LIGHTLY SUPERVISED GMM VAD TO USE AUDIOBOOK FOR SPEECH SYNTHESISER

Yoshitaka Mamiya, Junichi Yamagishi, Oliver Watts, Robert A.J. Clark, Simon King, and Adriana Stan[1]

The Centre for Speech Technology Research, University of Edinburgh,
10 Crichton Street, Edinburgh EH8 9AB United Kingdom
[1] Communications Department, Technical University of Cluj-Napoca,
G. Baritiu 26-28, 400027, Cluj-Napoca,Romania

ABSTRACT

Audiobooks have been focused on as promising data for training Text-to-Speech (TTS) systems. However, they usually do not have a correspondence between audio and text data. Moreover, they are usually divided only into chapter units. In practice, we have to make a correspondence of audio and text data before we use them for building TTS synthesisers. However aligning audio and text data is time-consuming and involves manual labor. It also requires persons skilled in speech processing.

Previously, we have proposed to use graphemes for automatically aligning speech and text data. This paper further integrates a lightly supervised voice activity detection (VAD) technique to detect sentence boundaries as a pre-processing step before the grapheme approach. This lightly supervised technique requires time stamps of speech and silence only for the first fifty sentences. Combining those, we can semi-automatically build TTS systems from audiobooks with minimum manual intervention. From subjective evaluations we analyse how the grapheme-based aligner and/or the proposed VAD technique impact the quality of HMM-based speech synthesisers trained on audiobooks.

Index Terms— voice activity detection, lightly supervised, audiobook, HMM-based speech synthesis

1. INTRODUCTION

TTS synthesis is a technique for generating intelligible, natural-sounding artificial speech for a given input text. It has been used widely in various fields including in-car navigation systems, e-book readers, voice-over functions for the visually impaired, and communication aids for the speech impaired. For some applications such as e-book readers and communication aids, TTS systems are required to read out texts with expressivity.

In order to construct such expressive TTS systems, the use of speech data recoded for audiobooks would be one possible solution because speech in audiobooks naturally includes various types of expressive speech. Some of audiobooks are available online for free such as Librivox (http://librivox.org) and they have a much larger amount of speech data available than previously used for speech synthesisers. These aspects of audiobook data have attracted various speech synthesis researchers recently [1–11].

However, the audiobook data is "found" data and was not perfectly designed for building speech synthesisers. For instance, for training the acoustic models of HMM-based speech synthesisers [12] from audiobook data, we need to know the correspondence between audio and text data. In practice we have to align audio and text data manually or automatically before we use them for building TTS synthesisers.

For this purpose, we proposed to use graphemes for automatically aligning speech and text data [9]. This approach automatically aligns a sequence of words corresponding to given audio segments. We have demonstrated that this method can correctly transcribe around 55% of the initial speech data. To use this approach, however, we need to segment a long audio file into smaller segments corresponding preferably to individual sentences beforehand.

Therefore this paper integrates a lightly supervised VAD technique to automatically detect sentence boundaries, that is, *end-silences* (silence at the end of a sentence) as a preprocessing step of the grapheme-based audiobook aligner. This lightly supervised technique requires time stamps of speech and silence only for the first fifty sentences. Combining those, we can semi-automatically build TTS systems from the audiobook data with minimum manual intervention. We perform a subjective evaluation and analyse how the grapheme-based aligner and/or the VAD technique impact the quality of HMM-based speech synthesisers trained on audiobook data.

This paper is organised as follows: Section 2 overviews the grapheme based speech and text aligner and we mention how we integrate the lightly supervised VAD technique to automatically detect sentence boundaries in Section 3. Section 4 shows experimental results and we summarise our findings in Section 5.

