DATA SELECTION FOR IMPROVING NATURALNESS OF TTS VOICES TRAINED ON SMALL FOUND CORPUSES

F.-Y. Kuo, S. Aryal, G. Degottex, S. Kang, P. Lanchantin, I. Ouyang

ObEN, Inc.

{fang-yu, sandesh, gilles, sam, pierre, iris}@oben.com

ABSTRACT

This work investigates techniques that select training data from small, found corpuses in order to improve the naturalness of synthesized text-to-speech voices. The approach outlined in this paper examines different metrics to detect and reject segments of training data that can degrade the performance of the system. We conducted experiments on two small datasets extracted from Mandarin Chinese audiobooks that have different characteristics in terms of recording conditions, narrator, and transcriptions. We show that using a even smaller, yet carefully selected, set of data can lead to a text-to-speech system able to generate more natural speech than a system trained on the complete dataset. Three metrics related to the narrator's articulation proposed in the paper give significant improvements in naturalness.

Index Terms— TTS, data selection, found data, audiobooks.

1. INTRODUCTION

Conventional text-to-speech systems are trained on dedicated datasets including speech recordings with their corresponding transcriptions. The acoustic environment conditions, the pronunciation and speaking style of the phonetically balanced set of recordings are carefully controlled in order to optimize the quality and the naturalness of the generated speech. The recordings are carefully annotated. In some use cases it is not possible to build such a controlled dataset, due to the unavailability of the target speaker. For instance, in the case of low resources language it might be necessary to use rare existing recordings of potentially extinct languages because we don't have access to a native speaker. Also, we might want to build a personalized TTS system with the voice of a celebrity who is not available because he or she is too busy or has passed away. Finally, we might want to personalize a voice output communication aid (VOCA [1]) using old recordings of people now afflicted with Multiple Sclerosis, Parkinsons and Motor Neuron Disease (MND). A large quantity of *found data* is available, such as data dedicated to the training of automatic speech recognition (ASR) systems, web data, public speeches, news and radio broadcasts, Youtube videos, phone conversations, or audiobooks. Those data can be of interest because a large quantity can be available for free and the naturalness of the speech can sometimes be higher than dataset dedicated to TTS (e.g. spontaneous speech). However, training TTS from found data is not straightforward because this kind of data can be really diverse in term of audio characteristics and available transcription. Being not controlled, the speech recordings can have a lot of variations, be inconsistent, be highly expressive, be multi-speaker, and have background noise. Also, text transcriptions can be inaccurate, incomplete or unavailable. Audiobooks are a popular source of found data useful for the building of TTS systems [2–6]. Usually mono-speaker and read by a professional narrator, they provide a large quantity of speech from the same speaker. Transcription is provided by the book and is usually accurate. They can be highly expressive, having large variations in prosody, intonation as well as speech rate. Speaking style may also vary when a single narrator tries to mimic different characters. All those variations tend to be difficult to model in current speech synthesis systems. Previous methods have looked at building speech synthesis systems from audiobooks [3, 5] and ASR corpora [7] while concentrating on techniques for building average voices using a variety of speaker adaptation techniques.

It is then of interest to develop data selection methods for producing high-quality voices from heterogeneous data sources. Different selection approaches were proposed for the data selection in audiobooks in previous work. In [6], authors controlled the recording conditions by producing a recording-condition-based clustering and only using utterances from one cluster. Variance in speaking style was also controlled by removing outliers of mean and standard deviation of pitch. They also discarded sentences with a low alignment score in order to remove both poorly-aligned. Combination of these approaches for a unit-selection system resulted in a better voice. Similarly, [4] controlled misalignment by selecting only utterances with high automatic alignment confidence score and created a module for selecting utterances with uniform speaking style in order to build a corpus of 60 hours of speech from audiobooks in 14 different languages for the purpose of building HMM-based voices for these languages. In [2], low-confidence utterances were discarded based on ASR confidence rather than alignment. They developed an automatic method for determining utterance naturalness as well. Despite discarding nearly half the original data, they found that the HMM voices they trained using both of these methods were judged to be significantly better than using all of the data in a preference test. In [8], the authors showed that selecting a smaller, cleaner subset for voice building give better results and is less time-consuming than building from a full noisy dataset. They did experiments on three datasets, an artificially degraded set of clean speech, a single speaker database of found speech, and a multi-speaker database of found speech. They proposed various utterance level metrics which would be indicators of the measure of goodness of an utterance. In [9], the authors produced subsets of utterances based on different metrics as well and showed that removing hyper-articulated outlier utterances and combining hypo-articulation with low mean F0 could lead to significantly more natural voice than the baseline. In [10], they found that selecting from utterances based on metrics such as standard deviation of F0, fast speaking rate and hypo-articulation produces the most intelligible voices.

