic model of the network allows the audio segment to be matched anywhere within the transcript, but limits the output to a consecutive sequence of words from it. To minimise the search space, for each utterance we define a text window around the approximate location of the audio. The text window is based on the average speaking rate computed from the length of the speech data and number of words in the transcript. The architecture is similar to that used in [11] for ASR training, but a full FST framework is not required due to the use of text windows. Also, the skip network does not allow insertions, a fact which also limits the errors within the recognised output.

We define two types of skip network which differ only in the number of words they can skip, and therefore the type of audio errors they allow. The *1SKIP network* (Figure 3a) allows the models to match the audio at any starting point within the network, but once this point is determined, they can only skip to the next consecutive word or to the sentence end marker. This type of network helps to approximately identify the location of the utterance within the text window, but it does not account for audio deletions. The *3SKIP network* (Figure 3b) has the same skip-in and skip-out arcs, but it also allows audio deletions of up to 2 words. To ensure a better representation of the text in the word networks, we also used a naïve language model: a list of bigrams generated from the original text which ensures that the arcs which are not in the bigram list are deleted (Figure 3c). These skip networks are used in conjunction with the acoustic models and make the alignment of large amounts of available speech data possible.

Since the publication of our previous work [1], the skip network model has

been applied, with minor modifications, in two other scenarios [27, 28]. Their results show an even higher accuracy of the recognised output when good acoustic models are also available.

6. Acoustic Model Training

In the current scenario, the acoustic model training must take into consideration a series of hypotheses which are not usually met within the conventional training setup. The hypotheses pertain to the lack of a phonetic dictionary, the unknown alignment between speech and text, and the imperfect transcripts. In the following subsections we describe the ideas and methods which make the alignment of large amounts of speech data possible, without the need to fall back on supervised, expert knowledge-based procedures. There are two major parts to the acoustic model training: the first part consists of an iterative grapheme-level acoustic model training procedure, starting from only 10 minutes of orthographically transcribed speech. The second part aims at lowering the sentence error rate of these models by using discriminative methods and tri-grapheme models.

6.1. Lightly supervised ML training with imperfect data

[29] introduces an iterative method for acoustic model training starting from a limited set of transcribed data. This method fits our scenario perfectly, with the exception of requiring a phonetic dictionary. However, several previous studies [30, 31] have shown that for most languages, recognition accuracy does not drop significantly when using graphemes instead of phones.[5] But by combining two highly error-prone methods of training acoustic models, the recognition output might include additional errors which would potentially make the iterative procedure fail. The solution comes from the works of [32, 33, 26, 34, 29, 35, 36]. These studies evaluated the influence of transcript errors present in the training speech database, and concluded that within certain limits, the drop in accuracy is negligible.

Following these previous findings, we considered that a system using incremental self-training, grapheme-based models and relying on imperfect transcripts should be accurate enough for our purposes. We therefore relied on such a system to provide a starting point for the alignment. The full description of the procedure is as follows: the initial training data with manual orthographic transcripts is used to build the initial grapheme models. The initial models are mono-graphemes with 5 states in a left-to-right configuration, 8 mixture components per state, and no state or mixture tying. These models in combination with a highly restricted word network (presented in Section 5) output a preliminary recognition result. The result is then filtered using a confidence measure,

[5] The exception being, of course, English. This is one of the reasons why our evaluations use English as the primary language: what can be achieved on English data should be easily achievable in other languages with simpler grapheme-to-phoneme conversion rules

and the confident data is then used to retrain a new set of acoustic models. In our initial studies we determined that repeating these steps more than once is computationally expensive, and did not give significant improvements in the amount of confident data or its accuracy.

6.2. Lightly supervised MMI training

In conventional maximum likelihood (ML) acoustic model training, the objective function does not take into account the classification decisions made by the acoustic models. This leads to a higher sentence error rate (SER), the score which we are trying to reduce. An alternative is to use discriminative training with the Maximum Mutual Information (MMI) criterion [37, 38]. It can be shown that the expected error rate converges to the model-free expected error rate as the amount of training data increases: in other words, MMI does not require the correct generative model to be used in order to be effective.

But again, for the practical application of MMI training in the low-resource lightly supervised setting there are a series of problems which arise. The first is that as the MMI objective function is based on the difference between numerator and denominator terms and it is important to avoid giving negative weights to the correct utterances through the use of incorrect training labels. This problem has been considered by [39].

Secondly, the denominator lattices are usually built using a weakened version of the language model. But in this scenario, the availability of such a model is highly unlikely for most languages. An alternative is to generate denominator lattices over graphemes, similar to early approaches to discriminative training for ASR, such as [40] where phone-level lattices were used. To test this hypothesis, in Section 9.3 we also compare performance when using word-level denominator lattices. Both word– and grapheme–level training used a bigram language model derived from the original text.

6.3. Tri-grapheme re-estimation and data-driven state tying

One other conventional method for improving the accuracy of the acoustic models is to extend the context of the acoustic units, most commonly through the use of immediate left and right context, i.e. triphones, or in our case *tri-graphemes*. As no phonetic information was available in our setup, the list of tri-graphemes was extracted directly from the text. This did not allow the models to generalise well to out-of-vocabulary words, but it did ensure good performance on the available data.

As in most cases the amount of training data is limited, and so the number of acoustic models whose parameters need to be estimated should be reduced. To achieve this, we modelled only within-word contexts, and data-driven state-tying was employed. This used a decision tree built with questions regarding only the identity of graphemes in a unit's trigrapheme context.