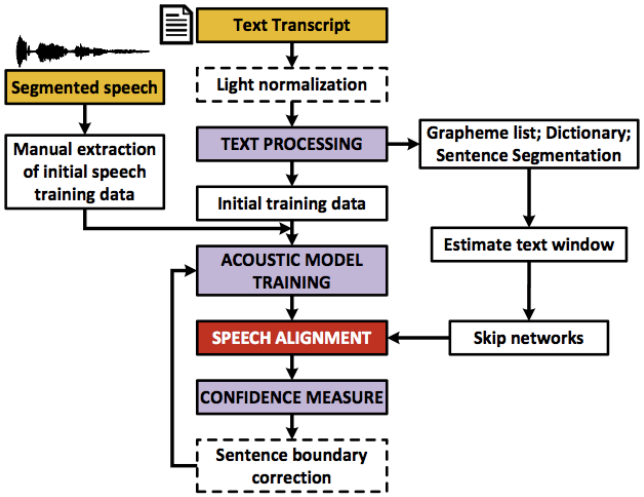


Fig. 1. Overview of the grapheme based speech and text aligner, which we call *AlignTK*.

2. GRAPHEME BASED SPEECH AND TEXT ALIGNER

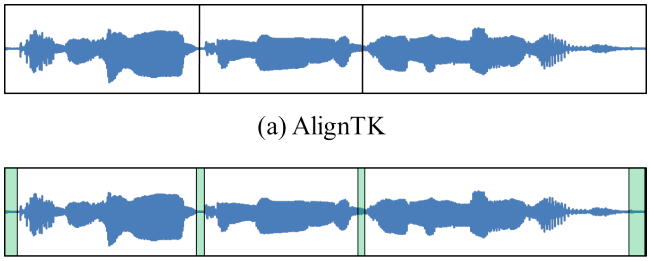
The aim of the audiobook aligner mentioned in this paper is not necessarily to identify correct transcriptions for all the speech data available, but to jointly select audio data and corresponding transcriptions from a larger set of data, in an unsupervised manner, using no prior knowledge of the language or additional resources.

We overview our previous approach that uses a "skip network", which is a finite-state network that allows audio segments to be automatically labelled with fragments of the text transcription based on only a relatively poor grapheme-based acoustic model [9].

The grapheme HMMs are first trained using a small number of initial training sentences manually located in the audio. These initial models are five-state, left-to-right, monographemes with eight mixture components per state, and no state tying. We then construct the skip network that constrains a Viterbi decoding of each audio segment such that it allows the audio segment to be matched to any point of the audiobook text, but constrains the output to be a consecutive sequence of words from the audiobook text.

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. The overview of this approach is shown in Figure 1.

However, in order to use the Viterbi decoding with skip networks for audiobook data, we need in practice to segment a long audio file into smaller segments corresponding preferably to individual sentences beforehand (otherwise it is too long!). In [9], the sentence boundary detection and correction was simply done using standard Voice Activity Detection (VAD) [13–21].



(a) AlignTK

(b) proposed

Fig. 2. Lightly supervised data required for (a) the grapheme based speech and text aligner *AlignTK* and (b) proposed method that additionally has VAD for sentence boundary detection. Green regions show the regions of sentence boundaries marked.

3. A LIGHTLY SUPERVISED GMM VAD TO DETECT SENTENCE BOUNDARIES

3.1. Motivation

In our informal experiments, we have observed that the standard VAD wrongly detected relatively long intra-sentence pauses as sentence boundaries.

This error was critical for prosody modeling: F0 generally becomes lower at the end of sentences. However, if one sentence is split into two sentences at the intra-sentence pause wrongly detected by VAD, F0 values at the end of the first sentence are higher than usual and F0 values at the beginning of the second sentence are lower than usual. This will increase variance of F0 model distributions, resulting in flat and unnatural prosody.

It is also worth mentioning that audiobook data is recorded at different environments (e.g. recording room) in various conditions (e.g. microphone, SNR) and this also makes VAD less accurate.

Therefore we hypothesise that building VAD on audiobook data directly would contribute to the quality of synthetic speech.

3.2. Proposed lightly supervised data

As a lightly supervised technique, the grapheme-based speech and text aligner, which we call *AlignTK* requires the first 50 sentences to be manually and correctly segmented from the long audio file. This is equivalent to providing the time stamps of sentence boundaries of the first 50 sentences as shown in Figure 2 (a).