In this work, we explore the use of different selection metrics on two small corpuses of 1.5h extracted from two Mandarin Chinese audiobooks. Our goal here is to propose and study different selection metrics and to show that some of them are efficient to improve the naturalness of the synthesized speech, even with smaller datasets. This paper is organized as follows. In Section 2, we introduce the approach used for the forced-alignment of complete audiobooks and present the different metrics which are used for selection, in Section 3, we present the experiments on two Mandarin Chinese audiobooks. Finally, Section 4 concludes our findings and discusses our future work in this direction.

2. PROPOSED APPROACH

Found data may consist in long audio recordings of several hours such as political speeches, radio/tv-shows, or audiobooks. Transcriptions can vary in terms of precision, reliability and completeness. We present the lightly supervised approach used to segment those long audio recordings with their associate transcriptions into pairs of text-audio sentence-level segment which can then be used for the training of statistical TTS system. We then introduce the different metrics used for the data selection including confidence measures derived from the segmentation process.

2.1. Lightly supervised alignment

We developed an audio alignment system specifically for Mandarin Chinese, following the same path than the one initially proposed in [11, 12] and refined in [13–22]. The goal of an audio alignment system is to recover time-stamps from the audio for words or syllables in the audio's transcripts. Mandarin Chinese is a tone language, where prosodic properties distinguish words from one another. More specifically, four prosodic patterns commonly referred to as "tones" function as phonemes. In other words, changes in prosody can change the meaning of the sentence in Mandarin Chinese. A Time Delay Neural Network based ASR system (TDNN [23]) was trained on a subset of the Mandarin Chinese RASC863 training dataset [24] considering tones as additional acoustic features. This subset is composed of 25h of clean read speech for 80 speakers from 4 dialectal regions (20 each). During the alignment process, the audio recording is first segmented into speech and non-speech segments using a baseline HMM-based speech activity detector trained on the TIMIT Corpus. Text transcripts are normalized and tokenized. The alignment is based on a lightly supervised approach [13] which consists in biasing the recognizer's language model (LM) to the content of the transcript. One biased language model (LM) at Chinese character-level is estimated on the transcription. The vocabulary was chosen to ensure coverage of characters from the transcription. Each segment is decoded using the biased LM as language model and the TDNN models as acoustic models. The resulting time-aligned text transcription is then aligned with the original text transcript. The aim is to associate time-stamps to the original text transcript and to partition the text and audio into smaller segments in order to reduce the alignment of a complete audiobook to the alignment of a set of small segments. The splitting must be performed in areas where we are highly confident that the original text transcript is correct, and it should not occur in the middle of an utterance. On the one hand, matching sequences of characters (also called anchor points [12]) between the decoding output and the original text transcript are a good indicator that the original transcript is correct. On the other hand, text transcripts provide a segmentation of the text into chunks of consecutively characters delimited by punctuation (full stop, exclamation, and question mark). In our approach, punctuation marks are used as potential split points, and a split is performed if positioned between two sequences of three consecutive matching characters and if the length of corresponding silence between the two matching sequences is longer than 150ms. Each segment of the obtained partition is then forced-aligned.

