INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Selection and Training Schemes for Improving TTS Voice Built on Found Data

F.-Y. Kuo, I. C. Ouyang, S. Aryal, P. Lanchantin

ObEN Inc., CA, USA

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

Abstract

This work investigates different selection and training schemes to improve the naturalness of synthesized text-to-speech voices built on found data. The approach outlined in this paper examines the combinations of different metrics to detect and reject segments of training data that can degrade the performance of the system. We conducted a series of objective and subjective experiments on two 24-hour single-speaker corpuses of found data collected from diverse sources. We show that using an 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. Moreover, we show that training the system by fine-tuning from the system trained on the whole dataset leads to additional improvement in naturalness by allowing a more aggressive selection of training data.
Index Terms: TTS, data selection, found data

1. Introduction

A large quantity of *found data* is available, such as web data, public speeches, news and radio broadcasts, YouTube videos, phone conversations, and audiobooks. Those data can be of interest because a large quantity can be available for free and the naturalness of found data (e.g. spontaneous speech) can sometimes be higher than dataset dedicated to Text-To-Speech (TTS). Also, sometimes it is not possible to record the voice of a speaker because he doesn't have time, he's passed away or is for instance afflicted with Multiple Sclerosis, Parkinsons or Motor Neuron Disease (MND [1]). However, training TTS systems from found data is not straightforward because this kind of data can be really diverse in terms of audio characteristics and available transcription. Being not controlled, the speech recordings can show 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. Thus, it is of interest to develop data selection methods for producing high-quality voices from heterogeneous data sources.

Different selection approaches have been proposed for data selection in previous work. In [2], 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 based on the mean and standard deviation of pitch. They also discarded sentences with a low alignment score in order to remove poorly-aligned utterances. Combination of these approaches for a unit-selection system resulted in a better voice. Audiobooks are an other popular source of found data useful for the building of TTS systems [2–6]. In [5] controlled misalignment by selecting only utterances with high automatic alignment confidence scores and created a module for selecting utterances with a 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 [3], 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 [7], the authors showed that selecting a smaller, cleaner subset for voice building gave better results and was 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 were indicators of the measure of goodness of an utterance. In [8], 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 [9], 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 [10] they found that selecting training data based on speaker features rather than on individual-utterance features leads to the creation of overall more intelligible voices. They also found that an even more substantial improvement in intelligibility can be made by selecting training speakers based on a number of acoustic features combined. The significant improvements over the baseline obtained by more intelligently selecting training data indicates that some parts of a larger corpus, and in fact some speakers, are better suited for TTS than others.

In [11] we investigated several metrics to detect different categories of errors in segments of training data that could degrade the performance of a TTS system. Small, 1.5-hour datasets were extracted from Mandarin Chinese audiobooks that had different characteristics (recording condition, narrator, transcription). Our experiments showed that three metrics related to the narrator's articulation (Non-fluency, Std. syl dur and Std. F0) could significantly improve the naturalness of TTS systems trained on expressive audiobooks. In the current work, we explore different combinations of selection metrics and training schemes in order to obtain additional improvements in naturalness of TTS. This paper is organized as follows. In Section 2, we present the different metrics which are used for selection, in Section 3, we present the experiments on two Mandarin Chinese datasets of found data. Finally, Section 4 concludes our findings and discusses our future work in this direction.

2. Proposed approach

2.1. Dataset

We collected data for one male speaker (Speaker A) and one female speaker (Speaker B). Both of them are politicians so a large quantity of found data is available online. 30 hours of data

http://dx.doi.org/10.21437/Interspeech.2019-2816

were collected for each speaker. Speaker A's data consist of 68 different recordings and Speaker B's data include 114 different recordings. The quality and styles of their recordings vary a lot due to multiple factors. For example, the speeches given outdoors, at a stadium, in a lecture hall or in a studio have different noise levels and reverberation. One-on-one interviews are calmer, while political campaigns are emotionally heightened. Important occasions like international political events and casual occasions like guest lectures also lead to different speaking styles. In terms of recording devices, some of them were recorded by mobile phones, while some others were recorded by professional equipment such as shotgun microphones. Since the recordings spanned over a decade, their voices changed over time. Both speakers have Southern Chinese accents, or more precisely Taiwanese accents. All recordings were loosely transcribed to get rough transcripts which might include errors. Time boundaries of the utterances were extracted using a long text-audio alignment system for Mandarin Chinese based on a *lightly supervised approach* described in [11], following the same path as the ones proposed in [12–23], and about 27 hours of data were left for each speaker after the procedure.

