Using Audio Books for Training a Text-to-Speech System

Aimilios Chalamandaris, Pirros Tsiakoulis, Sotiris Karabetsos, Spyros Raptis
[1] INNOETICS LTD, Athens, Greece
[2] Institute for Language and Speech Processing / Research Center "Athena", Athens, Greece
Artemidos 6, GR-15125, Marousi, Greece
E-mail: {aimilios,ptsiak,sotoskar,spy}@ilsp.gr

Abstract

Creating new voices for a TTS system often requires a costly procedure of designing and recording an audio corpus, a time consuming and effort intensive task. Using publicly available audiobooks as the raw material of a spoken corpus for such systems creates new perspectives regarding the possibility of creating new synthetic voices quickly and with limited effort. This paper addresses the issue of creating new synthetic voices based on audiobook data in an automated method. As an audiobook includes several types of speech, such as narration, character playing etc., special care is given in identifying the data subset that leads to a more neutral and general purpose synthetic voice. Part of the work described in this paper was carried out for the participation of our TTS system in the Blizzard Challenge 2013 where the developed TTS system was ranked top in overall impression in one of the two experimental hubs (Chalamandaris et al 2013). The main goal is to identify and address the effect the audiobook speech diversity on the resulting TTS system. Along with the methodology for coping with this diversity in the speech data, we also describe a set of experiments performed in order to investigate the efficiency of different approaches for automatic data pruning. Further plans for exploiting the diversity of the speech incorporated in an audiobook are also described in the final section and conclusions are drawn.

Keywords: speech synthesis, audio book, synthetic voice creation, TTS, unit-selection, training data pruning, expressive speech synthesis

1. Introduction

Typically, a state-of-the-art unit-selection text-to-speech (TTS) system relies on a database which usually derives from a large number of recordings by a professional voice talent, carefully designed, executed, and annotated at a phonetic level, either automatically or manually. The creation of a synthetic voice for a TTS system is an expensive and time consuming process, with three main aspects to address: a) the design of the textual corpus in order to cover most linguistic and acoustic properties of the target domain, b) the execution of the recordings and c) the annotation of the recordings, the accuracy of which is a major factor in the delivered synthetic speech, as far as intelligibility and naturalness are concerned.

Most of the times, the recordings are hyper-articulated with a neutral speaking-style in order to achieve highly intelligible and consistent synthetic speech. It has been demonstrated that larger databases generally lead to better quality synthetic voices, and by keeping this in mind, the aforementioned process is often an attempt to identify the best trade-off balance between quantity and cost (resources, effort, and money). Currently, state-of-the-art unit-selection TTS systems incorporate databases of several hours of running speech, something that makes the creation of a new voice an expensive project.

Audiobooks have been used as a raw material in speech synthesis research during the last few years (Bruer et al 2006; Zhao et al 2006; Szekely et al 2012; Eyben et al 2012; Braunschweiler et al 2011), as more audiobook resources become available, especially free of charge and with no rights reserved, and as they provide excellent training material for expressive speech synthesis research. There are already many resources of freely available audiobooks, many of which are produced from professional voice talents.

Although audiobooks have been employed as source material for emotional and expressive speech synthesis in several research attempts, one could utilize audiobooks for creating synthetic voices not only for producing new material, but for a general-purpose TTS system that could be used in other tasks, such as telecom applications or news-reading. The listening tests described in this paper include stimuli derived from novels and from newscasts in order to investigate the appropriateness of the TTS system in both domains – audiobooks and news reading.

The quality of the database, and more specifically the quality of the recordings and the accuracy of their annotation, is of crucial significance in the overall TTS quality. Therefore our work aims to address these two issues, i.e. the overall quality of the used recordings and the accuracy of their annotation. In this paper we describe our approach in crafting new synthetic voices for a unit-selection TTS system, based on the recordings of audiobooks. It is a fully automated process, which allows us to create new synthetic voices from freely available audiobooks with minimal effort and manual labor. Experiments have shown that the outcome of the approach provides high-quality synthetic speech, appropriate not only for narrative purposes, but for different tasks such as news-reading or simple conversational voice prompts.

This paper is organized as follows: Section 2 provides information about the audiobooks used in this study as they were provided in the Blizzard 2013 challenge. It also describes our methodology for the creation of a synthetic voice database for our system, while addressing issues that arise from the fact that the training material is not inherently appropriate for this task. Special focus is given on the pruning of the audio material in order to alleviate three different sources of errors in the TTS database. Section 3 describes the experiments and the listening tests

for assessing the different approaches for pruning the audio data. Section 4 provides a discussion of the results, a brief summary of our methodology and the future work to be carried out in the framework of this study.