2.2. Selection Metrics

We broadly categorize the error types in found data into four main types: misalignment error, variation in channel conditions, features computation error and variation in articulation. *Misalignment errors* are due to bad transcription, poor acoustic models used for automatic alignment, poor acoustic condition or poor pronunciation. *Variation in channel conditions* are due to microphone conditions, channel noise, etc... As mentioned in [8], while misalignment errors are never good, the second ones might in some cases be good for training, adding to the diversity. *Features computation errors* are due to the poor estimation of speech features such as pitch or spectral envelope. Finally, consistency in speech in the training data is important for the naturalness of the synthesized speech. Highly expressive speech or unusual speaking style can result to lower naturalness of the synthesized speech and we categorize those as *variation in articulation*. In [25] the authors note that slow speaking rate and louder speech are associated with hyper-articulation. In [9], the authors hypothesized that hypo and hyper-articulation of training utterances might have an effect on the naturalness of the resultant voice. They showed that removing hyper-articulated outlier utterances improve significantly the naturalness of synthesized speech. The same authors showed in [10] that selecting from utterances based on metrics such as standard deviation of F0, fast speaking rate and hypo-articulation produce the most intelligible voices.

We now introduce the different metrics used to identify the four types of errors.

- *Phone Matching Error Rate* (PMER [26]) is used to assess the reliability of the transcripts and to detect potential misalignment errors resulting from the lightly supervised alignment procedure described in Section 2.1. It is computed by scoring the lightly supervised decoding output of a speech segment against the corresponding aligned transcripts used as reference (it is described as matched error rate since there are no accurate transcripts to be used as reference). If both text transcripts differ strongly, the PMER value is high and the original transcript is therefore considered unreliable. This does not necessarily mean that the transcript is incorrect given that the difference could be due to the poor performance of the speech recognizer systems for particular acoustic conditions.

- *Voiced/Unvoiced Mismatch Rate* is used to assess reliability of F0 estimation (feature estimation error) and phoneme boundaries precision (misalignment error). Each frame can be mapped to a phoneme based on forced-alignment output. A frame is voiced/unvoiced mismatched if it is a voiced phoneme but with zero F0 or it is an unvoiced phoneme but with non-zero F0. It is computed as given in Eq. 1. Its value is high if the performance of F0 estimation or forced alignment is poor.

$$\%\text{V/UV mismatch} = \frac{\#\text{V/UV mismatched frames}}{\#\text{non-silence frames}} \tag{1}$$

- *Signal-to-Noise Ratio* (SNR) is used to assess the recording quality (variation in channel condition). Voice activity detector is used to separate the segments into signal (speech) and noise (non-speech). Then SNR is computed as given in Eq. 2 where P_{signal} and P_{noise} are the signal power and the noise power respectively. Segments with low SNR are usually unwanted since they can induce poor alignment and bad feature estimates.

$$\text{SNR} = \frac{P_{signal}}{P_{noise}} \tag{2}$$

- *Articulation* [9] is used to assess variation in articulation and detect abnormal articulation. It is computed as given in Eq. 3 where P_{signal} is the power of speech part extracted by voice activity detector, and average syllable duration is computed based on forced alignment. Segments with high articulation are hyper-articulated. Hyper-articulated segments sound unnatural since they contain speech with slow speaking rate and high energy. Relation between articulation and naturalness of speech is discussed in [9, 25].

$$\text{Articulation} = P_{signal} \times \text{Avg. syl. dur.} \tag{3}$$

- *Standard Deviation of Syllable Duration* (std. syllable duration) is used to assess the consistency of speaking rate (variation in articulation). Duration of each syllable is acquired according to forced-alignment result. If standard deviation of syllable duration is high for a segment, it indicates that the narrator spoke sometimes fast and sometimes slow in that segment. Thus, it reflects the inconsistency of speaking rate within a segment.

- *Non-fluency* is used to assess the reading fluency (variation in articulation) and the quality of the alignment procedure (misalignment error). It is defined as in Eq. 4. Maximum internal silence duration and average syllable duration are computed according to the result of the forced-alignment. Internal silences are silences other than the start and end ones of the segment. The value of non-fluency is high if there is a long pause within a sentence which reflects the non-fluency. Note that a high value may also be due to the lightly supervised alignment process when a split between two sentences is not realized due to non-matching character sequence between the output of the lightly supervised decoding and the transcription.

$$\text{Non-fluency} = \frac{\max(\text{internal silence duration})}{\text{Avg. syllable duration}} \tag{4}$$

