Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality

Norbert Braunschweiler, Sabine Buchholz

Toshiba Research Europe Ltd., Cambridge Research Laboratory, Cambridge, United Kingdom

{norbert.braunschweiler,sabine.buchholz}@crl.toshiba.co.uk

Abstract

Using publicly available audiobooks for HMM-TTS poses new challenges. This paper addresses the issue of diverse speech in audiobooks. The aim is to identify diverse speech likely to have a negative effect on HMM-TTS quality. Manual removal of diverse speech was found to yield better synthesis quality despite halving the training corpus. To handle large amounts of data an automatic approach is proposed. The approach uses a small set of acoustic and text based features. A series of listening tests showed that the manual selection is most preferred, while the automatic selection showed significant preference over the full training set.

Index Terms: speech synthesis, HMM-TTS, corpus creation, diverse speech, speaking styles, audiobooks

1. Introduction

Traditionally synthetic voices have been built on specifically designed corpora carefully recorded by a professional voice talent in a single speaking style, often a neutral style with moderate voice modulation. This ensured the highest synthesis quality and a homogenous speaking style at synthesis time, albeit often too uniform for longer texts. In an effort to realize more expressive synthetic speech, especially for long coherent texts, the use of other speech corpora like audiobooks has become the focus of research for synthetic voice building [1]. However, publicly available audiobooks are often read by non-professional readers under non-optimal recording conditions. Many of these audiobooks include a wide variety of speaking styles with one or more speakers.

Work on using publicly available audiobooks for research and for building synthetic voices has so far been focused on preparing these large corpora to make them accessible for voice building [1] [2] [3]. In [3] an approach has been developed to automatically segment and select sentences from audiobooks based on their correspondence between book text and speech recognition result. However, the presence of diverse speech as a challenge for synthetic voice building has not been widely addressed.

Diverse speech in publicly available audiobooks can include character voices, highly emotional speech, and speech with audibly different loudness, rate or quality. Here the term "audibly different" refers to some baseline considered to be 'neutral' speech with homogenous characteristics in terms of style, loudness, rate and quality. Although it has been mentioned in the literature that it is not possible to speak without expression [4] the concept of 'neutral' speech is still helpful as a baseline. Typically 'neutral' speech can be associated with traditional TTS corpora, while many audiobooks include a wide range of 'non-neutral' speech.

While using publicly available audiobooks for HMM-TTS voice building it was noticed that synthesis quality was degraded compared to standard TTS corpora. To test whether the presence of diverse speech affected this result it was decided to run a pilot experiment (described in detail in the next section) which separated 'neutral' speech from 'non-neutral' or potentially detrimental (for TTS voice building) speech. It was found that the removal of 'non-neutral' speech improved synthesis quality despite halving the training corpus. In order to handle large amounts of audiobooks an automatic approach was created to detect diverse speech likely to have a negative effect on synthesis quality. This approach is introduced in section 3. The approach is then evaluated in a series of listening tests in section 4 showing its performance; followed by the conclusion sketching some further directions.

2. Manual selection of 'neutral' speech

A pilot experiment was conducted to test whether an HMM-TTS voice based on a manually selected subset of sentences considered to be suitable for synthetic voice building is preferred to a voice based on all sentences extracted from an audiobook.

2.1. Audiobook corpus

The audiobook used here is based on the American English book "A Tramp Abroad" written by Mark Twain and publicly available from *librivox.org* [5]. The recording condition is acceptable, albeit there are some differences across recordings. The book describes the experiences of an American tourist while travelling through Europe around 1880. Its total running time is 15 hours and 46 minutes. The narrator (John Greenman) uses a wide range of speaking styles to express funny, humorous, bizarre, etc. passages.

The audiobook was split into sentences by the lightly supervised sentence segmentation and selection approach [3]. This approach compares the recognition result of a sentence with its original orthographic version in the book. When the comparison results in a word accuracy lower than the specified threshold the sentence will be discarded. A 75% confidence threshold was used which resulted in 6949 or 87.5% of sentences selected (15 hours, 891 minutes).

