AUTOMATIC SYNCHRONIZATION OF ELECTRONIC AND AUDIO BOOKS VIA TTS ALIGNMENT AND SILENCE FILTERING

Xavier Anguera, Nestor Perez, Andreu Urruela and Nuria Oliver

Telefonica Research,
Torre Telefonica Diagonal 00,
08019, Barcelona, Spain
{xanguera, andreu, nuriao}@tid.es

ABSTRACT

The e-book industry is starting to flourish due, in part, to the availability of affordable and user-friendly e-book readers. As users are increasingly moving from traditional paper books to e-books, there is an opportunity to *reinvent* and *enhance* their reading experience, for example, by leveraging the multimedia capabilities of these devices in order to turn the act of reading into a real multimedia experience. In this paper, we focus on the augmentation of the written text with its associated audiobook, so that users can listen to the book they are (currently) reading. We propose an audiobook-to-ebook alignment system by applying a Text-to-Speech (TTS)-based text to audio alignment algorithm, and enhance it with a silence filtering algorithm to cope with the difference on reading style between the TTS output and the speakers in the ebook environment. Experiments done using 12 five-minute excerpts of 6 different audio-books (read by men and women) yield usable word alignment errors below 120ms for 90% of the words. Finally, we also show a user interface implementation in the Ipad for synchronized e-book reading while listening to the associated audiobook.

Index Terms— e-book, audiobook, multimodal synchronization, audio processing, TTS alignment

1. INTRODUCTION

We are currently facing a revolution in the media publishing industry with the recent introduction – and adoption – of a wide range of e-book readers and portable devices with e-book reading capabilities (*e.g.* Amazon's Kindle[1] or Apple's iPad[2]). In 2009, device sales more than tripled by the end of the year and content sales increased by 175% [3]. Novel, engaging and appealing ways of presenting multimedia content through these devices need to be investigated in order to attract more readers into using these new technologies [1]. Another important trend in the reading space are audiobooks. The audiobook industry has been growing steadily for more than a decade with an estimated size close to 1 billion USD in 2009[4], and a significant increase on the percentage of downloadable audiobooks. Most of the sales in audiobooks are for unabridged titles (85%), meaning that material is not condensed, and in many cases recorded by professional actors from the original book.

In this paper, we leverage the audio capabilities of most e-book readers and combine audiobooks with e-books in order to transform the reading task into an audio-visual experience. For this task we propose the usage of a TTS-based alignment algorithm and improve it by using a speech-silence intermediate mapping to adapt the TTS output reading style to that of the audiobook recording. In the proposed implementation we carry such alignment offline and generate an alignment file where each word in the e-book is assigned a starting time where it is spoken in the audiobook. The audiobook, the e-book and their associated alignment files are then used by a prototype application we call N'Sink (acronym for Naturally-Spoken Ink) which presents the user with an e-book interface, highlighting the words in the text as they are read by the audiobook.

We envision that N'Sink can be used as a tool to teach how to read, to enhance learning a new language, to help people with reading impairments – including the elderly –, to enable the concept of *personalized audiobooks* where users can create audiobooks for their loved ones that will be listened in-synch with their associated e-book, and to offer users an augmented reading experience. Ultimately, N'Sink could be also used alternatively as an e-book or as an audio-book player depending on the situation the user is in, being able to switch between modalities seamlessly. Note that N'Sink significantly differs from using the TTS engine that is sometimes available in e-book readers, because it combines two sources of the same book (audio and text) where the audio is typically read by an actor, yielding a much more pleasant and less tiring experience than listening to the robotic voice of TTS engines.

[1] www.amazon.com/kindle

[2] www.apple.com/ipad

[3] December 1st 2009, paidcontent.org

[4] www.audiopub.org/2009SalesSurveyRelease.pdf

The remaining of the paper is structured as follows. Section 2 summarizes the different current algorithms available for audio-to-text alignment. Section 3 describes N'Sink's proposed alignment algorithm and user interface. Results of the evaluation of the proposed system against a manually annotated database are presented in Section 4, where we evaluate the accuracy of the produced alignments for 12 five-minute audiobook excerpts. Finally, we draw conclusions and highlight our lines of future work.