6.4. Full Acoustic Model Training Procedure

As there are several steps involved in the acoustic model training, we re-iterate them, and give a more compact description. Figure 4 presents the full acoustic model training procedure. The seven acoustic models shown in Figure 4 will be referred to using following short identifiers:

① **ML-G** – iterative ML training of mono-grapheme models. There are two ML-trained acoustic models built in this step: the initial ones built on only 10 minutes of manually transcribed training data (**G0-ML**), and the models trained on the confident data obtained from the prior ones (**G1-ML**);

② **G-MMI-GL** – iterative ML training of mono-grapheme models followed by MMI with grapheme-level lattices;

③ **TG-MMI-GL**- iterative ML training of mono-grapheme models followed by MMI with grapheme-level lattices and tri-grapheme re-estimation;

④ **TTG-MMI-GL** – iterative ML training of mono-grapheme models followed by MMI with grapheme-level lattices and unsupervised state-tying tri-grapheme re-estimation;

⑤ **G-MMI-WL** – iterative ML training of mono-grapheme models followed by MMI with word-level lattices;

⑥ **TG-ML** – iterative ML training of mono-grapheme models followed by tri-grapheme re-estimation;

⑦ **TG-MMI-WL** – iterative ML training of mono-grapheme models followed by MMI with word-level lattices and tri-grapheme re-estimation.

The three branches are used for different acoustic model comparisons. The first branch (models ①, ②, ③ and ④) represents the configuration found to achieve best results. Models ② and ⑤ are used to compare the lattice building method for the MMI discriminative training. The third branch (models ⑥ and ⑦) is used to determine whether the discriminative training should be performed before or after the tri-grapheme re-estimation.

7. Confidence Measure

Estimating the correctly recognised speech in an ASR system requires a purpose-designed confidence measure [41]. In our scenario, the transcription errors that might occur during recognition are critical as the resulting data will be used to build other speech-enabled systems. If for speech recognition systems these errors are less important,[6] for speech synthesis [42] showed that above a

[6] We are in fact relying on this to iteratively train our initial acoustic models.

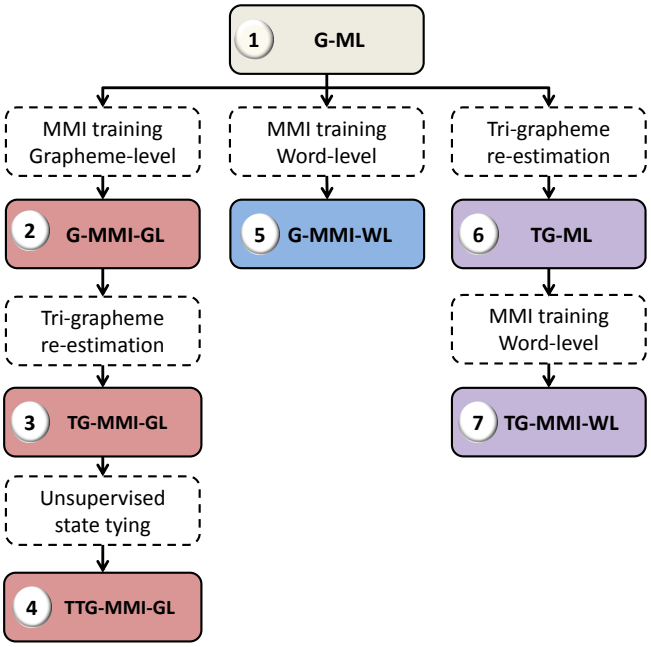


Figure 4: Schematic diagram of acoustic model training steps.

certain threshold, these errors can introduce additional artefacts and produce an unnatural voice.

In standard confidence measures a posterior probability is computed against a general language model [41]. However, the availability of a language model in most languages cannot be assumed. We therefore resorted to our grapheme-level acoustic models and the skip networks to offer some insight into the ranking of confident data. The structure of a 1SKIP network cannot take into account the deletions made by the reader, and therefore cannot determine correctly if these deletions did not actually occur in the speech material. But when allowing the network to skip one or several words (such as in a 3SKIP structure), and the recognised output is the same as the one obtained with the 1SKIP, we can assume that the models are confident enough with regard to the recognised output. This means that the acoustic likelihood scores obtained with 1SKIP and 3SKIP networks should be roughly the same for a confidently recognised utterance.

It is also important to establish the fact that the output is above chance-level, and that the 1- and 3SKIP networks did not enforce the same wrong state path for the models. A so-called *background model* – a single, fully connected, ergodic HMM with five states and eight mixture components per state, trained on all the speech material – can be used to provide a baseline acoustic score. This model is similar to the one used in [43].

As a result, we devised a confidence measure based on several acoustic likelihood scores, as follows:

- **S1** - recognition with 1SKIP network;
- **S2** - recognition with 3SKIP network;
- **S3** - recognition with a background acoustic model;

Confident utterances were considered those for which the following conditions apply:[7]

$$(S1 \approx S2) \ \wedge \ (S1 \ > \ S3) \tag{1}$$

Two additional constraints ensured an even better performance of the confidence measure:

- the length of the recognised utterance has to be above a given number of words, as shorter utterances are more likely to be misrecognised;
- the average likelihood score within each recognised word has to be above a precomputed threshold. This makes it possible to detect potential audio insertions, and especially those at the beginning and end of utterances.

This confidence measure makes it possible to select only the utterances which are most likely to have been correctly recognised, and can therefore either be used to re-estimate the acoustic models, or provide the final aligned data set (see Figure 1).

8. Pre- and Post-Processing

Given the lightly-supervised nature of the entire alignment method, a number of pre- and post-processing steps are required to make best use of the poor acoustic models. These steps are also lightly supervised, and developed to provide as much language independence as possible.