2.2. Manual selection

To select neutral sentences from the audiobook a human labeler was instructed to listen to each sentence and judge whether it sounds neutral or whether it should be discarded. The labeler used the following guidelines: select sentences spoken in the speaker's neutral style which is not strictly defined as the absence of all expression, but may include a neutral-narrative style, but not audibly expressive styles like shouting, whispering and speech sounding angry, happy, sad, etc. Also discard sentences potentially harmful for building a TTS system. In doubt, imagine a traditional unit selection

TTS corpus and ask yourself whether the sentence would have a negative effect on such a corpus, i.e. sentences including obvious problems such as differences between text and audio, audio files with parts cut off, no leading and/or trailing silences, and files including distortions.

This manual selection resulted in a subset of 3354 sentences representing 48.3% of the original corpus (6.4 hours, 387 minutes). This subset was used to build an HMM-TTS voice henceforth called NEUTRAL_hand.

2.3. Experimental results

Two HMM-TTS voices were built, one based on the full set of 6949 sentences (henceforth called voice FULL) and the other based on the manually selected neutral subset of 3354 sentences (NEUTRAL_hand). Both voices were built fully automatically and separately, e.g. automatic phone alignments were created for each corpus individually by using HMM models built from a flat-start [6].

70 sentences were synthesized using each voice. Sentences were selected from various domains, e.g. news, email, navigation and included several sentence types like yes/no-questions, wh-question, imperatives and statements.

A preference test was conducted via crowdsourcing including 24 subjects. Each subject listened to a randomly selected subset of 30 sentences and had to listen to the same sentence in two synthesized versions: one from voice NEUTRAL_hand the other from voice FULL. Listeners had the option to choose "no preference". Sample order as well as the order of sentences were both randomized for subjects.

Table 2.1 shows the mean preference scores. Voice NEUTRAL_hand was preferred to voice FULL (53.9% vs. 32.5%). This difference is statistically significant ($p<0.01$).

FULL	NEUTRAL_hand	Neither
32.5%	53.9%	13.6%

Table 2.1. *Mean preference for the comparison of HMM-TTS voices FULL vs. NEUTRAL_hand.*

The result showed that listeners preferred the voice based on the manually selected subset over the full training set voice, despite the latter having more than twice the amount of training data. Since the manual selection was very time consuming and is impractical for large amounts of data, an automatic method was developed. The next section introduces the automatic approach taken.

3. Automatic 'neutral' sentence selection

To automatically select a subset of neutral sentences and discard potentially harmful ones a combination of acoustic and text based features was used as both are available from the preparation step. Because of restrictions in time the most straightforward approach was chosen, which was a rule based approach utilizing a relatively simple set of features. Additional features based on e.g. phone alignments were not considered because the method was intended to be applied before any phone alignments become available.

Limited time was spent to fine tune threshold values as the focus was set on the removal of the more extreme ends of the diverse speech spectrum. The aim was to end up with a relatively 'clean' set of speech even at the expense of discarding slightly too many sentences. The latter is not considered to be problematic because the initial amount of data is typically large.

3.1. Acoustic features

Table 3.1 shows the acoustic features used in the automatic approach. Features are designed to discard sentences including audibly extreme f0 patterns and sentences that sound too loud or are barely audible. Four features check the f0 range, another four features check the range of RMS-amplitude. One feature calculates the percentage of voiced frames, a feature intended to detect, e.g. whispered speech.

The f0 features check whether there are any extreme outliers by comparing upper and lower extremes from individual sentences against f0 mean values calculated across the whole corpus. $f0_{mean\_max}$ is the average of all $f0_{max}$ values calculated for each sentence in the corpus. $f0_{mean\_mean}$ is the average of all non-zero f0 values calculated across the whole corpus. A similar strategy is chosen for the RMS features.