We propose to slightly extend the lightly supervised data such that we can discriminate the end silences and intra-sentence pauses more accurately. More specifically, we as-

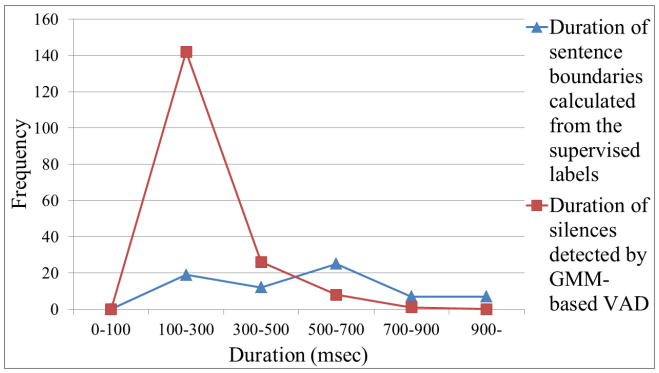


Fig. 3. Histograms of duration of sentence boundaries calculated from the supervised labels and duration of silences detected by GMM-based VAD. As expected, we see that GMM-based VAD detects not only the end silences but also very short pauses frequently.

sume that start and end times of the boundaries of the first 50 sentences are given as shown in Figure 2 (b) where green regions show the regions of sentence boundaries marked.

This additional task is very straightforward: marking the regions of sentence boundaries can be done without any technical knowledge such as pronunciation.

3.3. VAD with Gaussian mixture models

For VAD, we have used the standard Gaussian mixture models (GMM) with 16 mixture. We trained two GMMs, for speech and silence respectively, using the supervised 50 sentences. The observation vectors consist of energy, 12 dimensional MFCC, their deltas and the number of zero crosses in a frame. To judge speech or silence, we calculated the log likelihood ratio (LLR) of each frame, followed by a moving median filter for smoothing.

A key point is to discriminate pauses from sentence boundaries as mentioned above. For this purpose, we first calculate duration of sentence boundary silence of the supervised 50 sentences and fit a Gaussian pdf to the duration distribution of sentence boundary silence. Second we run VAD with the trained GMM on the same 50 sentences, calculate duration of detected non-speech parts, and fit a Gaussian pdf to the duration distribution in a similar way.

Figure 3 shows the histograms of duration of sentence boundary silence calculated from the supervised labels and duration of silences detected by GMM-based VAD. As expected, we see that GMM-based VAD detects not only the end silences but also very short pauses very frequently. To eliminate these short pauses, we calculate the intersection point between the two fitted normal distributions and use the value as a duration threshold for sentence boundary silence on the test data. The flow chart of the proposed lightly supervised VAD is shown Figure 4.

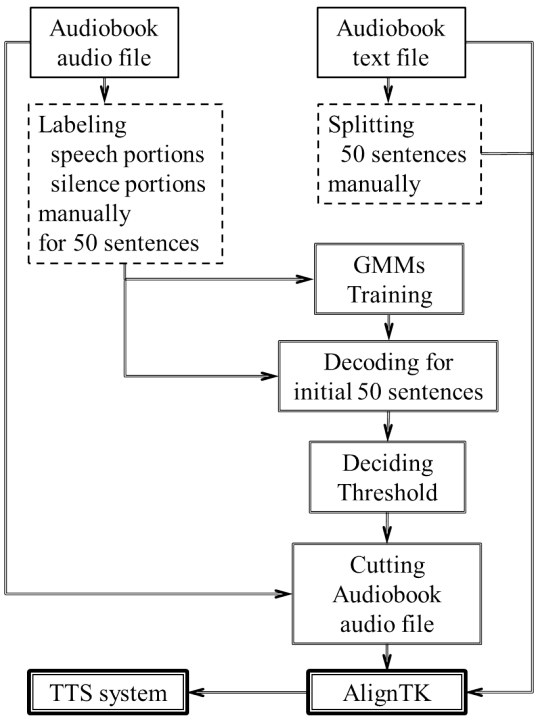


Fig. 4. Flow chart of the proposed lightly supervised VAD system to construct TTS system from audiobook data semi-automatically.

4. EXPERIMENT

In order to analyse how the grapheme-based aligner and/or the VAD technique impact the quality of HMM-based speech synthesisers trained on audiobook data, we performed a subjective listening test. The HMM-based speech synthesisers were trained using an audiobook in public domain,"A Tramp Abroad" by Mark Twain. The details of the HMM-based speech synthesisers can be found in [22]. The audiobook has about 10 hours of speech data uttered by an American male speaker in a relatively quiet environment.