2. PREVIOUS WORK

Previously, several works have addressed the challenge of audio and text synchronization in different acoustic domains (*e.g.* broadcast news, parliamentary recordings, music, etc.), following mainly two approaches: (1) *Speech recognition-based systems.* Most proposed algorithms [2, 3, 4, 5] use an automatic speech recognition system to perform the alignment. In [2, 3] a *forced alignment* of the audio to the text is carried out by means of the Viterbi algorithm. Others [4, 5] first run the recognizer to obtain a transcript that is then aligned with the text (text-to-text alignment). These methods have the main advantage of reaching high alignment accuracies (measured usually at the phoneme level), but they require the availability of language-dependent text preprocessing modules and preexistent acoustic models for each language, trained on large quantities of pre-labeled data, and adapted to each individual speaker of the audio to be aligned. Some researchers [6] have shown that the previous constraints can be relaxed by automatically training acoustic models from graphemes (*i.e.* the letters) using unlabeled data. However, such an approach is not applicable in our case as the audio needs to be split in many short sentences in order to train the acoustic models (not possible in our case) and the models are typically highly speaker dependent; (2) *TTS-based systems.* Fewer systems [7, 8, 9] use the output of a TTS synthesis engine to obtain an acoustic signal from the text, and then align the audio with the TTS audio by means of pattern matching techniques (audio-to-audio alignment). These systems still require the knowledge of the language the text is written in, but have the advantage that such knowledge can be condensed inside the TTS engine. In [9] a comparison between TTS-based and speech recognition-based algorithms is performed for phoneme-based labeling, showing that both can achieve fairly good alignment accuracies.

Previously, [10] and others, have proposed the creation of digital talking books by synchronizing audiobooks and ebooks. Unlike our proposal, in their approach they use a speech recognition-based system which has been extensively hand-tuned for the application. They also present a user interface for the alignment, although this is based on a website in a PC setting.

Currently, there are TTS engines for tens of languages through projects such as Mbrola [11] (with over 30 languages freely available) or Festival [12] which makes it a readily and convenient source for the TTS-based alignment. In addition, by using this method, no manually labeled data is necessary and much less tuning of parameters is required. For these reasons, in this paper we follow the TTS-based approach for the alignment of text and audio. Unlike most prior works, our goal is finding the alignment at the *word* level and for *very long acoustic sequences* (of several minutes or more). This differs from previous research, where the alignment is done at the phoneme level and for short sentences, which are easier to bootstrap the system with, if little or no manual alignments are available. On the contrary, when aligning long sequences special care is needed, as localized errors in the alignment may become unrecoverable. In this paper we propose a novel algorithm for doing this, by adding an inner silence filtering step to deal with variable speech/silence distributions in the audio signals considered(audiobook and audio generated by the TTS engine).

3. N'SINK: NATURALLY-SPOKEN INK

The N'Sink system is composed of an offline alignment module and a prototype application for Apple's iPad that highlights the text on the e-book as it is being read in the audiobook. Next we present each of the components in the system prototype.

