A GRAPHEME-BASED METHOD FOR AUTOMATIC ALIGNMENT OF SPEECH AND TEXT DATA

Adriana Stan

Communications Department
Technical University of Cluj-Napoca
G. Baritiu 26-28, 400027, Cluj-Napoca
Romania

Peter Bell, Simon King

The Centre for Speech Technology Research
University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
United Kingdom

ABSTRACT

This paper introduces a method for automatic alignment of speech data with unsynchronised, imperfect transcripts, for a domain where no initial acoustic models are available. Using grapheme-based acoustic models, word skip networks and orthographic speech transcripts, we are able to harvest 55% of the speech with a 93% utterance-level accuracy and 99% word accuracy for the produced transcriptions. The work is based on the assumption that there is a high degree of correspondence between the speech and text, and that a full transcription of all of the speech is not required. The method is language independent and the only prior knowledge and resources required are the speech and text transcripts, and a few minor user interventions.

***Index Terms*—** speech alignment, imperfect transcripts, grapheme-based models, word networks

1. INTRODUCTION

One of the most important problems when building text-to-speech synthesis (TTS) or automatic speech recognition (ASR) systems for a new domain or new language is the lack of training data with accurate orthographic transcripts. To avoid the high cost of recording or transcribing speech for the new domain, it is useful to be able to fully exploit readily available data, such as speech with imperfect orthographic text transcriptions. In this category there are, for most of the world's languages, a large number of online resources which could be used, including audiobooks, podcasts or video streams with subtitles. To use such resources, it is necessary to align segments of audio data with the corresponding portion of the transcription. However, for most of these resources, the correspondence between speech and text may be imperfect. This means that, taking the case of an audiobook as an example, the reader might have omitted, inserted or substituted words or sequences of words (with respect to the text), leading to poor performance of forced alignment tools such as the one described in [1]. It is therefore necessary to find a method of alignment which selects only those portions of speech data for which an accurate transcription exists somewhere in the text.

In the field of TTS, there have been several attempts to align speech with unsynchronised or imperfect transcripts. In [2] the authors use 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. Speaker-independent phone-level acoustic models are also used in [3], whilst [4] detects vowels and fricatives in speech and text and uses dynamic programming for alignment.

These previous approaches all have in common a reliance on expert knowledge of the language in question and the existence of suitable acoustic models or clean training corpora; however, these resources are not available for many – in fact, probably most – languages. Therefore it is necessary to develop a means through which any speech resources which *are* available can be exploited in a simple, language-independent manner whilst sidestepping the problem of specially collecting carefully-read speech or manually transcribing speech.

The aim of this work is not necessarily to identify correct transcriptions for all the speech data available, but to jointly select audio data and corresponding transcriptions from a larger set of data, in an unsupervised manner, using no prior knowledge of the language or additional resources. For this task, we propose a "skip network", which is a finite-state network that allows audio segments to be automatically labelled with fragments of the text transcription using only a relatively poor grapheme-based acoustic model. On audiobook data from a single speaker, we demonstrate that this method can correctly transcribe around 55% of the initial speech data.

Our primary objective is to build single-speaker acoustic models for HMM-based TTS, where having clean transcriptions is particularly important [5]. Although our eventual aim is to apply the technique for TTS in under-resourced languages such as Romanian, the results presented in this paper use an English audiobook as the source of text data, primarily to allow benchmarking against previously-reported

978-1-4673-5126-3/12/$31.00 ©2012 IEEE

results. Audiobooks are attractive resources in general since they are readily available in many languages and have an average of 90% correspondence[1] with the book text. The choice of English presents a big challenge for acoustic modelling in a näive manner at the grapheme level [6, 7], since it has particularly weak correspondence between graphemes and phonemes.