The scaling of the threshold values was done by listening to sentences for which the individual features had been extracted and then deciding whether they should be raised or lowered based on the perceptual impression of the sentence sounding neutral or not. This was conducted on a separate training set (all audiobooks) and included 5 male and 5 female speakers.

Feature name	Definition
f0 max too high	$f0_{max} > f0_{mean\_max} * 1.40$
f0 max too low	$f0_{max} < f0_{mean\_mean} * 1.35$
f0 mean too high	$f0_{mean} > f0_{mean\_mean} * 1.50$
f0 mean too low	$f0_{mean} < f0_{mean\_mean} / 1.38$
Min % voiced	ratio frames voiced/unvoiced
RMS max too high	$RMS_{max} > RMS_{mean\_max} * 2.0$
RMS max too low	$RMS_{max} < RMS_{mean\_max} * 1.1$
RMS mean too high	$RMS_{mean} > RMS_{mean\_mean} * 1.9$
RMS mean too low	$RMS_{mean} < RMS_{mean\_mean} / 2.8$

Table 3.1. *Acoustic features used in the automatic neutral sentence selection approach.*

3.2. Text based features

The following text based features are used in addition to acoustic features. The features are split into (a) features identifying non-neutral sentence types, i.e. sentences likely to be spoken in a non-neutral style, such as the following

- Quotation marks → likely to be direct speech
- Interjections, e.g. "Oh, Ah, Hm" → expressive speech
- Starting with lowercase → sentence fragment
- Sequence of 3 full stops: "…" → fragment
- Ending in comma, colon or semi-colon → fragment

and (b) features identifying sentences likely to include text normalization (TN) errors:

- Ampersand "&" → likely to cause problems for TN
- Digit in square brackets, e.g. "[1]" → TN problems
- Year digits, e.g. "1880" → TN problems

The presence of quotation marks is likely to indicate direct speech - often realized in a non-neutral style. Therefore all sentences including quotation marks are discarded. However, since the current approach evaluates sentences individually and does not detect opening and/or closing

quotation marks across single sentences, some direct speech sentences will not be detected.

TN errors could result in incorrect phone alignments having a negative influence on synthesis quality.

The percentage of sentences containing these kinds of text features can vary considerably per book and depends on the text genre, writing style of the author, formatting, etc.

3.3. Additional features to exclude problematic files

Additional features were added as follows because of known problems in HMM-TTS with certain training material:

- <25 ms leading and/or trailing silence
- Duration of audio file >15 sec or <0.8
- Duration of audio file >Dur_{mean} * 5 or <Dur_{mean} / 6

When mixing training data with and without leading/trailing silences the inconsistency of the training data increases and subsequently the variance of the model. This in turn results in over-smoothing in the synthesized speech.

Audio files >15 seconds could include multiple sentences and might be a result of mislabeling sentence boundaries. Such long audio files have a higher chance of having incorrect phone alignments. The limit of 15 seconds was determined by checking the longest sentences in typical TTS corpora. Very short audio files (<800 ms) can include word fragments caused by misalignments or interjections.

Finally two features compare the duration of each audio file with the average audio file duration across the corpus and check whether a file is a very long outlier (duration > Dur_{mean} * 5) or a very short outlier (duration < Dur_{mean} / 6). The next section presents the percentage of sentences detected by each of the features when applied to an audiobook.

3.4. Effect of features

Table 3.2 shows an overview of the acoustic features and the percentage of sentences discarded for corpus "A Tramp Abroad" (cf. section 2.1). Some features result in a relatively large number of sentences being discarded, e.g. *total duration too long*, *f0 maximum too low/high*, whereas others are just discarding a small fraction of sentences, e.g. *RMS mean too high*, *RMS max too low*, *percentage of voiced too low*. Duration based features and features checking the f0 are discarding more sentences than the RMS based features.

Table 3.3 shows the effect of text based features. Features with the highest number of detected sentences are *double quotes*, *ends in comma or (semi-)colon*, and *lowercase sentence start*. The other features contribute relatively little to the overall detection rate of the text based features.