The **pre-processing** included mostly text processing components, such as:

(a) *text cleaning* - all non-alphabetic characters are removed from the original text, and the non-ASCII characters are replaced by two letter substitutes.
(b) *a list of graphemes* – the text is trivially scanned for all distinct alphabetic characters.

[7] Note that as opposed to the confidence measure used in our previous work [1, 3], here we relax the equality condition between the $S1$ and $S2$ acoustic scores, and round them to the first decimal digit. This proved to be a more efficient way to harvest the correctly recognised utterances, and we show in the results section that their accuracy remains within previously stated limits.

(c) *a list of bigrams* – bigrams are used in the word skip network building, so a list of all pairs of consecutive words in the text is extracted. No sentence boundary markers are used.
(d) *a grapheme dictionary* - all distinct words are expanded into their constituent graphemes for use in the speech recognition part.

Post-processing refers to minimising the errors which occur in the selection of confident data, and would therefore limit its usability in various applications:

(a) *sentence boundary correction* - based on the text-based sentence-level segmentation, the recognition output was corrected for sentence-initial and sentence-final word insertions or deletions. Errors involving short words at the beginnings and ends of sentences are the most common type in the recognised output.

(a) *punctuation restoration* - using the original text transcript, we restored punctuation to the recognised output. This ensures the possibility to use the aligned data in TTS systems where punctuation provides important features for predicting pause locations.

9. Objective evaluation

9.1. Data description and preparation

For evaluation of the proposed method, we used a public domain audiobook, *A Tramp Abroad* by Mark Twain[8] and its corresponding text.[9] It contains around 15 hours of speech, uttered by a male speaker and segmented into 50 chapters and 6 appendices. Error rates of the alignment were computed against a sentence-level segmentation of speech supplied for the 2012 Blizzard Challenge.[10] Gold standard (GOLD) transcripts were kindly provided by Toshiba Research Europe Limited, Cambridge Research Laboratory. [8] reports a 5.4% word error rate when comparing the book text with the GOLD transcript.

Additional tests were carried out using a French audiobook, *Candide ou l'optimisme* by Voltaire.[11] It contains around 4 hours of speeech, uttered by a male speaker and segmented into 30 chapters. To derive a gold standard (GOLD) segmentation and transcription, we used the closed captions available on CCProse's Youtube channel.[12] The time alignment was manually checked, and the data was segmented at the sentence-level using this annotation. The word error rate for this audiobook, as computed against the manual transcription is 3.6%.

[8] `http://librivox.org/a-tramp-abroad-by-mark-twain/`
[9] `http://www.gutenberg.org/ebooks/119`
[10] `http://www.synsig.org/index.php/Blizzard_Challenge_2012`
[11] Audio: `http://librivox.org/candide-by-voltaire-2/`
Text: `http://www.gutenberg.org/cache/epub/4650/pg4650.txt`
[12] `http://www.ccprose.com/`

Table 1: Results of the GMM-based sentence-level segmentation. The indices are computed against the GOLD segmentation of the English audiobook.

FEC	**MSC**	**OVER**	**NDS**	**CORR**
0.30%	1.12%	2.05%	0.27%	96.26%

The first 10 minutes of both audiobooks were manually labelled with inter-sentence silence segments. The segmented sentences were then orthographically transcribed. The text window was set to 2800 words for the English audiobook, and 2200 for the French one. The length of the window is based on the average word duration and the length of both audio and text. Initial experiments using average grapheme duration-based length of the text window did not drastically reduce the search space.

9.2. Objective evaluation of the segmentation

The performance of the sentence-level segmentation GMM-based algorithm was compared against the provided segmentation for the English audiobook. The number of correctly detected sentence boundaries is **74%** of the total number of sentences. And although we determined a threshold to discriminate between the short pauses and the sentence boundaries, there were still a number of intra-sentence silence segments which are considered sentence boundaries. But this is inevitable, as the speaking style within the book is highly expressive, and thus the pause duration may vary from chapter to chapter.

To evaluate the GMM-based silence detection algorithm, we resort to the standard measures for end-point detection algorithms [23, 24, 25]. Table 1 shows these results for our method. The measures represent:

- **FEC** - Front End Clipping - speech classified as silence when passing from silence to speech;
- **MSC** - Mid Speech Clipping - speech classified as silence within a speech sequence;
- **OVER** - silence interpreted as speech at the end of a speech segment;
- **NDS** - Noise Detected as Speech - silence interpreted as speech within a silence part.

And CORR is computed as:

$$CORR = 100 - (FEC + MSC + OVER + NDS) \tag{2}$$

In our scenario, the MSC and OVER measures are most relevant due to the use of the segmented data for building other speech-enabled systems.

Table 2: Comparison of sentence (SER) and word error rates (WER) for the *G0-ML* and *G1-ML* acoustic models, when using a language model built from the book text, or the skip networks. For the 3SKIP network we also show the results when the skips in the network are constrained to be valid bigrams within the book text (see details in Section 5). The results are computed for the English audiobook against the GOLD standard transcription.

Language model	**G0-ML**		**G1-ML**	
	SER [%]	WER [%]	SER [%]	WER [%]
Book LM	95.34	46.53	90.12	38.67
1SKIP	22.51	2.98	21.13	3.33
3SKIP	79.51	11.65	62.12	10.45
3SKIP with LM	50.56	5.50	47.20	5.12

9.3. Objective evaluation of the alignment results

The objective analysis of the alignment algorithm was performed against the GOLD transcripts of the data, and using the supervised segmentation. We evaluate all of the main components and steps of the alignment method, and provide results in terms of word error rate (WER) and sentence error rate (SER). The latter is essential when using the data for text-to-speech synthesis systems. All results are reported mainly for the English audiobook, while the French audiobook is used as additional reference for the critical steps of the method.