- *Standard Deviation of F0* (std. F0) is used to assess the variation of F0 and to detect potential estimation errors. Higher standard deviation of F0 might due to more expressive speech but can also be associated to poor performance of F0 estimator. Since octave error can occur in F0 it can affect standard deviation of F0 in higher range order compared to the expressive speech. Therefore, F0 is can be considered unreliable if its standard deviation is high.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

Two sets of audios were used, recorded by a Male narrator and a Female narrator respectively (henceforth *Audiobook A* and *Audiobook B*). **Audiobook A** was the recording of 60 passages from *Putonghua Shuiping Ceshi Shishi Gangyao* (, the official guide to the official test of spoken proficiency in Mandarin Chinese, intended for native speakers of Chinese languages) [27]. The (male) narrator was a Mandarin Chinese teacher; he read clearly and neutrally at a regular pace. However, some audios had perceptible noises in the background. In contrast, **Audiobook B** was the recording of Xun Lu ()'s novels and essays: *Call to Arms* (), *Chao Hua Si She* (), and *Hot Wind* () [28–30]. The (female) narrator was more expressive, and the recording environment was quieter compared to Audiobook A. With regard to the length of data, Audiobook A originally was 2.2 hours with a total of 29,096 Chinese characters; Audiobook B originally was 9.4 hours with a total of 122,644 Chinese characters. To simulate small corpuses, we randomly selected 1.5 hours of data from each audiobook.

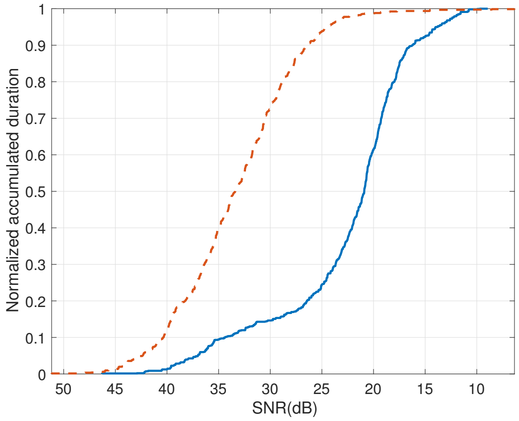
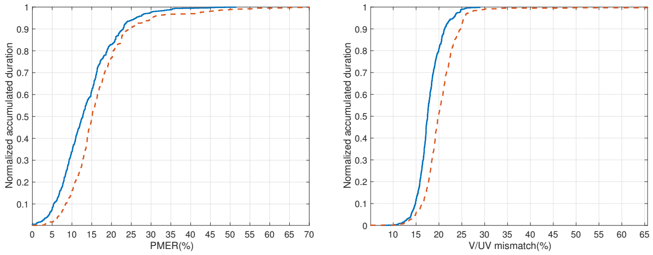


Fig. 1: Cumulative duration of the training data according to SNR; blue line: Audiobook A, dashed red line: Audiobook B.; threshold values corresponding to a selection 95% of data for Audiobook A and B are 14.03dB and 24.56dB respectively.

3.2. Data Selection

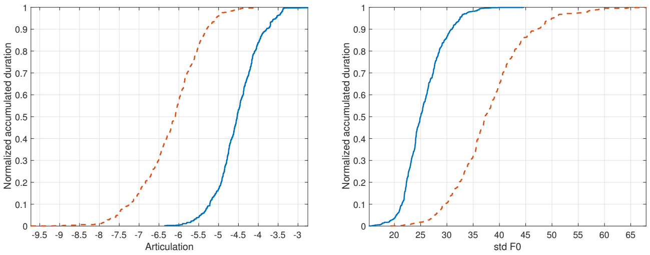


(a) PMER

(b) %V/UVmismatch

Fig. 2: Cumulative duration of the training data according to PMER and %V/UVmismatch; blue line: Audiobook A, dashed red line: Audiobook B. Threshold values corresponding to a selection 95% of data for Audiobook A and B are: a) 25.81% and 30.49%, b) 22.92% and 25.48%.