2.2. Selection Metrics

In [11] we broadly categorized the errors and variations in found data into four main types: *Misalignment errors* due to bad transcription, poor acoustic models used for automatic alignment, poor acoustic condition, or poor pronunciation; *Features computation errors* due to poor estimation of speech features such as pitch or spectral envelope; *Variation in channel conditions* due to microphone conditions, channel noise, etc. And finally, *variation in articulation*, due to highly expressive speech or unusual speaking styles, can result in lower naturalness of the synthesized speech. To identify those errors and variations for data selection purposes, we use a set of 14 metrics defined in the following. *Phone Matching Error Rate* (PMER [24]) is used to assess the reliability of the transcripts and to detect potential misalignment errors resulting from the lightly supervised alignment procedure. It is computed by scoring the lightly supervised decoding output of a speech segment against the corresponding aligned transcripts used as reference. *Voiced/Unvoiced Mismatch Rate* is used to assess reliability of F0 estimation (feature estimation error) and phoneme boundary 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 the ratio of the number of voiced/unvoiced mismatched frames to the number of non-silence frames. *Utterance duration* is used to reject sentences that are too long or short which may not be desirable for the training. *Signal-to-Noise Ratio* (SNR) is used to assess the recording quality (variation in channel condition). It is computed as the ratio of the signal power to the noise power. *Mean Opinion Score Listening Quality Objective (MOS-LQO)* is used to assess the recording quality and the reliability of acoustic feature estimation from vocoder. MOS-LQO [25] is a MOS-scaled value transformed from ITU-T P.862 perceptual evaluation of speech quality (PESQ), which is an objective method for end-to-end speech quality assessment of telephone networks and speech codecs. Here, the original recording is used as the reference and its copy-synthesized audio is used as the degraded version. *Articulation* [8, 26] is used to assess variation in articulation and detect abnormal articulation. It is defined as the power of the speech portion multiplied by the average syllable duration. *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 the ratio of the average internal silence (other than the start and end ones of the segment) duration to the average syllable duration. *Standard Deviation Energy* is used to assess the consistency of energy. *Mean, standard deviation and mean absolute slope (MAS) of F0* are used to assess the range and the variation of F0 of the target speaker and to detect potential estimation errors (octave-error). Compared to Std. F0, MAS F0 focuses on the differences between adjacent frames. *Voiced Rate* is used to assess F0 estimation and to improve the intelligibility of TTS. It is computed as the ratio of the number of voiced frames to the number of non-silence frames. For Mandarin Chinese, each syllable consists of either only sonorants or sonorants preceded by an obstruent, so it is very likely to have F0 estimation error if the voiced rate is low. Moreover, it shows that training TTS with higher voiced rate data leads to more intelligible voice in [10]. Finally, *Mean and standard deviation of Syllable Duration* are used to assess the range and the consistency (variation in articulation) of speaking rate of the target speaker.

2.3. DNN architecture and training description

In order to evaluate the effect of the data selection made according to different combination of the metrics on the synthesized speech, several TTS systems were trained and evaluated as described in Section 3. For each system we used a deep feed-forward neural networks (DNNs) as a deep conditional model to map linguistic features to acoustic features directly [27]. The input features for all neural networks consisted of 616 linguistic features. 613 of these represented the linguistic context including quinphone identity, part-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. WORLD [28] was used to extract 60-dimensional Mel-Cepstral Coefficients (MCCs), 3-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 [29] was used for the training. The DNN model used for the mapping consists of 9 feed-forward hidden layers; the first 2 and the last 2 hidden layers have 1024 hyperbolic tangent units, and the rest have 512 hyperbolic tangent units. The output layer consists of linear activation function. Besides, the 3rd, 6th and 9th hidden layers includes batch normalization. Learning rate was fixed at 0.001, warm-up momentum was 0.5, dropout rate was 0.05, and number of training epochs was 100 while the batch size was 512. 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. Systems were trained on the selected data *from scratch* or by *fine-tuning* the acoustic model trained on full data. Our *fine-tuning* approach consists of continued training of a previously trained model in which the pa-