Starting from the 10 minutes of manually transcribed speech, an initial set of grapheme-level ML acoustic models, *G0-ML*, was trained. The models are mono-grapheme models, having 5 states and 8 mixtures per state with a left-right structure. These models are used in conjunction with the skip-networks to perform a first pass over the data, and thus obtain a first set of alignments. These alignments will in turn be used as training data for a new iteration of the acoustic model building and data alignment, *G1-ML*.

Table 2 shows the error rate improvement when using the skip networks, as opposed to a simple language model built from the book text. It can be seen that given poor acoustic models trained only on 10 minutes of data, the accuracy of the simple LM is very low (i.e. less than 5% sentence-level accuracy). However, when restricting the output to consecutive words within the assigned text window (i.e. using the 1SKIP network) the accuracy increases to over 75%. The 3SKIP network, however, attains an accuracy which is closer to the trigram LM. This low accuracy is much improved when restricting the skips within the network using a bigram list obtained from the book text. In this case, the relative increase in accuracy is over 100%.

Based on these initial results, we expect the confidence measure to be able to extract a percentage of data which is approximately equal to the accuracy of the 3SKIP with LM output. However, slight differences might occur due to the additional conditions imposed for the confident files, such as the number of words within the utterance, or low likelihood scores for long silence segments.

Table 3: Comparison between the percentage of confident utterances, sentence (SER) and word error rates (WER) obtained from the *G0-ML* and *G1-ML* acoustic models. The results are computed for the English audiobook against the GOLD transcript.

Acoustic model	**Percent** [%]	**SER** [%]	**WER** [%]
G0-ML	48.23	12.14	0.74
G1-ML	56.98	11.15	0.58

Table 4: The influence of the amount and quality of training data transcription over the accuracy of the grapheme MMI model. The error rates are computed for the 3SKIP with LM network.

Training Data	**SER** [%]	**WER** [%]
GOLD	54.12	4.86
CONF	56.30	5.94
ALL	57.98	5.97
G0-ML	60.12	6.55

Table 3 shows the percentage of confident data extracted using the *G0-ML* and *G1-ML* acoustic models, as well as its accuracy at sentence and word-level. It can be seen that even with these preliminary acoustic models, but with the help of the skip networks, more than half of the original data can be aligned, its SER and WER being 11% and 0.6%, respectively.

The next step of the ALISA tool consists of discriminative training. However, using imperfect training data with it requires us to establish the influence of the transcription errors on the performance of the acoustic models. To do this, we applied MMI training to initial ML-trained mono-grapheme models using different sets of the reference transcript. Table 4 presents results in four cases: `GOLD` denotes the entire audiobook data with the ground-truth transcript; `CONF` represents only the confident utterances selected by the alignment procedure using the ML mono-grapheme acoustic model (approx. 54% of the data); `ALL` is the entire data with transcriptions obtained using the ML mono-grapheme acoustic model; finally, `G0-ML` represents the recognised output of the initial ML acoustic model.

It may be observed that the MMI training is relatively robust to the use of possibly incorrect transcriptions. Even when using the output of the `G0-ML` acoustic models versus the GOLD transcript, the error rate does not increase significantly. For the following experiments we adopt the use of the `CONF` set as training data.

Also, as described in Section 6.2, there are two possible types of lattice for MMI training, *word* and *grapheme*-level lattices. We evaluated both methods

Table 5: Evaluation of the performance of all acoustic models shown in Figure 4. The results for All data are evaluated using the 3SKIP with LM network.

Acoustic model	All data		Confident data		
	SER [%]	WER [%]	Percent [%]	SER [%]	WER [%]
① **G1-ML**	47.20	5.12	56.98	11.15	0.58
② **G-MMI-GL**	56.30	5.94	71.63	18.74	1.42
③ **TG-MMI-GL**	24.41	2.67	78.55	7.42	0.48
④ **TTG-MMI-GL**	21.82	2.30	79.34	6.83	0.44
⑤ **G-MMI-WL**	57.34	6.00	69.40	19.13	1.59
⑥ **TG-ML**	56.83	5.87	70.34	17.20	1.32
⑦ **TG-MMI-WL**	25.66	2.95	70.37	7.54	0.82

in terms of SER and WER, as well as the amount and quality of the confident data extracted with these models. Table 5 presents this evaluation. The acoustic model numbers are in accordance with those in Figure 4. Although the difference between grapheme and word level lattices in terms of SER and WER for the entire data is minor, the amount of confident data obtained with the `G-MMI-GL` model is higher. This means that the grapheme-level lattices determine the acoustic models to be more confident even when using a 3SKIP with LM network, and thus fulfilling the confidence measure's conditions.

Tri-grapheme re-estimation is also a way to improve the recognition accuracy of the acoustic models. However, it is important to establish if it should be performed before or after the discriminative training, given the fact that the discriminative training might introduce additional errors. Yet when adding context to the models, starting from good baseline models is essential. Therefore, as the numbers in Table 5 show, tri-grapheme re-estimation should be performed after the MMI training (compare the results of `G-MMI-GL` and `TG-ML`). Data-driven state-tying (model `TTG-MMI-GL`) brings even more benefits with a 4% increase in the amount of confident data, and a 40% relative increase over the ML models, for the same levels of error. These results are comparable in WER to those presented in [8], where good acoustic and language models are used to align the data.

To conclude the evaluation of the various steps in ALISA, in Tables 6 and 7 we show an overview of the results obtained with both audiobooks (i.e. English and French), and for each acoustic model that is present in the main processing

Table 6: Error rates for the objective evaluation of the alignment method for the **English** audiobook. Acoustic model numbers correspond to those in Figure 4. The *All Data* results are reported using the 3SKIP with LM network.