2. The Speech Corpus and TTS Voice Construction

2.1. The audiobooks

For this study we used a set of 21 audiobooks provided for research purposes in the framework of the Blizzard Challenge 2013.

The total speech time was over 140 hours long and they were all recorded by a professional American English native female speaker. They cover a wide range of expressive styles as they are commercially available audiobooks.

The employed TTS system for this study is the one described in (Raptis et al, 2012) with a unit-selection approach and a diphone based concatenation technique. As such, the schema and the characteristics of the voice database crafted during this study were compatible with the characteristics of the aforementioned TTS system. As the total length of the available speech was over 140 hours, manual supervision was not possible. Hence, we used an automated method for filtering-out the audio segments from the recordings that would not "qualify" to be incorporated in the TTS database. The pruning method is based on three main criteria which are described in the following sections.

2.2. Alignment between text and audio

One of the main issues that need to be addressed for the creation of a database from an audiobook is the alignment of the audio recordings with the actual spoken text at the sentence level (Braunschweiler, et al 2010). The annotation of any necessary information for the TTS system heavily depends on it.

As we use an HMM alignment mechanism – it is described in (Raptis et al, 2012) – between the audio and the text parts of the audiobooks, we performed an iterative process of aligning them, using smaller and smaller parts of the recordings each time, in order to increase the accuracy of the alignment. Starting from large chunks of speech, ranging from 20 minutes to 1,5h long (most of the times chapters or bigger sections of an audiobook), we ran the audio alignment again on a second phase on a sentence level, after the initial alignment. This approach provides better and more accurate results (Braunschweiler, et al 2010) as in the second stage we were also able to build an acoustic model or the HMM aligner based on the actual recordings, while in the first stage, the acoustic model used for the HMM aligner was one of a different speaker of the same gender.

The alignment has been performed without any significant change or supervision of the input text, meaning that possible mispronunciations or disagreements between the text and the recorded speech could and did exist. Most of these misalignments would be excluded automatically at a later stage of the database crafting process, where the audio pruning is performed. In Figure 1 the segmentation stages are illustrated.

2.3. Audio Segments Pruning

The pruning of the audio segments after the automatic alignment of the recordings with the respective text aims at alleviating three main factors that introduce errors and inconsistencies in the TTS database: a) misalignments between audio and text, mainly because the narrator has not read the exact same text, b) phrases of prominent reading where the narrator makes imitations, character's voices, etc., and c) different environmental settings during the recordings which often lead to spectral mismatches during synthesis. In the following sections we describe the method for coping with each problem source independently.

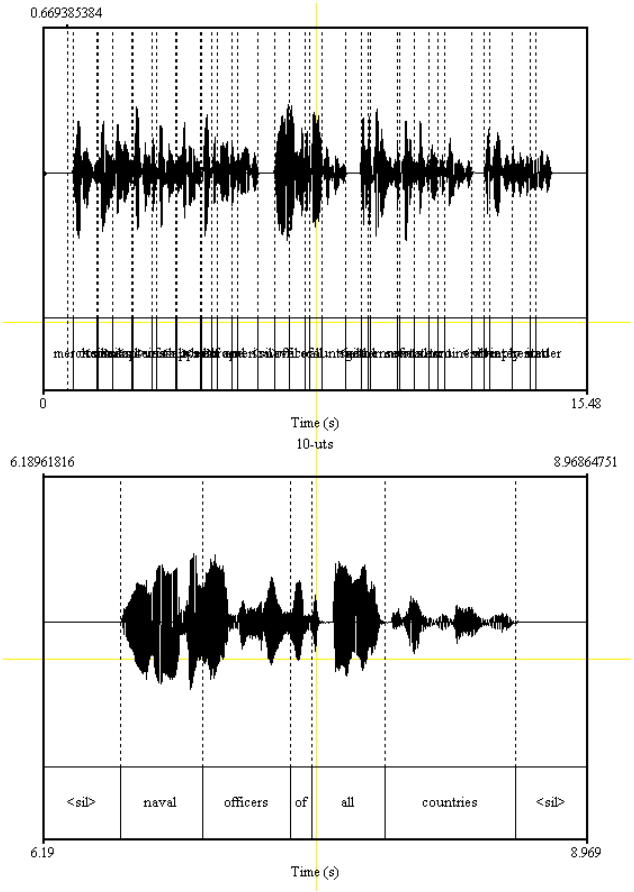

Figure 1: The segmentation and alignment of the audio data is performed in an iterative method, starting from long audio chunks (e.g. chapters or sections) and concluding to a phrase level (identified as the chunk between two consecutive recognized speech pauses). Phrases are then aligned with the text and annotated at the phonetic level.