To compare the quality of the TTS systems using the grapheme-based aligner and/or the proposed VAD technique, we used hand aligned labels and hand segmented audio files supplied for the Blizzard Challenge 2012. These "GOLD" transcriptions and audio segments were kindly provided by Toshiba Research Europe Limited, Cambridge Research Laboratory, UK.

4.1. Objective evaluation of VAD

It is important that we confirm the performance of the VAD objectively. Therefore using the GOLD transcripts as references, we assess the performance of the VAD.

The number of VAD detected positions at the correct sentence boundaries was 3764, which is **74 %**. Although we have thresholds to try to discriminate the short pauses and the

Table 1. Indices for evaluating the GMM VAD.

FEC	MSC	OVER	NDS	Correct	SAD error
0.30%	1.12%	2.05%	0.27%	96.26%	4.09%

sentence boundaries, the VAD still incorrectly detected 941 short pauses as sentence boundaries. On the other hand, there was no silence detection within speech regions. And the number of the sentence boundaries where the VAD did not detect was 1354, which is **26 %**.

There are several general indices for evaluating the performance of VAD [19–21]. Table 1 shows results of our VAD measured in the major six indices.

The MSC index represents the ratio of the non-silence portions that the VAD misrecognise as silence, and the OVER index represents the ratio of the silence portions that the VAD misrecognise as non-silence continuously with the correct non-silence portions. Higher values of these indices means that miss-cutting at the short pause portions and miss-connecting at the silence portions occurred. These indices would be the most relevant to the quality of the TTS system. For details of other indices, please refer to [19–21].

4.2. Subjective evaluation of the quality of synthetic speech generated from HMM audiobook TTS systems

In oder to separately asses the impact of the grapheme-based aligner and/or the VAD technique onto the quality of HMM-based speech synthesisers, we constructed three types of HMM-based TTS systems as follows:

A) HMM TTS system constructed from audiobook data where speech and text data are *automatically aligned* using the grapheme approach and the sentence boundaries are *automatically detected* using the proposed lightly supervised GMM VAD.

B) HMM TTS system constructed from audiobook data where speech and text data are *automatically aligned* using the grapheme approach, but, correct sentence boundaries are used to segment audio files.

C) HMM TTS system constructed from audiobook data where speech and text data are correctly aligned and correct sentence boundaries are used, that is, full supervised GOLD topline system.

We adopted a standard preference tests for evaluating the three systems in three combinations, that is A vs B, B vs C, and A vs C. The listening test was done using 90 sentences extracted from "The Adventures of Tom Sawyer", which is a novel by the same author of "A Tramp Abroad". Each combination had 30 pairs of stimuli, and 20 normal-hearing paid subjects listened to 90 pairs each in sound proof booths using headphones and judged which ones they prefer.

The results with 95 % confidence intervals of the listening test are shown in Figure 5. Overall the results obtained are

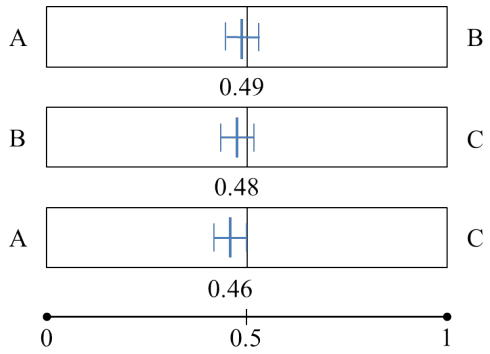


Fig. 5. Preference scores of the listening test.

very good – If we compare A with B, we can see the impact of the proposed lightly supervised automatic sentence-boundary detection compared with fully supervised sentence-boundary detection. This was found to be statistically in significant. If we compare B with C, we can see the impact of the lightly supervised grapheme-based speech/text aligner compared with fully supervised (i.e. manual) speech/text alignment. This was also found to be statistically in significant.