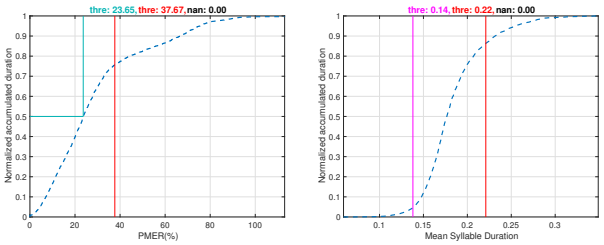

Figure 1: *Cumulative durations of the training data according to PMER (left) and mean syllable duration (right) for speaker A with knee points (magenta and red) and half-data point (cyan) thresholds.*

Table 1: *Metric thresholds used for selection, selection metric sets and corresponding quantity of data.*

	Metric thresholds				Selection metric sets					
Measure(x)	Speaker A		Speaker B		v1	v2	v3	v1	v2	v3
PMER	x<37.7	x<23.6	x<39.5	x<27.3						
V/UV	x<37.6	x<30.0	x<41.4	x<32.4						
SNR	x>12.0	x>27.8	x>16.9	x>23.3						
MOS-LQO	x>1.2	x>2.1	x>1.2	x>1.8						
Utt Dur	x∈]1;30[	x∈]1;30[	x∈]1;30[	x∈]1;30[						
Non-fluency	x<2.04		x<2.3							
Std Syl Dur	x∈]0.06;0.11[		x∈]0.05;0.10[							
Std F0	x∈]23.9;51.9[		x∈]30.2;76.7[							
Mean Syl Dur	x∈]0.14;0.22[		x∈]0.15;0.20[							
Mean F0	x∈]111.7;158.6[		x∈]138.8;211.4[							
MAS F0	x<9.0		x<12.2							
Std Energy	x<10.5		x<27.5							
Articulation	x<265.6		x∈]17.5;425.6[							
Voiced Rate	x>73.6		x>65.5							
			Speaker A (h)		11.5	7.1	4.8	2.7	2.0	1.5
			Speaker B (h)		12.1	8.1	5.4	4.6	3.2	2.3

rameters from first 4 hidden layers are frozen.

3. Experiments and Results

20 utterances (about 2.6 minutes) were selected from the 27-hour dataset as the test set. For each selection we selected the validation dataset in a 1:9 valid/train ratio. The selection threshold of each metric was determined based on the cumulative plot of the metrics as illustrated in Figure 1 for PMER and Mean Syllable Duration. We considered the *knee points* of the curve as the threshold to discard the worst data from one or both ends depending on the metric as shown in Table 1. For some measures (blue cells in the table), we also considered *half-data point* as the threshold for the selection. Half-data points were more aggressive and led to smaller selections than knee points. By combining the different metrics, we obtained 6 selection sets per speaker. A TTS system was built for each selection, trained from scratch and by fine-tuning from the system trained on the full dataset as described in Section 2, leading to 12 systems + 1 system trained on the full dataset per speaker.

3.1. Objective Evaluation

We evaluated the 12+1 systems trained for both speakers' datasets in terms of Mel-Cepstral distortion (MCD), root mean square error of pitch (F0-RMSE), root mean square error of duration in frames (DUR-RMSE) and voice/unvoiced mismatch (VUV). The results are presented in Table 2. In general, systems trained by fine-tuning lead to better results in terms of MCD and F0-RMSE. For speaker A it appears that globally data selection leads to better results in terms of F0-RMSE, duration and V/UV error, but worst in MCD. Conversely, for speaker B results are better in terms of MCD but worst in F0-RMSE and V/UV error except for the system trained on data selected by v1 half-data threshold using fine-tuning which achieves the best results.