2.3.1 Audible differences between recordings

As stated earlier, an issue that needs to be resolved during this process is the fact that different audiobooks are often recorded in different environmental settings, such as studio, recording devices etc. Sometimes, one can identify such differences even within the same audiobook, from one section to another. For alleviating this issue we performed a clustering of the audio files, based on the SFFT spectral content of the recordings and a K-means

clustering method and kept the sentences that were members of the same or of similar clusters.

	EH1	*Length*	*Further*	*DB Length*
1	black	4:42	35%	3:10
2	mansfield	13:08	32%	8:51
3	awakening	4:47	33%	3:15
4	black	4:36	32%	3:10
5	chatterley	10:18	31%	7:12
6	cityoz	0:42	30%	0:29
7	daisy	1:01	29%	0:40
8	dalloway	4:22	33%	2:58
9	emerald	0:42	32%	0:29
10	emma	14:05	30%	9:26
11	frankenstein	6:32	33%	4:22
12	jane	16:13	33%	11:40
13	leagues	8:24	28%	5:57
14	madding	12:32	29%	8:46
15	magi	0:12	30%	0:08
16	patchwork	2:09	31%	1:27
17	pride	11:08	32%	7:27
18	roomview	6:50	33%	4:38
19	scarlet	6:18	32%	4:17
20	treasure	6:13	32%	4:28
21	washington	6:00	28%	4:07
	Total	140:54:00	31%	97:07:16

Table 1. The duration for each audio book used and the resulted database after the pruning stage.

2.3.2. Prominent speech pruning

Audiobooks are exciting material for text-to-speech voice crafting as they include various linguistic and acoustic phenomena, enabling great potential for emotional and expressive text-to-speech systems if modeled and reproduced appropriately. However, since our aim is to build a general-purpose TTS voice, less emotional and more neutral, we attempted to identify and remove most prominent speech parts of the recordings from the database.

The issue of identifying different speech styles within audio recordings has been addressed previously (Wang et al, 2006; Székely et al, 2012; Cheng et al 2012; Charfuelan et al 2012) with different results depending on the set of acoustic and linguistic features extracted from the audio and text respectively (Raptis, 2013; Norrenbrock et al 2012).

Our methodology employed two acoustic features only, the mean and the variance of the F0 variable, on each phrase. By phrase we mean the audio part between two consecutive recognized silences or pauses by the voice alignment engine, and not a textual sentence. This is a reasonable choice, as the narrator tends to keep the same speaking style within a phrase and make a pause before changing speaking style. From the entire population of the recognized phrases, we keep only a portion of the phrases which is located closest to the centroid of the distribution based on a specific distance. In this study we discarded 10% of the audio data located in distance of the population's centroid.

2.3.1. Segmental errors pruning

During the automatic alignment process between the audio recordings and the actual text of the audiobooks there are segmental errors which are caused either by possible disagreements between the text and the spoken version of it or by possible fails of the actual aligning mechanism. In order to address this issue, a simple but efficient stage of pruning was introduced, which depends on the local HMM score of every phoneme and the overall score of every sentence. This criterion is used to decide whether an aligned sentence would have probably been appropriately segmented and annotated, and hence a good addition to the TTS database. By doing so, sentences or words that had received a low score by the HMM algorithm during alignment were removed from the recordings pool and the voice database crafting process.

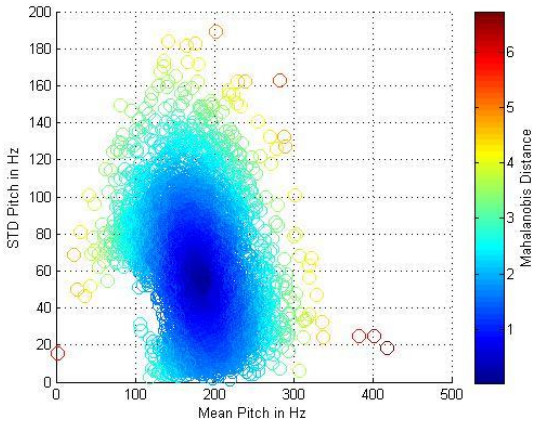

Figure 2: Distribution of the acoustic phrases based on their Mean Pitch value versus the Standard Deviation Pitch value. The Mahalanobis distance of each instance from the population's centroid is annotated via a color map.