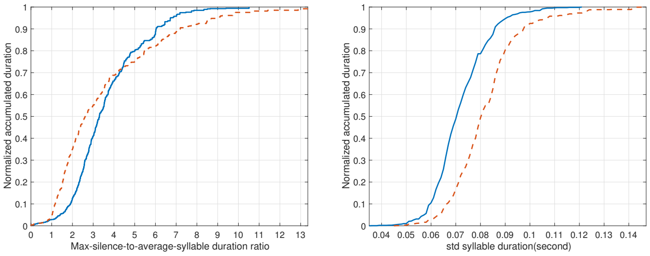
The cumulative duration of the training data for both audiobooks according to the different selection metrics presented in Section 2.2 is presented in Fig 1, 2, 3 and 4. Results related to Audiobook A and B are presented in blue full line and red dashed line respectively. For each metric we rejected the 5% of data corresponding to the segments with the worst metric's values. SNR-based selection results are presented in Fig. 1. It appears that the quantity of data with high SNR is larger for Audiobook B due to the cleaner recording condition. Regarding PMER-based selection results, Fig. 2a indicates that a larger quantity of data has a lower PMER for Audiobook A. This can be attributed to better transcription or to a better pronunciation by the narrator. Considering now the %V/UV mismatch-based se-



(a) Articulation

(b) Std. F0

Fig. 3: Cumulative duration of the training data according to articulation and Std. F0.; blue line: Audiobook A, dashed red line: Audiobook B. Threshold values corresponding to a selection 95% of data for Audiobook A and B are: a) -3.59dB and -5.05dB, b) 32.49Hz and 49.93Hz.



(a) Non-fluency

(b) Std. syllable duration

Fig. 4: Cumulative duration of the training data according to Non-fluency and Std. syllable durations; blue line: Audiobook A, dashed red line: Audiobook B. Threshold values corresponding to a selection 95% of data for Audiobook A and B are: a) 6.71 and 8.91, b) 0.09s and 0.10s.

lection results presented on Fig. 2b, there is more data with lower %V/UV mismatch for the Audiobook A which could be due either to the better alignment or to the better estimation of F0 (due to gender). Articulation-based selection results are presented on Fig. 3a. There is a larger quantity of data with higher articulation for Audiobook A. Hence is appears that the Male narrator has clearer articulatory movements while the Female narrator of Audiobook B sounds more expressive. This could be due to the fact that the Female narrator tends to reduce the speech strength in more places (hence more contrasts). Considering std. F0-based selection results presented on Fig. 3b, there is a larger quantity of data with low std. F0 value in Audiobook A than B. This enlights the fact that the Female narrator in B is more expressive. Non-fluency-based selection results presented on Fig. 4a show that there is more data for Audiobook B at the two ends compared to Audiobook A. This is mainly related to the fact the narrator in Audiobook A speaks more fluent and consistent. Finally, std. syllable duration-based selection presented on Fig. 4b indicates that there is a larger quantity of data with lower std. syllable duration for Audiobook A, which indicates that there are more variations in speaking rate in the Audiobook B which is indeed more expressive.

3.3. DNN architecture and training description

We use a deep feed-forward neural networks (DNNs) as a deep conditional model to map linguistic features to acoustic features directly [31]. The input features for all neural networks consisted of 737 linguistic features. 734 of these represented the linguistic context and the remaining 3 are for the position of the frame within the phone. The linguistic context includes quinphone identity, parts-of-speech and positional context of phoneme, syllable and word within a syllable, word and breath groups, respectively. The remaining 3 are within-phone positional information: the number of frames from the start and the end of the phone, and the total number of frames in the phone. The phone boundaries in the utterance was obtained from the lightly supervised forced-alignment procedure described in Section 2.1. WORLD [32] was used to extract 60-dimensional Mel-Cepstral Coefficients (MCCs), 5-dimensional band aperiodicities (BAPs), and fundamental frequency on log scale $\log F_0$ at 5 msec frame intervals. The output features of neural networks thus consisted of MCCs, BAPs, and $\log F_0$ with their deltas and delta-deltas, plus a voiced/unvoiced binary feature. Before training, the input features were normalized using min-max to the range [0.01, 0.99] and output features were normalized to zero mean and unit variance. The Merlin toolkit [33] was used for the training. The deep conditional model used for the mapping consists of 6 feed-forward hidden layers; each hidden layer has 1024 hyperbolic tangent units, with a linear activation function at the output layer, and a batch size of 64 for acoustic model. Learning rate was fixed at 0.002, warm-up momentum was 0.3, dropout rate was 0.05, and number of training epochs was 50. For duration model we used a Gradient Boosting Regression tree. At synthesis time, Maximum likelihood parameter generation (MLPG) was applied to generate smooth parameter trajectories from the de-normalized neural network outputs. Spectral enhancement in the cepstral domain was applied to the MCCs to enhance naturalness for subjective evaluation.