3.1. Automatic Audiobook to E-book Alignment

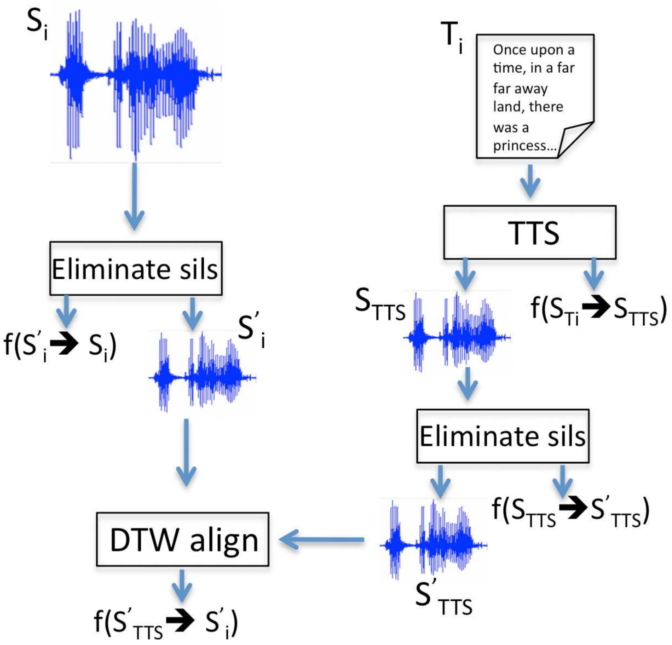


Fig. 1. *Block diagram of the steps involved in audiobook–e-book alignment.*

The goal of the audiobook to e-book alignment algorithm is to generate an alignment file where each word in the text is assigned a corresponding starting time in the audio. Figure 1 shows a block diagram of the steps involved in this alignment process. The input signals are waveform S_i (audiobook) and text T_i (e-book), which are processed independently to create two signals within the same modality – audio in this case – that can be compared to each other. As previously mentioned, we follow a TTS-based alignment approach, similar to [7, 8, 9], such that the text is first converted into an acoustic signal via a TTS engine and then aligned with the audiobook by means of an audio-to-audio alignment system. Unlike previous work, we aim at synchronizing both signals at the word level as our application scenario is an end-user and interactive e-book reading application. In addition, we aim at the alignment of long segments of text with its audio, with no synchronization points in the middle, and we use a novel silence filtering module which brings stability to the alignments. Following Figure 1, the alignment is done in the following steps:

1. Audiobook Processing: The acoustic signal (S_i) is first filtered through a silence detection algorithm that eliminates the segments of the signal with lowest energy (*i.e.* silence regions in a quiet environment), as shown on the left-hand side of Figure 1. Note that audiobooks typically include random amounts of silence in the recording due to interpretive reading, which can later jeopardize the alignment as the TTS waveform usually lacks long pauses. Although more sophisticated methods could be applied, we found that a simple energy thresholding method works well in this case. The short-time energy, obtained over a 200ms sliding window, is first computed for the entire signal. Then all energy values on the lower 1% of the overall energy range are eliminated from the output waveform. In addition, a mapping function $f(S_i^{'} \rightarrow S_i)$ is obtained to map frames from the original signal S_i to the trimmed signal $S_i^{'}$.

2. E-book Processing: The text from the e-book (T_i) is given to a TTS engine as input to produce a speech signal S_{TTS} and a mapping between the text and the signal $f(S_{T_i} \rightarrow S_{TTS})$, as shown on the right-hand side of Figure 1. In our current implementation, we use the Festival TTS engine [12] freely available online[5], using a *multisyn* male English voice [12]. We also apply the silence filtering step to S_{TTS} using the same parameters as above, and obtain an output signal $S_{TTS}^{'}$ and a mapping $f(S_{TTS} \rightarrow S_{TTS}^{'})$.

3. Alignment: Finally, we parameterize $S_{TTS}^{'}$ and $S_i^{'}$ using 10 Mel Frequency Cepstral Coefficients (MFCC) every 20ms and apply a Dynamic Time Warping (DTW) algorithm [13] in order to obtain an alignment between both signals (*i.e.* a mapping $f(S_{TTS}^{'} \rightarrow S_i^{'})$). The DTW has been implemented with standard local constraints, and a diagonal banding global constraint as in [13], with the maximum allowed deviation from the diagonal being set to 5 seconds. Once all steps have been performed we use all computed mapping functions to obtain the corresponding starting time in the audiobook for each word in the e-book text and save it in the output alignment file.