Acoustic model	**All data**		**Confident data**		
	SER [%]	WER [%]	Percent [%]	SER [%]	WER [%]
① **G0-ML**	50.56	5.50	48.23	12.14	0.74
① **G1-ML**	47.20	5.12	56.98	11.15	0.58
② **G-MMI-GL**	56.30	5.94	71.63	18.74	1.42
③ **TG-MMI-GL**	24.41	2.67	78.55	7.42	0.48
④ **TTG-MMI-GL**	21.82	2.30	79.34	6.83	0.44

pipeline.[13] It is important to notice that due to its shorter length,[14] the French audiobook achieves lower alignment percentages. However, the trend for each acoustic model is similar. One important thing to notice for the French audiobook is the high relative increase in accuracy of the tied trigrapheme models over the trigrapheme ones (55% relative increase). This is of course a result of the insuficient speech data versus the number of distinct trigraphemes, and this supports the idea that a simple left–right immediate context can have an important effect on the accuracy of the poor grapheme models.

Although we did not provide accurate evaluation of the alignments in our previous study [16], we showed that for various languages the algorithm is able to extract on average 68% of the original data. This data was then used to train speech synthesis systems, and the results showed good intelligibility scores for most languages.[15]

As the reported results in this paper are mostly for an English audiobook, an interesting comparison of ALISA's results is also against a state-of-the-art ASR system, trained in a supervised manner on large quantities of audio data from a different domain. For this purpose, we use a system trained on 150 hours of transcribed TED talks freely available online.[16] The system uses 6-layer deep neural networks, which are used to generate posterior probabilities over approximately 6,000 tied triphone states. These are converted to pseudo-likelihoods for use in a standard HMM-based decoder, using a trigram LM trained on 300M words of TED transcriptions, News Crawl and Gigaword text. Speaker adaptation to the audiobook data was performed using feature-space MLLR. The system is described more fully in [44].

[13] See the first branch in Figure 4.

[14] Only 4 hours, compared to the 15 hours of English

[15] We are now investigating the influence of the language, speaker and recording conditions to determine a correlation between these factors and the system's ratings.

[16] www.ted.com

Table 7: Error rates for the objective evaluation of the alignment method for the **French** audiobook. Acoustic model numbers correspond to those in Figure 4. The *All Data* results are reported using the 3SKIP with LM network.

Acoustic model	**All data**		**Confident data**		
	SER [%]	WER [%]	Percent [%]	SER [%]	WER [%]
① **G0-ML**	79.65	19.29	44.82	27.50	7.75
① **G1-ML**	52.79	6.98	51.66	12.76	0.62
② **G-MMI-GL**	54.32	6.17	53.50	19.50	0.96
③ **TG-MMI-GL**	49.33	10.84	57.51	14.22	0.69
④ **TTG-MMI-GL**	27.46	3.38	65.69	7.22	0.30

Table 8: Error rates of the entire speech data for state-of-the-art ASR system unadapted with a general language mode, state-of-the-art ASR system with adapted acoustic models, state-of-the-art ASR system with adapted acoustic models and biased language model, ALISA and ALISA with a supervised phonetic lexicon. The ALISA results are reported using the 3SKIP with LM network.

System	**SER** [%]	**WER** [%]
ASR unadapted, general LM	78.54	13.89
ASR adapted, general LM	74.25	11.29
ASR adapted, biased LM	24.48	2.18
ALISA	21.82	2.30
ALISA supervised lexicon	18.57	2.12

We compare results from this system with its original LM, and with a LM biased to the audiobook text, as well as a version of ALISA which uses an expert-specified lexicon instead of graphemes. Results for the recognition accuracy of these models are shown in Table 8.

It can be seen that ALISA's WER is similar to that of our best performing standard ASR system, and that ALISA's SER is better. This is a very positive result given the ALISA's lack of dependency on existing resources. For the configuration in which ALISA used an expert-constructed lexicon, the amount of confident data was 81.78% with a SER of 6.92% and WER of 0.55% – an improvement over the grapheme-based system, although the error rates are slightly higher. This was to be expected, especially for English, as the letter-to-phone mapping can be quite complex, and a correct phonetic transcription of the data enables the models to be estimated more accurately. However the small relative improvement does not justify making the use of a lexicon a definite requirement for ALISA.

10. Subjective Evaluation

A previous subjective evaluation of an earlier version of ALISA was published in [2]. The preference listening tests included 3 systems: fully automatic segmentation and alignment; supervised segmentation and automatic alignment; and fully supervised segmentation and alignment. The results showed no definite preference for any of the systems.

Given the changes to ALISA that have taken place since the time of the previous results, we re-ran the listening tests using a similar preference test. As the segmentation algorithm is the same, we only evaluate two conditions: fully automatic alignment and segmentation and fully supervised segmentation and alignment. The test is, however, performed using material in two languages, English and French. Conventional supervised front-ends were used to generate full-context labels for HMM-based speech synthesis training from all transcripts: both the ALISA generated ones as well as the GOLD standard ones. For the English ALISA and GOLD systems we used the English front-end distributed with the Festival system.

For the French ALISA and GOLD systems we used the web-based system described in [45].