3.4. Objective Results

The objective results of the 7 systems trained on 95% data selection + 1 system trained on the whole 1.5h dataset (baseline) for both audiobooks are presented in Table 1. For Audiobook A, all systems trained on the different proposed selections achieve better results than the baseline in terms of voice/unvoiced error. The system trained on the selection done according to Articulation achieves better results for both spectrum and pitch. However, the system trained on the selection done according to Std. F0 gives significantly worst results than the baseline. For Audiobook B, there are significant improvements in term of F0 for systems based on selection done according to %V/UV mismatch, Articulation, Fluency, Std. syl dur and Std. F0, the latter two giving the best improvements. Selections based on PMER and %V/UV mismatch give the best results in term of duration, due to the reject of segments containing misalignment errors, leading to a better estimation of durations.

3.5. Subjective Results

To evaluate the naturalness of the synthesized speech, we used a crowd-sourcing approach by publishing a listening test online and distributing it to 57 participants. We then rejected some assessments for the following reasons: those done using loudspeakers (kept only those done with headphones or earphones); if sounds were not assessed or assessed before the participant even listened to the corresponding audio files. We finally obtained 48 valid assessments. We restricted our survey questionnaire to native speakers of Mandarin Chinese. Our survey consisted of pairwise comparisons between the baseline TTS system (trained on the complete 1.5 hours) and each of the TTS systems trained on one of the 7 data selections. 40 sentences

Table 1: Comparison of objective results considering the different data selections for both audiobooks. MCD: Mel-Cepstral Distortion. F0-RMSE is calculated on linear scale. DUR-RMSE is calculated in frames. V/VUV: voice/unvoiced error. Distortion of band aperiodicities are not presented due to no significant differences between selections.

Audiobook A	MCD (dB)	F0-RMSE (Hz)	DUR-RMSE (frames)	V/UV (%)
1.5h	5.96	18.95	6.96	7.61
SNR	5.95	19.14	7.01	7.54
PMER	5.95	19.06	7.05	7.59
VUVmismatch	5.96	18.99	7.02	7.36
Articulation	5.94	18.86	6.98	7.38
Std. F0	5.98	19.40	7.01	7.50
Fluency	5.95	19.20	6.92	7.56
Std. syl dur	5.98	18.89	7.05	7.50

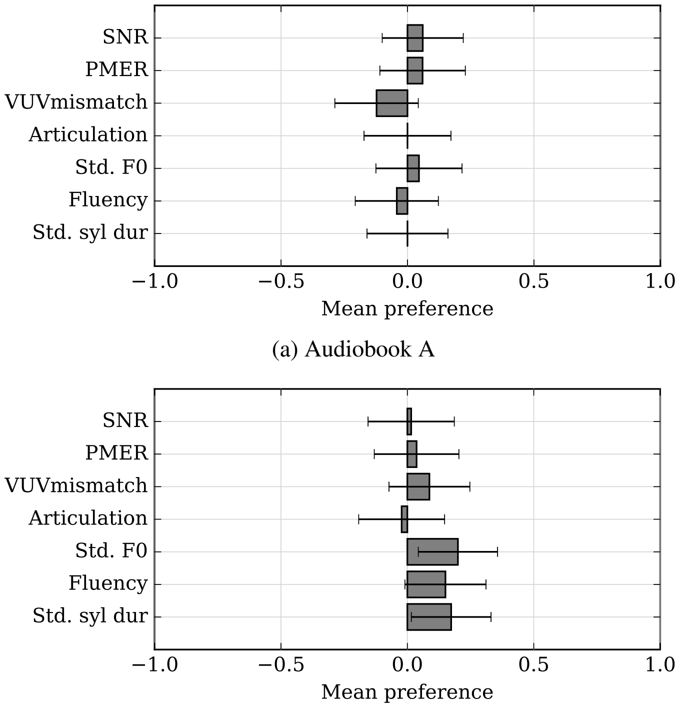
Audiobook B	MCD (dB)	F0-RMSE (Hz)	DUR-RMSE (frames)	V/UV (%)
1.5h	5.17	29.93	7.18	11.55
SNR	5.17	29.98	7.26	11.61
PMER	5.15	30.00	7.13	11.47
VUVmismatch	5.17	29.68	7.13	11.77
Articulation	5.19	29.58	7.21	11.78
Std. F0	5.17	29.11	7.24	11.44
Fluency	5.18	29.87	7.27	11.88
Std. syl dur	5.18	29.56	7.19	11.64