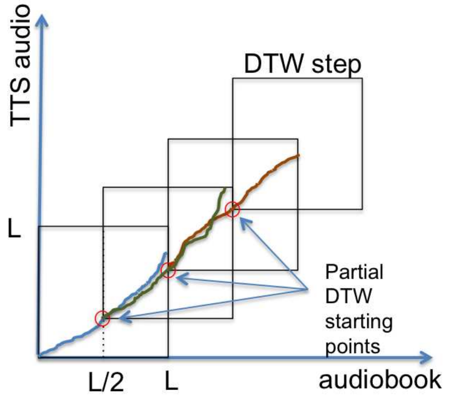


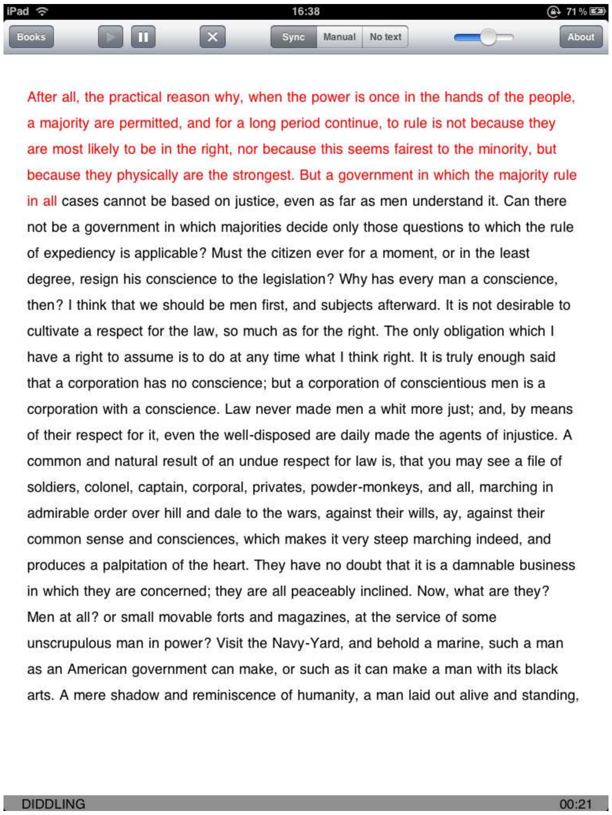
Fig. 2. Illustration of the partial DTW alignments for long files.

The processing complexity to perform the alignment between any amount of text and the corresponding audio is usually of linear complexity for the TTS conversion step and of quadratic complexity in the DTW step. The direct implementation of the above algorithm becomes unfeasible (both for computation and for memory requirements) when the length of the audio is too long (in the order of 15 minutes or more). In order to reduce such complexity we use a partial DTW alignment algorithm as shown in Figure 2. For a given DTW window step of size L (manageable by the system) we compute the DTW alignment setting all DTW constraints except for the ending point (*i.e.* we let the alignment finish in any point $(L, \cdot)$ or $(\cdot, L)$ allowed by the global constraints). Then we select the point $(\frac{L}{2}, DTW(\frac{L}{2}))$ as the new starting point for the new DTW window step, where $DTW(\frac{L}{2})$ is the optimum alignment in the TTS audio for the frame $\frac{L}{2}$ in the audiobook.

3.2. N'Sink's user interface

In order to test the alignment algorithm in a real-life scenario we developed an audio-ebook reading interface called N'Sink (Naturally-Spoken Ink) shown in Figure 3. It has been implemented as an objective-C native iPad/iPhone application. The application allows the user to select among different stored books and four playing modalities (text-only, audio-only, text+audio and synchronized text+audio) through the selection buttons on the upper part of the screen. These modalities can be used by the user to select his reading preferences depending on the context he is in (at home, in the bus, etc.), this way being able to just read the ebook, just listen to it or read it while listening to it. In N'Sink each stored book is composed of a bundle of files, corresponding to the e-book's

[5] www.cstr.ed.ac.uk/projects/festival



After all, the practical reason why, when the power is once in the hands of the people, a majority are permitted, and for a long period continue, to rule is not because they are most likely to be in the right, nor because this seems fairest to the minority, but because they physically are the strongest. But a government in which the majority rule in all cases cannot be based on justice, even as far as men understand it. Can there not be a government in which majorities decide only those questions to which the rule of expediency is applicable? Must the citizen ever for a moment, or in the least degree, resign his conscience to the legislation? Why has every man a conscience, then? I think that we should be men first, and subjects afterward. It is not desirable to cultivate a respect for the law, so much as for the right. The only obligation which I have a right to assume is to do at any time what I think right. It is truly enough said that a corporation has no conscience; but a corporation of conscientious men is a corporation with a conscience. Law never made men a whit more just; and, by means of their respect for it, even the well-disposed are daily made the agents of injustice. A common and natural result of an undue respect for law is, that you may see a file of soldiers, colonel, captain, corporal, privates, powder-monkeys, and all, marching in admirable order over hill and dale to the wars, against their wills, ay, against their common sense and consciences, which makes it very steep marching indeed, and produces a palpitation of the heart. They have no doubt that it is a damnable business in which they are concerned; they are all peaceably inclined. Now, what are they? Men at all? or small movable forts and magazines, at the service of some unscrupulous man in power? Visit the Navy-Yard, and behold a marine, such a man as an American government can make, or such as it can make a man with its black arts. A mere shadow and reminiscence of humanity, a man laid out alive and standing,

Fig. 3. Screenshot of the N'sink application running on an Ipad.

text, the associated audiobook's audio file, the alignment file between text and audio, and a plist file (common in Apple software) that includes information about the book such as title, author, publication date, etc...