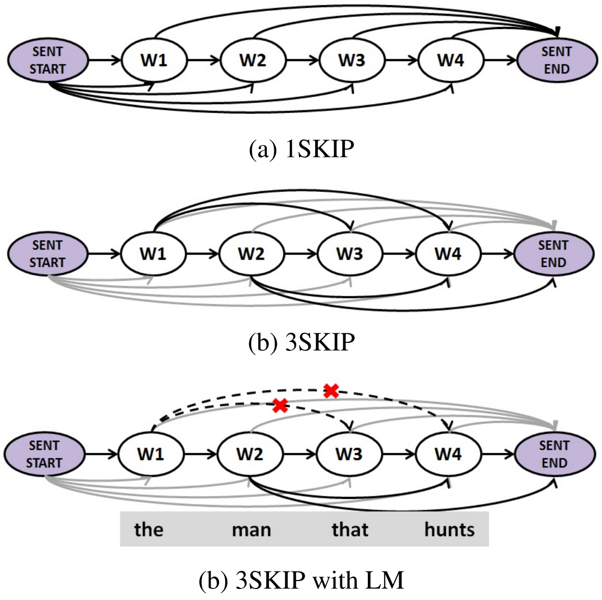
2. SPEECH ALIGNMENT WITH A NOISY TRANSCRIPTION

2.1. Speech alignment with a skip-network

In the case where the text and audio data are imperfectly matched, a common method for alignment is to use a lightly supervised approach, performing speech recognition with a biased language model (LM) primarily estimated from the noisy transcript. The recognition output is then aligned with the transcript using dynamic programming. This technique has been used successfully for both TTS [2] and ASR [8]. However, both these works used existing well-trained acoustic models to perform speech recognition. In our case, the acoustic models available are very weak grapheme-based models obtained from a small initial data set – hence the need to impose tighter constraints from the transcription, whilst retaining the flexibility to exclude incorrect portions.

We therefore propose to use a *skip network*, illustrated in Fig.1, to constrain a Viterbi decoding of each audio segment. This network allows the audio segment to be matched to any point of the transcript, but constrains the output to be a consecutive sequence of words from the transcript: this is a much tighter constraint than that imposed by a n-gram biased LM. Portions of the transcript that are not contained in the audio are automatically removed by this method. We construct networks individually for each utterance by selecting a large text window where the speech utterance is estimated to reside, using a measure of average speaking rate. The architecture is most similar to that used in [9] for ASR training, but the use of text windows to limit computational cost avoids the need for a full FST framework to be used; in addition, the skip network does not allow arbitrary word insertions which would introduce errors in the final transcription. The basic skip network ("1SKIP") does not account for deletions in the audio, with respect to the speech. The alternative "3SKIP" network (Fig.1b) has the same basic structure, but also allows for the speaker to delete sections of 1 or 2 words at a time. This is advantageous for computing confidence measures, discussed below.

To achieve a better representation of the text in the word networks, we also used a naïve language model. A list of bigrams is generated from the original text, and while building the skip networks, arcs which are not in the bigram list are removed (Fig.1c), effectively ensuring that skips occur only when linguistically plausible.



(a) 1SKIP

(b) 3SKIP

(b) 3SKIP with LM

Fig. 1. Word skip networks design.

2.2. Confidence Measure

As previously mentioned, for training TTS systems we desire absolutely correct transcripts. Although the results will show that the skip networks are able to provide accurate transcripts in general, it is necessary to select from these only the utterances with a fully correct transcription. A post-ASR confidence measure is used to identify these correctly aligned utterances. An additional benefit of performing this step is that we obtain a data set with high-accuracy transcripts which can then be used to train an improved set of acoustic models (compared to the initial model set trained on a very small amount of data); these models can in turn be used to re-align the data, increasing the overall harvesting rate.

We propose a confidence measure based on 6 acoustic likelihood scores computed for each utterance: **S1** - recognition with 1SKIP network; **S2** - forced alignment on the 1SKIP network output; **S3** - recognition with 3SKIP network; **S4** - forced alignment on the 3SKIP network output; **S5** - recognition with a background acoustic model; **S6** - forced alignment with the background model. The background model is explained in Section 2.4. The imposed conditions are:

$$(S1 = S3) \wedge (S2 = S4) \wedge (S1 > S2 > S5 > S6) \quad (1)$$

The first two conditions check that both the 1SKIP and 3SKIP networks produced matching word sequences, which means that the speaker did not delete any words with respect

[1] Our informal estimate

to the text. If the speaker did delete words, then we would expect the 3SKIP network to have a different (higher) likelihood than the 1SKIP network. The third condition ensures that the acoustic models have a high score relative to the background model.