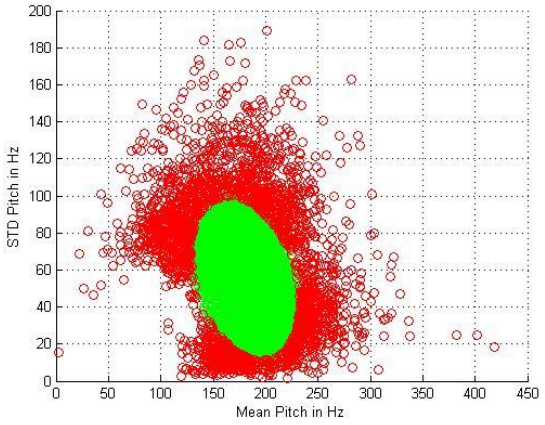

Figure 3: The phrases that were selected to be included in the database are denoted with a green circle. The selection criterion is based on a manually estimated threshold of the calculated distance.

3. Experiments

In order to investigate the validity of our background theory in the prominent speech data pruning process, a

pilot test was carried out and the listeners were asked to identify their preference among TTS systems that were based on the full dataset and on the pruned one, with a different level of pruning each time.
More specifically, we developed three different TTS systems, one based on the entire audio dataset without any pruning, a second one with pruning based on prosodic features on a phrase level as mentioned in the previous section (namely mean and standard deviation of the pitch value), and a third one from a dataset pruned on prosodic features (as in system B) and segmental criteria, based on the HMM score of every phoneme after the forced alignment process between the audio and the text. The threshold was set manually and 10% of the population of every phoneme was discarded, keeping in mind however that the coverage of all possible diphones should be kept intact.
The aforementioned systems are named Systems A, B and C respectively in the experiment.
The listening tests have been carried out with the Amazon Mechanical Turk web service (Amazon, 2012).

3.1 Experiment setup

To obtain objective preference measures, a set of 300 synthetic stimuli was produced, 100 stimuli from each TTS system, as described above. The stimuli were selected automatically from a larger set of synthetic stimuli, as defined in the Blizzard Challenge 2013, based on two criteria: a) to create comparison pairs with significant auditory difference (based on spectral distances between stimuli) and b) to include different types of domains (i.e. novels, news, book sentences and book paragraphs). The final audio material for the experiments consisted of 100 synthetic stimuli for each TTS system, containing sentences from news (52%), sentences from novels and books (34%) and entire paragraphs from literature books (14%) and which created 300 different preference test questions in total, i.e. comparison between A and B, A and C, and B and C.
The stimuli were shuffled automatically and 49 different listeners were asked to identify which of the two stimuli sounded better to them, having at the same time the alternative option to denote that none of them sounded better from the other. Every pair of stimuli was tested 10 times from 10 different listeners.

3.2 Results

An overview of the results is depicted in Table 2. Together with the overall results, the Table depicts the preference test results for every stimuli category, depending on the text domain they belong to, namely news sentences, book/novel sentences and book paragraphs.
It can be seen from the table that TTS system C, with both prosodic and segmental pruning, is significantly preferable to both A and B systems. The above observation is valid for every type of stimuli. TTS system B is preferable to system A only on news texts and inferior in literature texts, without however having obtained a statistical significant result in both cases.

3.3 Discussion

The most prominent conclusion of the experiment is that the third TTS system (TTS C) which integrated the pruned database with both prosodic and segmental criteria, is preferable in every type of text and this is attributed to the fact that system C did not suffer from segmental errors like systems A and B did.
At the same time, any comparison between TTS systems A and B seems to be inconclusive from this experiment as the two systems did not perform significantly different one from each other, although we expected system B to outperform system A. The most obvious explanation for this result is that although TTS system B did not suffer from prominent speech parts which would produce prosodic discontinuities in the synthetic stimuli, it did suffer from segmental errors like TTS system A. As segmental errors in synthetic stimuli are dominant in people's overall impression, TTS system B was not considered significantly different from TTS system A.

Stimulus Type	System A	System B	System C	N/P	p value (sign-test)
News	41,5%	48,3%	-	10,2%	0,204
	-	37,6%	**51,6%**	10,8%	0,033
	36,0%	-	**56,2%**	7,9%	0,004
Book Sentences	50,0%	37,7%	-	12,3%	0,112
	-	40,2%	47,9%	12,0%	0,230
	36,8%	-	**55,6%**	7,7%	0,021
Book Paragraphs	46,0%	44,0%	-	10,0%	0,444
	-	43,8%	54,2%	2,1%	0,235
	18,8%	-	**72,9%**	8,3%	<0,001
Total	45,0%	44,1%	_	10,9%	0,435
	-	39,3%	**50,7%**	10,0%	0,016
	33,8%	-	**58,3%**	7,9%	<0,001

Table 2. Preference scores (%) among speech samples synthesized from TTS systems A, B and C. N/P denotes "no preference". The systems which achieved significantly better preference at $p < 0.05$ level are in the bold font.