When the user selects the synchronized modality, the e-book text on the screen is first shown in dark grey and each word is progressively switched to red as they are read in the audiobook, making use of the synchronization information contained in the previously computed alignment file. When the audio reaches the end of the page, N'Sink automatically changes pages and moves onto the next page on the e-book. Finally, since iPads can be rotated to any orientation, N'Sink's interface can also rotate and reorient its look and text location to adapt the new orientation.

4. EXPERIMENTAL EVALUATION

In this section we evaluate the accuracy of N'Sink's automatic alignment algorithm. We randomly selected 12 five-minute text and audio excerpts chosen from 6 different e-books and their corresponding audiobooks. A summary of each of the excerpt's characteristics is shown in columns 1 to 4 – from left to right – in Table 1, which contain the title and author of the e-book, the number of words in the selected excerpt, the gender of the person reading the book in the audiobook and his/her reading speed (in words per minute). Only the book titled ”*How to win friends...*” was recorded by a professional speaker, while the rest of the audiobook excerpts were read by non-professionals. We opted to include non-professionally read audiobooks in order to test the quality of our alignment algorithm with varying speaking rates and dictions, as professional speakers tend to become very standardized in how understandably and how fast they speak. The variability in speed is shown in the fourth column of the Table, where the reader's speed of ”The adventure of the ...” is significantly faster than the rest. Additionally, we balanced the selection of male and female readers in order to analyze the effect of gender in the proposed system (third column of the Table). Finally, The last 3 columns in the Table summarize the alignment results obtained by the proposed algorithm, as explained below.

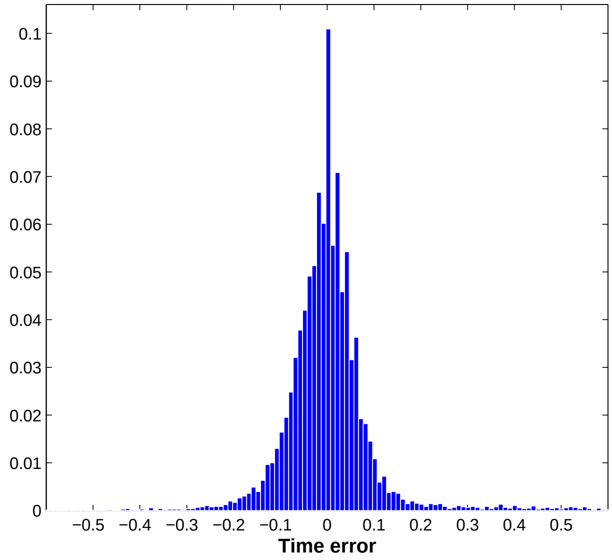


Fig. 4. Normalized histogram of errors $e[n]$ between ground-truth and automatic alignments.

The ground-truth alignment for each file was manually generated indicating the time when each word in the e-book text was produced in the audiobook. We then compared this ground-truth with the output of the automatic alignment algorithm under two conditions: using the silence filtering step and not using it. For each excerpt and condition, we computed the error $e[n]$ between the start time in the ground-truth alignment $t_{gt}[n]$ and the one estimated automatically $t_a[n]$, computed for every word n in the e-book: $e[n] = t_{gt}[n] - t_a[n]$. In Figure 4 we show the normalized histogram of $e[n]$ values for all evaluated excerpts. We observe that most errors concentrate around 0, with only a small percentage with absolute values greater than 0.2 seconds, as seen in Table 1 for each excerpt individually. In addition, in some cases there are outliers in the histogram corresponding to points where the

Table 1. Summary of aligned excerpts and alignment results.

Title—author	# words	gender	reading speed	$\mu_{\|e\|}$ / error@90%	% silence	$\mu_{\|e\|}$ / error@90%
			(words/second)	no filter (in ms.)	TTS / human	with sil. filter (in ms.)
The adventure of the dancing men–A. C. Doyle	1091	male	220.6	75.8/152	2.6/2.2	63.8/128.2
The adventure of the dancing men–A. C. Doyle	1181	male	247.5	117.4/299.3	2.2/5.2	56.7/112.6
The empire of the ants–H. G. Wells	853	male	172.2	79/161	3.0/7.4	63.0/123.2
The empire of the ants–H. G. Wells	800	male	153.1	611.7/2000	2.4/11.0	**65.1/127.7**
Civil disobedience–H. D. Thoreau	806	female	167.1	102.6/284	2.4/8.9	75.4/118.4
Civil disobedience–H. D. Thoreau	890	female	174.0	97.5/330	3.1/11.7	50.3/80.0
Diddling–E. A. Poe	839	female	162.7	138.8/403.8	3.0/13.5	78.2/135.7
Diddling–E. A. Poe	842	female	170.7	490.6/1758.1	2.5/16.9	100.7/157.3
How to win friends & influence people–D. Carnegie	781	male	155.1	209.3/618.6	1.4/9.8	72.7/142.0
How to win friends & influence people–D. Carnegie	758	male	153.1	51.7/94.3	1.4/10.7	48.1/86.7
How the first letter was written–R. Kipling	921	female	184.6	137.2/410.1	2.5/18.0	80.1/117.0
How the first letter was written–R. Kipling	839	female	165.2	288.2/863.8	2.0/18.3	71.7/130.1
Average	883.4	n/a	177.1	200.0/591.1	2.4/11.1	68.8/121.4