Two other conditions are also used to refine the results: the utterance length in words and the average score per state for each word. Aligned utterances with less than a set number of words are discarded, as well as those which contain words with an average acoustic score per state above a set threshold. The utterance length is important, because shorter utterances are more likely to be misrecognised; while a high average score per state can signal an audio insertion.

2.3. Text Processing

Because our method is aimed at obtaining speech with orthographic transcripts with minimal language knowledge, the entire system uses graphemes instead of phonemes. The text is processed in the following stages:

(a) *a list of graphemes* – the text is trivially scanned for all distinct alphabetic characters. Using simple handwritten rules, diacritics are substituted with a 2 character sequence (e.g. in Romanian, *ş* becomes *sh*), while rare characters which do not belong to that language are simply replaced by an in-alphabet one (e.g. in English, *â* becomes *a*)

(b) *a sentence level segmentation* – using simplified punctuation rules, a naïve sentence segmentation can be obtained. This segmentation is used in the selection of the initial training data.

(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 beginning or ending markers are used, because the sentence level segmentation might not be reliable.

(d) *text for initial training data* – a very small set of sentences is selected such that each grapheme has a specified minimum number of occurrences.

2.4. Acoustic Model Training

The initial training text mentioned in (d) above is manually located in the audio, and oracle alignments are produced for it. Using these alignments, grapheme HMMs are trained, following the procedure of [10]. These initial models are five-state, left-to-right, mono-graphemes with eight mixture components per state, and no state tying. In combination with the skip networks, these initial acoustic models are used to find utterances with high-confidence aligned transcriptions from the entire training corpus – this expanded dataset is then used to re-estimate a new acoustic model set, followed by finding more utterances with high-confidence aligned transcriptions.. An overview of the entire process is presented in Fig. 2.

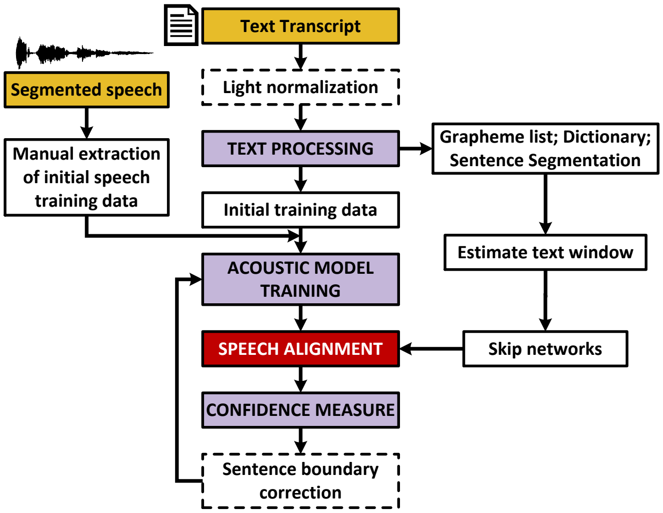


Fig. 2. Flowchart of the training and alignment process.

A background model is needed for use in the confidence measure described in Section 2.2. It consists of a single fully connected ergodic HMM with five states and eight mixture components per state and is trained on all the speech material.

3. EVALUATION AND RESULTS

In order to evaluate the proposed method, we used a public domain audiobook, *A Tramp Abroad* by Mark Twain[2]. It contains around 15 hours of speech, segmented in 50 chapters and 6 appendices. To obtain accuracy results, we used the Blizzard Challenge 2012[3] supplied alignment between text and speech. GOLD transcripts were kindly provided by Toshiba Research Europe Limited, Cambridge Research Laboratory. [2] reports the word and sentence overlap for comparison of GOLD vs. book text.

The text of the book was lightly normalised: all the non-alphabetic characters were discarded from the text; some numbers and frequent abbreviations, such as *Mr., Mrs.* or *Dr.* are expanded; and parts which we suspect the reader might have omitted, such as the table of contents or Gutenberg licence, are stripped out. Because *A Tramp Abroad* contains text in French and German as well, the letters *â, ä, è, é, ê, ô, ö* and *ü* were assigned to *a,e,o,u* from the English alphabet. Using our initial training data selection tool, we obtained 50 utterances together containing at least 30 occurences of each grapheme. The total duration of the selected utterances was 9 minutes.

[2] Text: http://www.gutenberg.org/ebooks/119. Audio: http://librivox.org/a-tramp-abroad-by-mark-twain/.