The amount of potential training data obtained from the ALISA and GOLD alignments is different, as ALISA presents only the confident utterances for training while the gold standard annotation covers all the data. We wished to focus on the difference in TTS output due to alignment quality; to prevent the differing quantities of data obscuring this, we used the same duration of training data for different systems in either language. For French, this was the duration of ALISA's output after material was held out for testing: 84 minutes. For English we used five hours of data to build each voice. In the first evaluation we ran, the same utterances were used for building the GOLD and ALISA voices – that is, all utterances were from the set marked by ALISA's confidence measure to be correctly annotated, although in the GOLD case, the correct annotation is guaranteed. These subsets are termed *ALISA* and *GOLD-CONF*. A second separate evaluation was run, in which the GOLD voices did not depend on ALISA's confidence measure – the training utterances for these voices were selected at random from the GOLD set until the duration of the selected utterances and the ALISA sets were the same. These new GOLD subsets are termed *GOLD-RAND*. The results of this second evaluation can show whether the confidence measure we have devised has any benefit beyond determining the correctness of transcriptions. For all voices, the speech data were parameterised as described in [46] using the high-quality STRAIGHT vocoder [47]. For all voices, speaker-dependent acoustic models were trained on this parameterised speech data using a speaker-dependent model-building recipe essentially the same as that described in [48].

For testing, the text of 90 medium-length sentences (consisting of between 5 and 31 words) were taken from the held-out GOLD transcripts in each language and used as input to the trained models to produce speech waveforms. For both evaluations (comparing voices built on the *GOLD-CONF* and *GOLD-RAND*

sets with voices built on *ALISA*, respectively), the following procedure was followed. For each language, 10 paid native speakers of the target language were asked to listen to 90 pairs of these stimuli and asked to choose the more natural one. In each of the 90 pairs the same text was synthesised by the GOLD and ALISA systems. Both the ordering of the 90 pairs and the order of systems within each pair was randomised separately for each listener. The listening test was conducted in purpose-built listening booths using high-quality headphones. Different listeners were employed for each of the two evaluations (*GOLD-CONF* and *GOLD-RAND*).

Figure 10 shows the results of the two evaluations. The leftmost bar of the figure shows that there is a small but significant preference for the English system built using the correct transcripts of utterances selected by ALISA over the one using ALISA's transcripts. This is in contrast to earlier results where there was no statistically significant difference between voices built on similar data sets [2]. The second bar, however, shows that the preference is not significantly different from chance (50%) in the case of the French voices. It is possible that the difference in performance between the French and English systems is due to the fact that the French voice was trained on less data and of lower overall quality, whereas the English voice was of sufficient quality for the improved transcript to have a perceptible impact on the quality of its output. However, the two bars on the right of Figure 10 show that ALISA's confidence measure is doing something useful beyond determining whether utterances are correctly transcribed or not. In both English and French, using the set of utterances determined by ALISA's confidence measure results in synthetic speech which is significantly preferred over speech from systems trained using the same amount of correctly transcribed but randomly selected data. This result contrasts with results in ASR, where data selection using a confidence measure can lead to biased estimates and harm performance [49].

A different subjective evaluation of the ALISA tool was presented in [46] where the intelligibility of 5 TTS systems in 5 different languages was evaluated. However, the front-end of those systems was also lightly supervised and built according to the methods presented in [50], which means that the results are not directly comparable to those presented here.

11. Conclusions

Using speech resources available online is an important step in adapting existing speech-based technologies to a multitude of new languages. We have therefore introduced a two-step method for aligning speech with the type of imperfect transcripts often found in online resources. The two steps consist of a GMM-based sentence-level segmentation and an iterative grapheme-level acoustic model building and aligning. The method uses only 10 minutes of manually labelled data, and it is able to extract on average 70% of it, with a word error rate of less than 0.5%. This is possible due to the use of a segmentation model which is built from the original data itself, and the use of an iterative

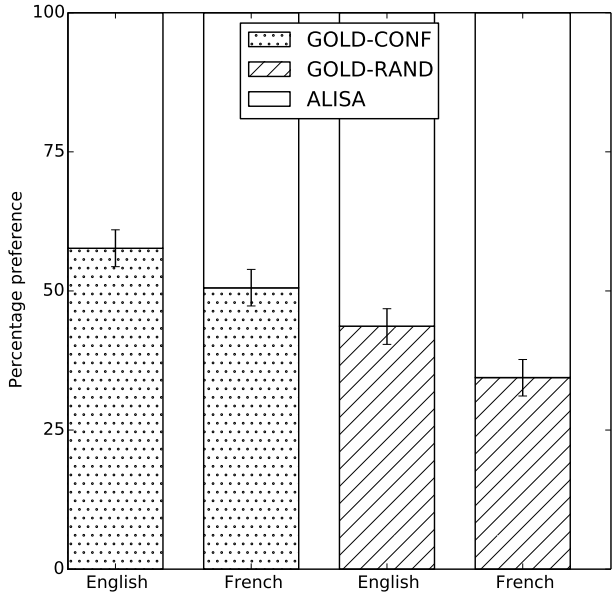


Figure 5: AB preference score results of the listening tests for English and French. ALISA systems use the training data obtained with fully automatic segmentation and alignment. GOLD-CONF systems use the same utterances as ALISA but use a gold-standard segmentation and alignment. GOLD-RAND systems use a random selection of utterances from the gold-standard segmentation and alignment of the same duration as the ALISA systems. 95% confidence intervals computed using a binomial test are shown.

acoustic model training procedure, which makes use of a highly restricted word network, called a *skip network*.

Due to the aim of aligning data in a number of different languages, the lack of expert or prior knowledge when training the acoustic models imposes a series of additional problems, such as defining a confidence measure for the recognised output, deciding between the use of word or grapheme level lattices for the discriminative training, selecting a proper tri-grapheme set, or defining a set of questions for the state tying. All these issues are analysed using an English audiobook, for which the GOLD standard segmentation and transcription are available. Additional results were presented for a French audiobook, but only for the processing pipeline previously found to be best. We also compared our results against a state-of-the-art speech recognition system, as well as a version of our tool which uses a supervised lexicon instead of graphemes. These results showed that ALISA with a supervised lexicon outperforms even our best ASR system.