alignment gets lost, although this usually happens very few times and after some time the alignment algorithm is able to regain alignment.

Based on the error $e[n]$ we report two performance metrics in Table 1: (1) The absolute error average $\mu_{|e|} = \frac{1}{N}\sum_{n=1}^{N}|e[n]|$ in ms., where N is the total number of words in the excerpt and $|\cdot|$ indicates the absolute value; and (2) the alignment error (in ms.) at 90% percentile (*i.e.* only 10% of the values are larger or equal to the error). The results with both metrics are shown in columns 5 and 7 of Table 1, depending on whether the silence filtering is used or not. Note that although prior work also uses these metrics to evaluate the alignment, the results are not directly comparable as those are computed on a per phone basis on generally short sentences, while we are computing them per word and from an alignment performed on several minutes of speech aligned all at once. Furthermore, the errors reported here have been computed using manually generated ground truth data, which in itself contains a non-quantified labeling accuracy variability. In addition, column 6 shows the percentage of frames that are labeled as silence by the silence filtering algorithm both in the TTS and audiobook audio signals. As expected, the TTS audio contains very stable values and small percentage of silence frames (2.4% in average), while the audiobook's audio exhibits larger variability and larger absolute values (11.1% in average), as they highly depend on the reader's style, the e-book's genre, etc.

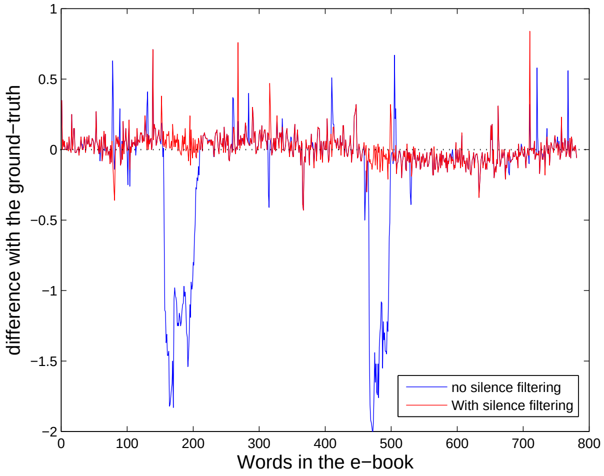


Fig. 5. Comparison of alignment errors depending on the use of a silence filtering.

Next, we focus on results regarding the use of the silence filtering step. Note the significant improvement in $\mu_{|e|}$ for all tested excerpts when using silence filtering (particularly noticeable on the fourth excerpt). Additionally, the alignment error at 90% percentile is also much smaller. We consider that these error values are reasonable for the intended application (only half a syllable in average, considering that an average syllable lasts around 200ms). We carried out preliminary informal tests of the N'Sink prototype with a few users and concluded that generally they were not able to perceive any alignment errors and hence considered the alignment to be perfect in most cases. We plan to carry out a more extensive and formal user study in future work. Regarding the effect of the TTS voice used (male speaker) on the alignment with respect to the audiobook reader (male or female), we obtain similar average errors for both conditions ($\mu_{|e|}$ is 61.56 for male and 76.06 for female, while the error@90% is 120.06 for male and 123.08 for female), highlighting the robustness of the proposed system to speaker variability.