Table 2: *Results from objective evaluation.*

Speaker A		MCD (dB)	F0-RMSE (Hz)	DUR-RMSE (frames)	V/UV (%)
full		6.17	33.56	12.97	15.64
v1 knee	scratch	6.19	32.08	12.70	16.53
	fine-tuned	6.13	31.97	12.70	15.50
v2 knee	scratch	6.23	31.95	12.67	15.31
	fine-tuned	6.16	31.60	12.67	15.23
v3 knee	scratch	6.29	31.61	12.94	15.04
	fine-tuned	6.22	32.47	12.94	15.16
v1 half	scratch	6.35	33.19	12.66	15.20
	fine-tuned	6.23	32.53	12.66	14.79
v2 half	scratch	6.39	32.06	12.62	15.01
	fine-tuned	6.24	31.94	12.62	15.15
v3 half	scratch	6.41	33.30	12.83	14.68
	fine-tuned	6.23	32.59	12.83	14.86

Speaker B		MCD (dB)	F0-RMSE (Hz)	DUR-RMSE (frames)	V/UV (%)
full		6.60	30.72	15.46	18.30
v1 knee	scratch	6.48	30.75	14.93	17.67
	fine-tuned	6.46	30.99	14.93	17.66
v2 knee	scratch	6.51	30.80	15.23	18.48
	fine-tuned	6.46	31.19	15.23	18.28
v3 knee	scratch	6.67	31.56	15.37	19.04
	fine-tuned	6.56	31.82	15.37	18.42
v1 half	scratch	6.43	30.49	14.86	19.19
	fine-tuned	6.34	30.48	14.86	17.91
v2 half	scratch	6.46	31.09	15.25	19.05
	fine-tuned	6.36	31.08	15.25	18.15
v3 half	scratch	6.59	31.28	15.47	18.95
	fine-tuned	6.45	30.81	15.47	18.90

3.2. Subjective Evaluation

To evaluate the naturalness of the synthesized speech, we conducted four listening tests (Tests 1-4) with native speakers of Mandarin Chinese. 35 sentences of varying length and structure, not part of the training set, were used to generate audios from each TTS system. Each test consisted of 36 pairs of audios, and were completed by 36-41 people. Participants were asked to compare the two audios in a pair and choose the one they found better overall, considering audio quality, enunciation, rhythm, as well as intonation. The following factors were fully randomized: the sentences used in each test, the order of presentation of the audio pairs in a test, and the left and right locations of the two audios in a pair. We analyzed participants' responses using logistic regression, a statistical model commonly used for binary outcome variables.

We first tested whether the TTS systems trained on selected data from scratch had performed better than the TTS systems trained on the full set of data. One of the two audios in each pair came from the TTS system trained on the full set of data; the other audio came from one of the following three kinds of TTS systems: systems trained from scratch on data selected by metrics set v1, v2, or v3. Results of test are presented in Figure 2(a). For each speaker, all evaluated data selections lead to significant improvement[1] in naturalness compared to the system trained on the full corpus. It also appears that the choice of thresholds and the speaker don't have effects on their own, but there is an interaction between them, e.g., for speaker A, synthesized speech is significantly more desirable in systems trained on larger selections made according to knee point thresholds rather than the more aggressive half-data thresholds (z=2.998, p=0.0109). Hence the main result of this test is that systems trained on selected data generate more natural speech than system trained on the whole dataset.

In the second listening test, we tested whether removing data according to knee-point thresholds or the more aggressive half-data thresholds had produced better TTS system. One of the two audios in each pair came from the TTS systems trained on data selected using knee point thresholds, and the other audio came from the TTS systems trained on data selected using half-data thresholds. Results are presented in Figure 2(b). Systems trained on smaller selections made according to the half-data thresholds are significantly more desirable than the ones according to the knee-point thresholds in Speaker B's fine-tuned

[1] full<v1 (z=3.917, p=8.98e-05), full<v2 (z=3.609, p=0.000307), full<v3 (z=2.960, p=0.003081)