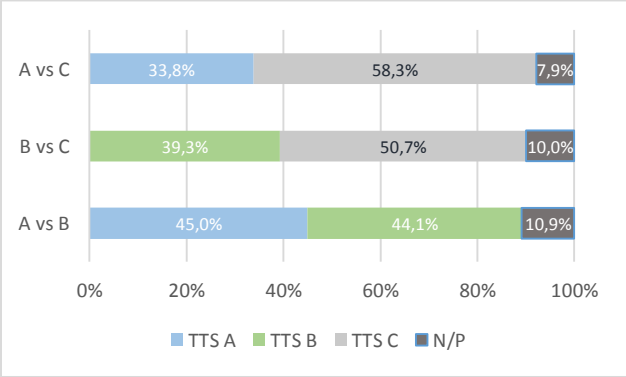

Figure 4. Preference scores (%) among speech samples synthesized from TTS systems A, B and C. N/P denotes "no preference". The preference listening tests were performed in pairs comparing each time two systems.

4. Conclusion

In this paper we described a method for developing a TTS synthetic voice from an audiobook. As an audiobook includes several types of speech like narration, character playing, etc., our main concern was to identify the data subset that generate better synthetic speech compared to the synthetic speech that would derive from the entire audiobook dataset. In this paper we investigated two methods for pruning this audio data and validated their performance through a listening test. The results have shown that although simple prosodic pruning does not result in a significantly better TTS voice, pruning based on both prosodic and segmental features lead to better synthetic speech, not only for texts of similar domain (i.e. books) but also in other domains, such as news sentences. As the results of the experiment we carried out were only partially statistically significant, we plan to investigate further methods for appropriately pruning audiobook data for creating TTS synthetic voices.

5. Acknowledgements

The research leading to these results has received funding from the POLYTROPON project (KRIPIS-GSRT, MIS: 448306).

6. References

Braunschweiler, Norbert, Mark JF Gales, and Sabine Buchholz, (2010). "Lightly supervised recognition for automatic alignment of large coherent speech recordings." Proceedings of the 11th Annual Conference of the International Speech Communication Association. Curran Associates, Inc.

Braunschweiler, Norbert, and Sabine Buchholz, (2011). "Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality." Proc. Interspeech.

Breuer, Stefan, et al., (2006) "Set-up of a Unit-Selection Synthesis with a Prominent Voice." Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), Genova..

Chalamandaris A., Tsiakoulis P., Karabetsos S., Raptis S., (2013) "The ILSP/INNOETICS Text-to-Speech System for the Blizzard Challenge 2013", In Proc. Blizzard Challenge 2013 Workshop (satellite of Interspeech).

Charfuelan, Marcela, and Marc Schroder, (2012). "Correlation analysis of sentiment analysis scores and acoustic features in audiobook narratives." ES3-LREC Workshop.

Chen, Langzhou, et al., (2012) "Exploring Rich Expressive Information from Audiobook Data Using Cluster Adaptive Training." INTERSPEECH.

Eyben, Florian, Sabine Buchholz, and Norbert Braunschweiler, (2012) "Unsupervised clustering of emotion and voice styles for expressive TTS." Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on IEEE.

Norrenbrock, Christoph R., et al. (2012) "Towards Perceptual Quality Modeling of Synthesized Audiobooks–Blizzard Challenge.

Raptis S., (2013) "Exploring latent structure in expressive speech", In Proc. 4th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2013), Budapest, Hungary, Dec. 2-5,

Szekely, E., et al. (2012) "Evaluating expressive speech synthesis from audiobook corpora for conversational phrases." Proc. of LREC, Istanbul.

Szekely, Eva, et al., (2011) "Clustering expressive speech styles in audiobooks using glottal source parameters." Proc. of Interspeech, Florence (2011): 2409-2412.

Szekely, Eva, et al. (2012) "Detecting a targeted voice style in an audiobook using voice quality features." Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012.

Szekely, Eva, et al., (2012) "Synthesizing expressive speech from amateur audiobook recordings." Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE, 2012.

Amazon Mechanical Turk. (2012): "Amazon Mechanical Turk." Retrieved August 17 2012.

Wang, Lijuan, et al. (2006) "Exploring expressive speech space in an audio-book." Proc. of Speech Prosody 2006.

Zhao, Yong, et al. (2006) "Constructing stylistic synthesis databases from audio books." Proceedings of Interspeech. 2006.