To further analyze the effect of silence filtering on the output, Figure 5 shows the alignment error $e[\cdot]$ for each word in the first excerpt from the e-book *"How to win friends..."*, both with and without silence filtering. In both cases we observe three kinds of alignment errors: (1) A high frequency ripple with errors around or lower than 100ms that we attribute to the inevitable variability of the manual labeling of the ground-truth data; (2) Low frequency deviations from 0 in the error signal, indicating the errors that the automatic system makes when deciding on a particular alignment path along the DTW cost matrix; and (3) Outlier error peaks corresponding to particular words loose their alignment (*i.e.* the DTW algorithm got lost). From the three types of errors the most dangerous one are the third kind as it can cause long term or permanent alignment loss. The development of alignment algorithms must try to reduce these outliers, and when they occur, allow the alignment to get back in synch as fast as possible so that the error perceived by the user gets minimized. Note how the silence filtering step presented in this paper can eliminate most of these outliers and when they occur the DTW algorithm can soon regain synchrony. On the contrary, without silence filtering, the system gets lost twice (for some time) in this example, although it is able to get realigned at some later point.

5. CONCLUSIONS AND FUTURE WORK

With the foreseen explosion of the e-book industry, there is an opportunity to *reinvent* and *enhance* the user e-book reading experience. In this paper, we present the N'Sink (Naturally-Spoken Ink) system that is composed of an offline alignment algorithm and a prototype user interface. The e-book/audiobook alignment algorithm is based on audio-to-audio alignment of the audiobook with the audio derived via text-to-speech from the e-book text. We propose a novel silence filtering algorithm that significantly improves the alignment by eliminating the variable amounts of silence in both audios due to their differences in speaking styles, and describe how we are able to align long audio excerpts using sightly modified dynamic programming techniques. The user interface for the application is built as an iPad prototype to illustrate the concept of audio-augmented e-books, where the text on the e-book is highlighted as it is being read by the audiobook. Although standard text-to-speech synthesis is able to perform this task, N'Sink allows users to listen to the original audiobook, which is typically read by professional (human) speakers, instead of using the robotic-sounding synthesized speech from a TTS engine. Moreover, the N'Sink application opens the path towards *personalized audiobooks* where users can create audiobooks for their loved ones that will be listened to in-synch with their associated e-book. Future work includes testing our alignment strategy with multiple languages, carrying out a user study to understand the pros and cons of this technology from a user-centric perspective, and increasing the efficiency of the alignment algorithm to allow for the alignment to be created directly from the audio and text within the device.

6. REFERENCES

[1] Mark T.J. Carden, "E-books are not books," in *BooksOnline '08: Proceeding of the 2008 ACM workshop on Research advances in large digital book repositories*, New York, NY, USA, 2008, pp. 9–12, ACM.

[2] Diamantino Caseiro, Hugo Meinedo, António Serralheiro, Isabel Trancoso, and ao Neto, Jo "Spoken book alignment using wfsts," in *Proceedings of the second international conference on Human Language Technology Research*, San Francisco, CA, USA, 2002, pp. 194–196, Morgan Kaufmann Publishers Inc.

[3] Pedro J. Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *in Proc. ICSLP*, 1998.

[4] Anthony F. Martone, Cuneyt M. Taskiran, and Edward J. Delp, "Automated closed-captioning using text alignment," in *in Proc. SPIE*, 2004.

[5] Konstantin Biatov, "Large text and audio data alignment for multimedia applications," in *Text, Speech and Dialogue, Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2003.

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

[7] Nick Campbell, "Autolabelling japanese tobi," in *In Proc. of ICSLP*, 1996, pp. 2399–2402.

[8] Sérgio Paulo and Luís C. Oliveira, "Improving the accuracy of the speech synthesis based phonetic alignment using multiple acoustic features," in *Computational Processing of the Portuguese Language, in Lecture Notes in Computer Science*, 2003.

[9] F. Malfrére, O. Deroo, T. Dutoit, and C. Ris, "Phonetic alignment: speech synthesis-based vs. viterbi-based," *Speech Communication*, vol. 40, pp. 503–515, 2003.

[10] L. Carrico, N. Guimaraes, C. Duarte, T. Chambel, and H. Simoes, "Spoken books: Multimodal interaction and information repurposing," in *in Proc. 10th International Human-Computer Interaction*, Crete, Greece, 2003.

[11] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vreken, "The mbrola project: Towards a set of high-quality speech synthesizers free of use for non- commercial purposes," in *in Proc. ICSLP*, 1996, vol. 3, pp. 1393–1396.

[12] Robert A.J. Clark, Korin Richmond, and Simon King, "Multisyn: Open-domain unit selection for the festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317 – 330, 2007.

[13] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. in ASSP*, vol. 26, pp. 43–49, 1978.