[3] http://www.synsig.org/index.php/Blizzard_Challenge_2012

Initial models M_0 were trained and speech alignment was performed using them. The text window comprised 2600 words around the estimated location of the utterance within the whole text, and represents the minimum length at which all utterances have their corresponding text in the text window. The total number of words in the lightly normalised book is 155,261, segmented into 7498 utterances. Average word duration was thus estimated to be 0.36 seconds. Because knowing the sentence-level error rate is essential for further use of data, the results are analysed using both sentence error rate (SER) and word error rate (WER) measures.

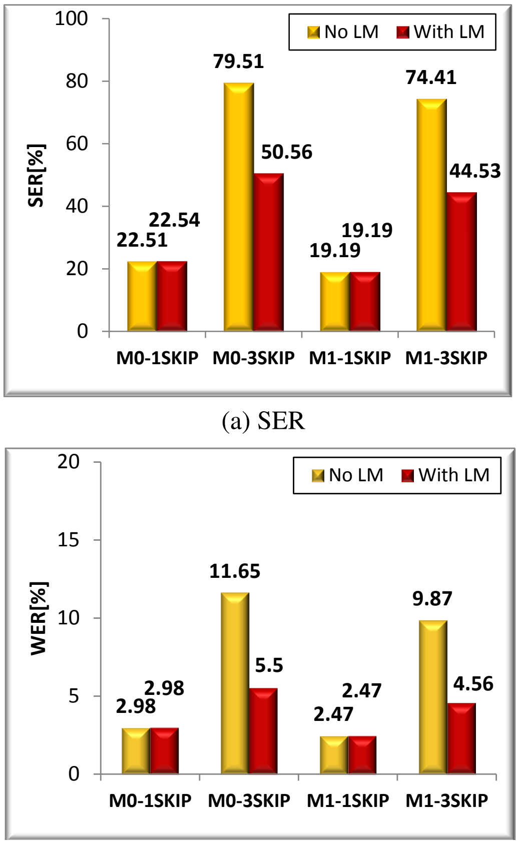
Using the 3574 confident utterances obtained with initial models, we then trained the re-estimated (M_1) models. Any further iterations of realignment and retraining gave only a slight decrease in SER. Fig. 3 summarises the results of the alignment using 1SKIP and 3SKIP networks with both initial (M_0) and re-estimated (M_1) models. When applying the bigram language model, the accuracy results for the 3SKIP network increased substantially. This was to be expected, as many of the earlier errors were a result of erroneous word skipping. The 1SKIP networks results are not affected, because the output is implicitly constrained to contain only bigrams present in the transcription.

The confidence measure was also evaluated with reference to the GOLD transcripts. The error rates of the confident utterances are presented in Fig. 4. The percentage of confident utterances is almost equal to the percentage of accurate ones obtained with a 3SKIP network, which means that the confidence measure works well. Slight differences occur due to the utterance word length constraint. The average SER for the confident utterances is 9.4% with a 0.65% WER. The bigram language model results in a slightly higher (worse) SER but harvests more than twice as many confident utterances.

When using no word length limitation or average acoustic score per state, the SER of the confident files increases by 2.5%. A minimum of 6 words per utterance and a threshold of 8 for the acoustic score per state provided a good balance between the number of confident files and their accuracy.

We observed that a major cause of loss in utterance accuracy is single-word insertions or deletions (particularly of short words) at the beginning or end of utterances. Therefore, given the approximate sentence splitting obtained in the text processing step, we automatically corrected the alignment of the utterances where the audio segmentation differed compared to the text one only in the the inital and final words. This resulted in a 4% decrease in SER. Some utterances which were actually correct after the alignment step could be mistakenly changed by this step if the audio segment boundaries were poorly placed, but an analysis showed that this affected only 1% of the corrected utterances. The final result is that we harvested 55% of the corpus as confident utterances, with a SER of 7% and WER of 0.5%.

Table 1 presents the results when varying the confidence threshold, giving a direct comparison with the results in [2]. Our WER is higher by 0.5% on average for the corresponding proportion of extracted sentences, but note that this has been achieved with no initial acoustic models or dictionary. The bold face row represents the best results achieved when balancing the amount of confident data and the error rates.