Subjective listening tests were conducted to determine how much the errors within the aligned dataset influence the quality of a text-to-speech synthesis system as this is the major goal of our work. In one case, our evaluations showed a statistically significant preference for voices built on error-free transcripts,

but this was small and in other tests, no statistically significant preferences for systems built on the supervised segmentation and alignment were found.

In conclusion, the ALISA tool can represent a starting point in building new speech datasets in various languages, starting from found data and with minimal user intervention.

As future work, we would like to investigate the use of non-alphabetic languages, multi-speaker databases, and suboptimal recording conditions of the speech data. One other important aspect would be the full alignment of the speech resource, including spoken insertions and substitutions, as a means of aligning small datasets.

12. Acknowledgement

The GOLD transcripts for *A Tramp Abroad* were kindly provided by Toshiba Research Europe Limited, Cambridge Research Laboratory. The GOLD transcripts for *Candide ou l'optimisme* were provided by CCProse.

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 287678 (Simple4All), as well as by EPSRC EP/I031022/1 (NST) and EP/J002526/1 (CAF).

The work presented here has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF: `http://www.ecdf.ed.ac.uk`). The ECDF is partially supported by the eDIKT initiative (`http://www.edikt.org.uk`).

References

[1] A. Stan, P. Bell, S. King, A Grapheme-based Method for Automatic Alignment of Speech and Text Data, in: Proc. of IEEE Workshop on Spoken Language Technology, Miami, Florida, USA, 2012, pp. 286–290.

[2] Y. Mamiya, J. Yamagishi, O. Watts, R. A. J. Clark, S. King, A. Stan, Lightly Supervised GMM VAD to use Audiobook for Speech Synthesiser, in: Proc. ICASSP, 2013.

[3] A. Stan, P. Bell, J. Yamagishi, S. King, Lightly Supervised Discriminative Training of Grapheme Models for Improved Sentence-level Alignment of Speech and Text Data, in: Proc. of Interspeech, 2013.

[4] X. Anguera, N. Perez, A. Urruela, N. Oliver, Automatic synchronization of electronic and audio books via TTS alignment and silence filtering, in: Proc. of ICME, 2011, pp. 1–6.

[5] J.-P. Goldman, EasyAlign: an automatic phonetic alignment tool under Praat, in: Proc. of Interspeech, 2011, pp. 3233–3236.

[6] C. Cerisara, O. Mella, D. Fohr, JTrans: an open-source software for semi-automatic text-to-speech alignment., in: Proc. of Interspeech, ISCA, 2009, pp. 1823–1826.

[7] P. J. Moreno, C. F. Joerg, J.-M. V. Thong, O. Glickman, A recursive algorithm for the forced alignment of very long audio segments, in: Poc. of ICSLP, 1998.

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

[9] G. Bordel, M. Peñagarikano, L. J. Rodríguez-Fuentes, A. Varona, A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions, in: Proc. of Interspeech, 2012.

[10] K. Prahallad, A. W. Black, Segmentation of Monologues in Audio Books for Building Synthetic Voices, IEEE Transactions on Audio, Speech & Language Processing 19 (5) (2011) 1444–1449.

[11] P. Moreno, C. Alberti, A factor automaton approach for the forced alignment of long speech recordings, in: Proc. of ICASSP, 2009, pp. 4869–4872.

[12] É. Székely, T. G. Csapó, B. Tóth, P. Mihajlik, J. Carson-Berndsen, Synthesizing expressive speech from amateur audiobook recordings, in: Proc. IEEE Workshop on Spoken Language Technology, Miami, Florida, USA, 2012, pp. 297–302.

[13] Y. Tao, X. Li, B. Wu, A dynamic alignment algorithm for imperfect speech and transcript, Comput. Sci. Inf. Syst. 7 (1) (2010) 75–84.

[14] A. Haubold, J. Kender, Alignment of speech to highly imperfect text transcriptions, in: Multimedia and Expo, 2007 IEEE International Conference on, 2007, pp. 224–227. doi:10.1109/ICME.2007.4284627.

[15] T. Hazen, Automatic alignment and error correction of human generated transcripts for long speech recordings, in: Proc. of Interspeech, 2006, pp. 1606–1609.

[16] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. J. Clark, J. Yamagishi, S. King, TUNDRA: A Multilingual Corpus of Found Data for TTS Research Created with Light Supervision, in: Proc. of Interspeech, 2013.

[17] L. Lamel, L. Rabiner, A. Rosenberg, J. Wilpon, An improved endpoint detector for isolated word recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-29 (4) (1981) 777–785.

[18] Z. Xiao-Jing, H. Xu-Chu, C. Hui-Juan, T. Kun, Voice activity detection based on LPCC and spectrum entropy, Telecommunications Engineering 50 (6) (2010) 41–45.

[19] M. Fujimoto, K. Ishizuka, T. Nakatani, A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme, in: Proc. ICASSP, Las Vegas, New Mexico, 2008, pp. 4441–4444.

[20] D. Ying, Y. Yan, J. Dang, F. Soong, Voice activity detection based on an unsupervised learning framework, IEEE Transactions on Audio, Speech, and Language Processing 19 (8) (2011) 2624–2633.

[21] F. Beritelli, S. Casale, A. Cavallaro, A robust voice activity detector for wireless communications using soft computing, IEEE Journal on Selected Areas in Communications 16 (9) (1998) 1818–1829.