Combining audio and text features and removing duplicates results in 42.1% of all sentences being discarded. Acoustic features used to identify non-neutral speaking styles (f0, RMS, voicing) categorize 15.7% of sentences while the others (duration, leading/trailing silences) discard 11.4%. The latter is mainly a result of discarding sentences >15 sec. In the acoustic features designed to identify non-neutral speaking styles the f0 features account for the majority of hits (>99%).

Text based features designed to identify non-neutral sentences are responsible for the majority of sentences detected by text based features (23.4%) while the remaining three features (ampersand, digit in bracket, year digits) contribute very little (<0.4%). The largest contribution (about two-third) stems from feature "double quotes".

Class	Feature	Detected
Silence	No leading/trailing silence	2.2%
Duration	total duration too long	10.7%
	total duration too short	0.3%
	relative duration too long	-
	relative duration too short	3.2%
F0	f0 max too high	6.4%
	f0 max too low	8.3%
	f0 mean too high	2.0%
	f0 mean too low	0.6%
RMS	RMS max too high	0.2%
	RMS max too low	0.1%
	RMS mean too high	<0.1%
	RMS mean too low	0.5%
Voicing	% of voiced too low	0.1%
All acoustic features (no duplicates)		27.2%

Table 3.2. *% of sentences discarded per acoustic feature. Field "All acoustic features" shows percentage after removing duplicate hits.*

Feature	Detected
Double quotes (")	15.8%
Interjections ("Ah, Oh")	0.4%
Lowercase start ("a man …")	3.8%
3 full stops ("…")	0.2%
Ends in comma, (semi-)colon	4.5%
Ampersand ("&")	<0.1%
Digit in bracket ("[1]")	<0.1%
Year digits ("1880")	0.3%
All text features (no duplicates)	23.8%

Table 3.3. *% of sentences discarded per text feature. Field "All text features" shows percentage after removing duplicate hits.*

4. Evaluation

To evaluate the automatic selection approach the following evaluations were conducted all of them using audiobook "A Tramp Abroad" as described in section 2.1.

4.1. Manual vs automatic selection of 'neutral' data

The automatic selection was first evaluated against the manual selection. Both objective and subjective comparisons were conducted.

Corpus NEUTRAL_hand was as described in section 2.2. Corpus NEUTRAL_auto including the automatically selected neutral subset comprised 4026 sentences, representing 57.9% of the original corpus (7.8 hours).

The automatic selection was compared to the manual one by treating the NEUTRAL_hand corpus as gold standard and calculating precision and recall. Precision was 58.9% indicating quite some discrepancy between human and automatic selection and showing the large number of

sentences incorrectly classified as 'neutral'. Recall was 70.6% showing that the number of false negatives was not extremely high. The automatic approach discarded about 10% fewer files and discarded almost 30% of NEUTRAL_hand sentences. More than 38% additional sentences are selected by the automatic approach. Since the human labeler used additional criteria like voice quality, speaking rate, distortions, etc., some of the discrepancy may be due to this.

A listening test was conducted comparing voices NEUTRAL_hand vs. NEUTRAL_auto. The preference test was done by 20 subjects via crowdsourcing. The result shows a significant preference ($p<0.05$) for voice NEUTRAL_hand (see Table 4.1).

NEUTRAL_hand	NEUTRAL_auto	Neither
47.7%	36.7%	15.6%

Table 4.1. *Mean preference for the comparison of voices NEUTRAL_hand vs. NEUTRAL_auto.*

While the automatic selection method does not achieve the performance of a human in this task, this is not very surprising given the relatively simple criteria used. Further refinements are possible to improve the automatic selection.

4.2. Automatic 'neutral' data vs full training

To test how the automatic selection produced voice compares to the full voice another listening test was conducted comparing voice FULL (cf. section 2.3) with voice NEUTRAL_auto.