(a) SER

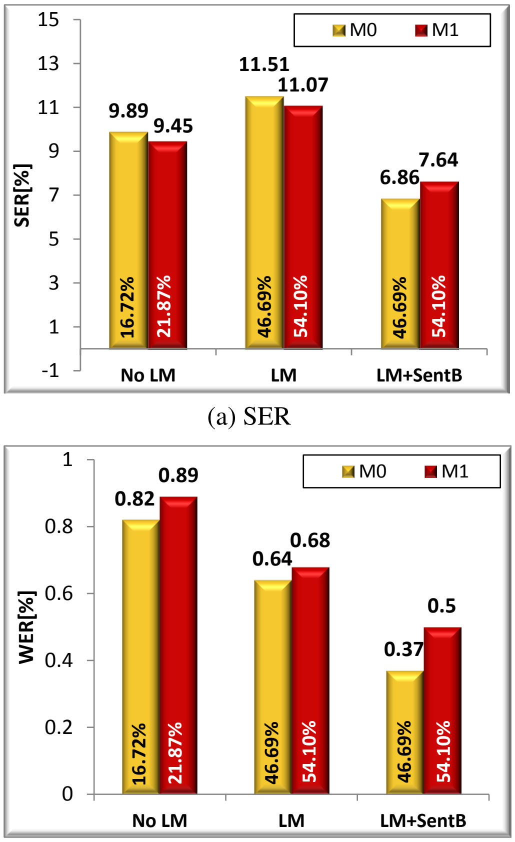
(b) WER

Fig. 3. (a) Sentence (SER) and (b) word error rates (WER) for initial (M_0) and (M_1) re-estimated models using 1SKIP and 3SKIP word networks, with and without a bigram language model. 7498 utterances were analysed.

4. CONCLUSIONS

This paper introduces an innovative method for speech alignment with imperfect text transcripts. The naïve manner in which the entire system is built makes it language independent and suitable for use in any scenario where speech and text are available, but their time alignment is not known. We have provided results that demonstrate good performance on audiobook data: 55% of the original data was harvested and assigned accurate transcriptions, with minimal user intervention.

Future work includes the possible use of tri-grapheme acoustic models, testing the method on more languages and the use of the harvested data for building TTS voices.



(a) SER

(b) WER

Fig. 4. (a) Sentence (SER) and (b) word error rates (WER) for **confident utterances** obtained from initial (M_0) and re-estimated (M_1) models, with (LM) and without a language model (NoLM). LM+SentB represents the accuracy using sentence boundary correction. The rotated numbers on the bars show the percentage of the entire data that has been harvested (higher is better).

5. ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287678. (Simple4All). The GOLD transcripts for *A Tramp Abroad* were very kindly provided by Toshiba Research Europe Limited, Cambridge Research Laboratory. The grapheme-based acoustic model training tools were provided by Oliver Watts.

6. REFERENCES

[1] R.A.J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

Table 1. Varying the confidence threshold changes the amount of sentences extracted (ExtrSent). Here we show SER and WER results for various confidence threshold settings. *No SentBound* refers to results before the sentence boundary correction, and *With SentBound* refers to results after the sentence boundary correction.

	No SentBound		With SentBound	
ExtrSent	SER	WER	SER	WER
100%	44.40%	5.25%	42.14%	5.14%
80%	42.87%	4.20%	40.22%	4.19%
55%	**10.85%**	**0.76%**	**7.42%**	**0.59%**
40%	10.24%	0.84%	6.83%	0.63%
10%	10.27%	1.28%	6.47%	0.97%

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

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

[4] A. Haubold and J.R. Kender, "Alignment of speech to highly imperfect text transcriptions," *CoRR*, vol. abs/cs/0612139, 2006.

[5] J. Ni and 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.

[6] M. Killer, S. Stüker, and T. Schultz, "Grapheme based speech recognition," in *Proc. of EUROSPEECH*, 2003, pp. 3141–3144.

[7] G.K. Anumanchipalli, K. Prahallad, and A.W. Black, "Significance of early tagged contextual graphemes in grapheme based speech synthesis and recognition systems," *Proc. of ICASSP*, pp. 4645–4648, 2008.

[8] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

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

[10] O. Watts, J. Yamagishi, and S. King, "Letter-based speech synthesis," in *Proc. of Speech Synthesis Workshop*, 2010, pp. 317–322.