If we compare A with C, we can see the combination effect of these lightly supervised approaches compared with the fully supervised case. The impact of speech/text alignment and sentence-boundary detection seems to be additive. However this was found to be about good as the topline system (only 4% drop in preference score). Overall this is very encouraging for our lightly supervised approaches.

Sound samples synthesised by each of the TTS systems are available at `http://homepages.inf.ed.ac.uk/v1ymamiy/icassp2013_sample.html`.

5. CONCLUSION

This paper integrated a lightly supervised VAD technique to automatically detect sentence boundaries as a pre-processing step for a grapheme-based audiobook aligner. Combining those techniques, we can semi-automatically build TTS systems from audiobook data with minimum manual intervention. We performed a subjective evaluation and analysed how the grapheme-based speech/text aligner and/or the VAD technique impact the quality of HMM-based speech synthesisers. The impacts were additive, but the quality of the semi-automatically built TTS system was at least about the fully supervised system.

6. ACKNOWLEDGEMENT

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), EPSRC EP/I031022/1 (NST) and EP/J002526/1 (CAF)

7. REFERENCES

[1] L. Chen, M.J.F. Gales, V. Wan, J. Latorre, and M. Akamine, "Exploring rich expressive information from audiobook data using cluster adaptive training," in *Proc. Interspeech*, Portland, Oregon, 2012, Tue.P4c.01.

[2] E. Szekely, J.P. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering expressive speech styles in audiobooks using glottal source parameters," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 2409–2412.

[3] K. Prahallad, A.R. Toth, and A.W. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2901–2904.

[4] K. Prahallad and A.W. Black, "Handling large audio files in audio books for building synthetic voices," in *Proc. the 7th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW7)*, Kyoto, Japan, 2010, pp. 148–153.

[5] I. Trancoso, C. Duarte, A. Serralheiro, D. Caseiro, L. Carrico, and C. Viana, "Spoken language technologies applied to digital talking books," in *Proc. Interspeech*, Pittsburgh, Pennsylvania, 2006, pp. 1990–1993.

[6] A. Xavier, P. Nestor, U. Andreu, and O. Nuria, "Automatic synchronization of electronic and audio books via TTS alignment and silence filtering," in *Proc. 2011 IEEE International Conference on Multimedia and Expo*, Barcelona, Spain, 2011, pp. 1–6.

[7] 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, no. 5, pp. 1444–1449, 2011.

[8] E. Szekely, J. Kane, S. Scherer, C. Gobl, and J. Carson-Berndsen, "Detecting a targeted voice style in an audiobook using voice quality features," in *Proc. ICASSP*, Kyoto, Japan, 2012, pp. 4593–4596.

[9] A. Stan, P. Bell, and S. King, "A grapheme-based method for automatic alignment of speech and text data," in *Proc. IEEE Workshop on Spoken Language Technology*, Miami, Florida, 2012, pp. 286–290.

[10] N. Braunscheweiler and S. Buchholz, "Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 1821–1824.

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

[12] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007, pp. 294–299.

[13] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, and J.G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 4, pp. 777–785, 1981.

[14] Z. Xiao-Jing, H. Xu-Chu, C. Hui-Juan, and T. Kun, "Voice activity detection based on LPCC and spectrum entropy," *Telecommunications Engineering*, vol. 50, no. 6, pp. 41–45, 2010.

[15] M. Fujimoto, K. Ishizuka, and 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.

[16] D. Ying, Y. Yan, J. Dang, and F.K. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2624–2633, 2011.

[17] F. Beritelli, S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 9, pp. 1818–1829, 1998.

[18] J. Ramirez, J.C. Segura, A. Torre C. Benitez, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[19] M.A. Henricus, *Segmentation, diarization and speech transcription : surprise data unraveled*, Ph.D. thesis, University of Twente, 2008.

[20] J. Richiardi and 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.

[21] D.K. Freeman, G. Cosier, C.B. Southcott, and I. Boyd, "The voice activity detector for the Pan-European digital cellular mobile telephone service," in *Proc. ICASSP*, Glasgow, UK, 1989, vol. 1, pp. 369–372.

[22] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS System for Blizzard Challenge," in *Proc. Blizzard Challenge 2010*, Sept. 2010.