The result of the preference test is shown in Table 4.2. It was run via crowdsourcing and included 24 subjects. The test showed a significant preference ($p<0.01$) for voice NEUTRAL_auto. This shows that the automatically selected sentences also resulted in improved HMM-TTS quality. However, the difference in average preference between FULL and NEUTRAL_auto is lower than the difference obtained in test FULL vs NEUTRAL_hand (cf. section 2.3), i.e. 16.4% vs. 21.4% respectively also confirming the results of section 4.1.

FULL	NEUTRAL_auto	Neither
32.9%	49.3%	17.7%

Table 4.2. *Mean preference for the comparison of HMM-TTS voices FULL vs. NEUTRAL_auto.*

4.3. Effect of confidence level

One might suspect that the confidence level of 75% in the lightly supervised approach affected the outcome of the listening test. To check this, a test was conducted using a 100% confidence threshold which provides the highest confidence that text and audio corresponds to each other.

FULL_100	NEUTRAL_auto100	Neither
34.7%	47.2%	18.1%

Table 4.3. *Mean preference for the comparison of HMM-TTS voices FULL_100 vs. NEUTRAL_auto100.*

At the 100% threshold 5169 sentences (69.0%) were extracted (943 minutes). Applying the automatic neutral sentence selection script to this subset resulted in 3200 (42.7%) sentences (354 minutes). Two voices were compared: voice FULL_100 (based on all 5169 sentences) vs. voice NEUTRAL_auto100 (based on the automatically generated subset of 3200 sentences).

The result of the test is shown in Table 4.3. It was run via crowdsourcing and included 32 subjects. It shows a significant preference ($p<0.01$) for voice NEUTRAL_auto100 showing that the effect of the initial choice of a 75% confidence threshold did not significantly affect the results.

5. Conclusions

If not handled, diverse speech in a speech corpus degrades TTS quality as shown in a pilot experiment. To improve synthesis quality of voices built on speech corpora including diverse speech an automatic sentence selection approach was presented. The new approach identifies (a) sentences spoken in a non-neutral style, and (b) sentences which are potentially detrimental for synthesis quality by using a small set of acoustic and text based features. The speech corpus used in this study was a publicly available audiobook including a wide range of speaking styles from a single speaker.

The accuracy of the automatic approach was evaluated by comparing it to a manually performed selection. The results showed that the synthesis voice based on the manual selection was preferred, although by a small margin. However, the automatic selection was still significantly preferred over training on the full set of sentences, showing the positive effect of the automatic method.

More refined features in the automatic approach could certainly boost performance. Candidates for this refinement are features such as speech rate, phone duration and spectral features related to voice quality.

6. Acknowledgments

The authors would like to thank the teams behind Librivox.org and Gutenberg.org for hosting voluntarily produced audiobooks and out of copyright books; John Greenman, who made his narration of "A Tramp Abroad" publicly accessible and Anette Laver for listening to "A Tramp Abroad" and identifying neutral sentences.

7. References

[1] Prahallad, K., Toth, A.R., and Black, A.W., "Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases", in Proc. of Interspeech, Antwerp, Belgium, 2007.

[2] Prahallad, K. and Black, A.W., "Handling Large Audio Files in Audio Books for Building Synthetic Voices", in Proc. of the 7th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW7), Kyoto, Japan, p.148-153, 2010.

[3] Braunschweiler, N., Gales, M.J.F., Buchholz, S. "Lightly supervised recognition for automatic alignment of large coherent speech recordings", in Proc. of Interspeech, Makuhari, Chiba, Japan, p.2222-2225, 2010.

[4] Tatham, M. and Morton, K. "Expression in Speech", Oxford University Press, 2004.

[5] Librivox audiobook "A Tramp Abroad" by Mark Twain read by John Greenman: http://librivox.org/a-tramp-abroad-by-mark-twain, cataloged on June 04, 2009.

[6] Buchholz, S., Braunschweiler, N., Morita, M., Webster, G., "The Toshiba entry for the 2007 Blizzard Challenge". In Proceedings of The Blizzard Challenge 2007: Workshop, Bonn, Germany, 2007.