[22] J. Ramirez, J. Segura, A. T. C. Benitez, A. Rubio, Efficient voice activity detection algorithms using long-term speech information, Speech Communication 42 (3-4) (2004) 271–287.

[23] M. Henricus, Segmentation, diarization and speech transcription : surprise data unraveled, Ph.D. thesis, University of Twente (2008).

[24] J. Richiardi, A. Drygajlo, Evaluation of speech quality measures for the purpose of speaker verification, in: Proc. Odyssey 2008: The Speaker and Language Recognition Workshop, Stellenbosch, South Africa, 2008, paper 005.

[25] D. Freeman, G. Cosier, C. Southcott, I. Boyd, The voice activity detector for the Pan-European digital cellular mobile telephone service, in: Proc. ICASSP, Vol. 1, Glasgow, UK, 1989, pp. 369–372.

[26] L. Lamel, J.-L. Gauvain, G. Adda, Lightly supervised and unsupervised acoustic model training, Computer Speech and Language 16 (2002) 115–129.

[27] J. Driesen, S. Renals, Lightly supervised automatic subtitling of weather forecasts, in: Proc. Automatic Speech Recognition and Understanding Workshop, Olomouc, Czech Republic, 2013.

[28] N. A. Tomashenko, Y. Y. Khokhlov, Fast algorithm for automatic alignment of speech and imperfect text data, in: Proc. of SPECOM, 2013, pp. 146–153.

[29] S. Novotney, R. M. Schwartz, Analysis of low-resource acoustic model self-training, in: Proc. of Interspeech, 2009, pp. 244–247.

[30] M. Killer, S. Stüker, T. Schultz, Grapheme based speech recognition, in: Proc. of Eurospeech, 2003, pp. 3141–3144.

[31] M. Magimai-Doss, R. Rasipuram, G. Aradilla, H. Bourlard, Grapheme-Based Automatic Speech Recognition Using KL-HMM, in: Proc. of Interspeech, ISCA, 2011, pp. 445–448.

[32] B. Lecouteux, G. Linarès, S. Oger, Integrating imperfect transcripts into speech recognition systems for building high-quality corpora, Computer Speech & Language 26 (2) (2012) 67–89.

[33] M. H. Davel, C. J. van Heerden, N. Kleynhans, E. Barnard, Efficient Harvesting of Internet Audio for Resource-Scarce ASR, in: Proc. of Interspeech, ISCA, 2011, pp. 3153–3156.

[34] C. Fox, T. Hain, Lightly supervised learning from a damaged natural speech corpus, in: Proc. ICASSP 2013, 2013.

[35] M. Alessandrini, G. Biagetti, A. Curzi, C. Turchetti, Semi-Automatic Acoustic Model Generation from Large Unsynchronized Audio and Text Chunks, in: Proc. of Interspeech, 2011, pp. 1681–1684.

[36] K. Yu, M. Gales, L. Wang, P. C. Woodland, Unsupervised training and directed manual transcription for LVCSR, Speech Communication 52 (7-8) (2010) 652–663.

[37] R. Schlüter, H. Ney, Model-based MCE bound to the true Bayes' error, Signal Processing Letters 8 (5).

[38] G. Bouchard, B. Triggs, The trade-off between generative and discriminative classifiers, in: Proceedings of 16th Symposium of IASC Computational Statistics, 2004, pp. 721–728.

[39] H. Chan, P. Woodland, Improving broadcast news transcription by lightly supervised discriminative training, in: Proc. of ICASSP, Vol. 1, 2004, pp. 737–740.

[40] J. Zheng, A. Stolcke, Improved discriminative training using phone lattices, in: Proc. of Interspeech, 2005, pp. 2125–2128.

[41] H. Jiang, Confidence measures for speech recognition: A survey, Speech Communication 45 (4) (2005) 455–470.

[42] J. Ni, H. Kawai, An Investigation of the Impact of Speech Transcript Errors on HMM Voices, in: Proc. of 7th ISCA Workshop on Speech Synthesis, 2010, pp. 246–251.

[43] S. Young, ATK - An Application Toolkit for HTK, Machine Intelligence Laboratory, Cambridge University Engineering Dept, 1st Edition (2004).
URL `http://mi.eng.cam.ac.uk/~{}sjy/ATK_Manual.pdf`

[44] P. Bell, H. Yamamoto, P. Swietojanski, Y. Wu, F. McInnes, C. Hori, S. Renals, A lecture transcription system combining neural network acoustic and language models, in: Proc. Interspeech, 2013.

[45] S. Roekhaut, S. Brognaux, R. Beaufort, T. Dutoit, eLite-HTS: a NLP tool for French HMM-based speech synthesis, in: Interspeech, 2014.

[46] O. Watts, A. Stan, R. Clark, Y. Mamiya, M. Giurgiu, J. Yamagishi, S. King, Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis, in: 8th ISCA Workshop on Speech Synthesis, Barcelona, Spain, 2013, pp. 121–126.

[47] H. Kawahara, I. Masuda-Katsuse, A. de Cheveign, Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based {F0} extraction: Possible role of a repetitive structure in sounds, Speech Communication 27 (34) (1999) 187 – 207.

[48] H. Zen, T. Toda, M. Nakamura, K. Tokuda, Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005, IEICE Trans. Inf. & Syst. E90-D (1) (2007) 325–333.

[49] R. Zhang, A. I. Rudnicky, A new data selection approach for semi-supervised acoustic modeling, in: Proc. of ICASSP, 2006.

[50] O. Watts, Unsupervised Learning for Text-to-Speech Synthesis, Ph.D. thesis, University of Edinburgh (2012).