of varying length, not part of the training set, were used to generate audios from each TTS system. Each participant listened to 21 sentences in the Audiobook A's voice and 21 sentences in the Audiobook B's voice, such that every participant rated a total of 42 audio pairs. Then, for each selected sentence, the participant assessed only one of the 7 systems against the baseline. Participants were asked to compare the two audio recordings and choose the one they find better overall (including the audio quality, whether the pronunciation is clear, and whether the intonation and rhythm is natural). The order of presentation of each pair was randomly swapped to obtain either an A/B or a B/A order. Participants were also given a forced choice between A or B, i.e. there was not a "no preference" option. For each comparison (baseline vs. one of the 7 systems of interest), we analyzed participants' preferences using logistic regression, a statistical model commonly used for binary outcome variables.

Table 2: Significance test for the proposed evaluation. Lines marked with a '*' indicate significant results.

	Audiobook A			Audiobook B		
	z-value	p-value		z-value	p-value	
SNR	1.3060	0.1920		-0.1110	0.9120	
PMER	0.9590	0.3380		0.3130	0.7540	
VUVmismatch	-2.2810	0.0225	*	1.1240	0.2610	
Articulation	0.3880	0.6980		-0.4100	0.6820	
Std. F0	0.6640	0.5070		2.0090	0.0445	*
Fluency	-0.4190	0.6760		2.0750	0.0380	*
Std. syl dur	0.1890	0.8500		2.1520	0.0314	*

The subjective results of the 7 systems trained on the different 95% data selections for each audiobook are presented in Fig. 5. For Audiobook A, results don't show significant improvement for any of the selection metrics, even for articulation which was shown to give improvement in the objective results. Also, selection based on 1.5h is significantly preferred (p-value<0.05) over the system trained on the selection based on the %V/UV mismatch metric. Also, despite the fact that the recording conditions are more noisy, the SNR selection doesn't give significant improvement in naturalness.



(a) Audiobook A

(b) Audiobook B

Fig. 5: Results of the subjective evaluations on 48 listeners; a) Audiobook A, b) Audiobook B

For Audiobook B, system trained on selections made according to Std. F0, Fluency and Std. syl dur are significantly preferred to the baseline (p-value<0.05), which goes in the sense of the results obtained during the objective evaluation. Thus, articulation related selection metrics (i.e Std. F0, Fluency and Std. syl) give significant improvements on the Female audiobook, for which the speech is more expressive and hence subject to more variations which could potentially degrade the synthesized speech.

4. CONCLUSION AND FUTURE WORK

We investigated several metrics to detect different categories of errors in segments of training data that can degrade a text-to-speech system's performance. Small, 1.5-hour datasets were extracted from Mandarin Chinese audiobooks that had different characteristics (recording conditions, narrator, transcriptions). Our experiments showed that three metrics related to the narrator's articulation (Std. F0, Fluency and Std. syl dur) can significantly improve the naturalness of TTS systems trained on expressive audiobooks.

We are currently exploring combinations of the proposed selection metrics (rather than using them independently) with the aim to obtain additional improvements in naturalness. For future work, we plan to run experiments on more challenging data in terms of misalignment errors and channel conditions.

5. REFERENCES

[1] P. Lanchantin, C. Veaux, M.J.F Gales, S. King, and J. Yamagishi, "Reconstructing voices within the multiple-average-voice-model framework," in *Proc. Interspeech*, Dresden, Germany, 2015.

[2] N. Braunschweiler and S. Buchholz, "Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality," in *Proc. Interspeech*, 2011.

[3] 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.

[4] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. Clark, J. Yamagishi, and S. King, "TUNDRA: a multilingual corpus of found data for TTS research created with light supervision," in *Proc. Interspeech*, 2013.

[5] O. Watts, A. Stan, R. Clark, Y. Mamiya, M. Giurgiu, J. Yamagishi, and S. King, "Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from found data: evaluation and analysis," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013, pp. 101–106.

[6] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, , and S. Raptis, "Using audio books for training a text-to-speech system," in *Proc. LREC*, 2014.

[7] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, and Y.-J. Wu et al., "Thousands of voices for HMM-based speech synthesis-analysis and application of TTS systems built on various asr corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 984–1004, 2010.

[8] P Baljekar and A.W. Black, "Utterance selection techniques for TTS systems using found speech," in *Proc. SSW*, 2016.

[9] E. Cooper, Y. Levitan, and J. Hirschberg, "Data selection for naturalness in HMM-based speech synthesis," in *Proc. Speech Prosody*, Boston, Massachusetts, June 2016.

[10] E. Cooper, X. Wang, A. Chang, Y. Levitan, and J. Hirschberg, "Utterance selection for optimizing intelligibility of TTS voices trained on ASR data," in *Proc. Interspeech*, Stockholm, Sweden, August 2017.

[11] J. Robert-Ribes and R. Mukhtar, "Automatic generation of hyperlinks between audio and transcript," in *Proc. Eurospeech*, 1997.

[12] P. J. Moreno, C. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *International Conference on Spoken Language Processing*, 1998, vol. 8.

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

[14] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, VRR. Gadde, and J. Zheng, "An efficient repair procedure for quick transcriptions," in *Proc. ICSLP*, 2004.

[15] H.Y. Chan and P.C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proc. ICASSP*, 2004, vol. 1, pp. 737–740.

[16] L. Mathias, G. Yegnanarayanan, and J. Fritsch, "Discriminative training of acoustic models applied to domains with unreliable transcripts," in *Proc. ICASSP*, 2005, vol. 1, pp. 109–112.

[17] B. Lecouteux, G. Linares, P. Nocera, and J.F. Bonastre, "Imperfect transcript driven speech recognition," in *Proc. InterSpeech*, 2006, pp. 1626–1629.

[18] A. Haubold and J. Kender, "Alignment of speech to highly imperfect text transcriptions," in *IEEE International Conference on Multimedia and Expo*, 2007, pp. 224–227.

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

[20] S. Hoffman and B. Pfister, "Text-to-speech alignment of long recordings using universal phone models," in *Proc. Interspeech*, 2013, pp. 1520–1524.

[21] O. Boeffard, L. Charonnat, S. L. Maguer, D. Lolive, and G. Vidal, "Towards fully automatic annotation of audiobooks for tts," in *International Conference on Language Resources and Evaluation*, 2012.

[22] P. Lanchantin, P. Karanasou, M.J.F. Gales, X. Liu, L. Wang, Y. Qian, and C. Zhang, "The Development of the Cambridge University Alignment Systems for the Multi-Genre Broadcast Challenge," in *Proc. ASRU*, Scottsdale, USA, 2015.

[23] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[24] A-J Li and Z-G Yin, "Standardization of speech corpus," *Data Science Journal*, , no. 6, pp. S806–S812, 2007.

[25] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech Communication*, vol. 43, pp. 155–175, 2004.

[26] Y. Long, M. J. F. Gales, P. Lanchantin, X. Liu, M. S. Seigel, and P. C. Woodland, "Improving lightly supervised training for broadcast transcriptions," in *Proc. Interspeech*, 2013.

[27] "Putonghua Shuiping Ceshi Shishi Gangyao," http://www.beijingputonghua.com/psc/ldzp/ldzp.htm.

[28] "Call to Arms," https://librivox.org/call-to-arms-by-xun-luhttps://librivox.org/call-to-arms-by-xun-lu.

[29] "Chao Hua Si She," https://librivox.org/chao-hua-si-she-by-lu-xun.

[30] "Hot Wind," https://librivox.org/hot-wind-by-lu-xun.

[31] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[32] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, 2016.

[33] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. 9th ISCA Speech Synthesis Workshop*, 2016.