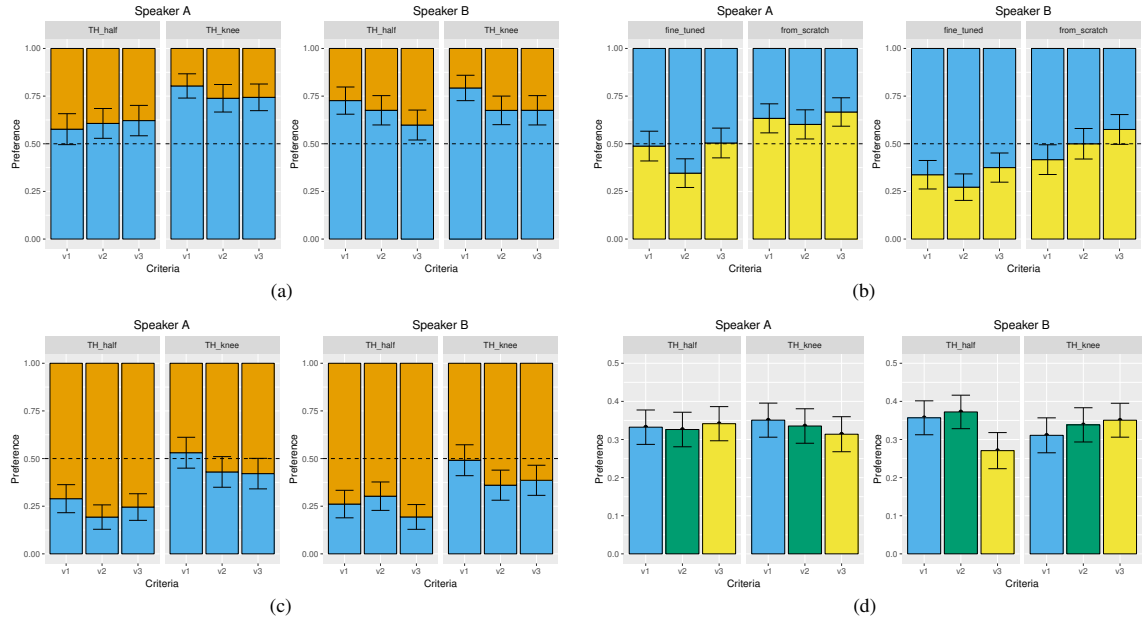

Figure 2: *Results of the subjective evaluation: (a) Test 1: full data (orange), selected data (blue); (b) Test 2: half-data (blue), knee point (yellow); (c) Test 3: fine-tuned (orange), from scratch (blue); and, (d) Test 4: v1 (blue), v2 (green), v3 (yellow)*

models (z=3.837, p=0.000497). However, systems trained on selections made according to the knee-point thresholds are significantly more desirable than the ones according to the half-data thresholds in Speaker A's trained-from-scratch models (z=-3.025, p=0.009953). Hence the main result of this test is that for systems trained using fine-tuning, selection can be more aggressive to select better data because less data is needed, whereas for systems trained from scratch, there is a trade-off between the amount of data and the quality of the selected data.

In the third test, we sought to determine if fine-tuned systems had produced more natural speech than the ones trained from scratch for all the selection sets. One of the two audios in each pair came from TTS systems trained from scratch using selected data, and the other audio came from TTS systems that were trained on the full dataset and fine-tuned on the selected data. The summary of the listener's preference are presented in Figure 2(c). It shows that speech generated from fine-tuned systems is significantly more preferable to trained-from-scratch systems (z=-1.979, p=0.0479). Also, the effect of fine-tuning is significantly larger in speech generated from smaller selection according to the half-data thresholds compared to the knee-point thresholds (z=3.218, p=0.00129). Finally, it appears that the effect of preference toward fine-tuned model is also modulated by the metrics set [2]. Those results are valid for both speakers as there is no significant difference between them. Looking at results of the second and third tests combined, performance of our TTS systems trained on the four kinds of selected data can be ranked in the following order, from highest to lowest: fine-tuned half threshold, fine-tuned knee thresholds, from-scratch knee thresholds, and from-scratch half thresholds.

In the last test, we attempted to determine the best set of selection metrics. Since the third test had shown that fine-tuned systems performed better than trained-from-scratch systems, we evaluated the metrics on the fine-tuned TTS systems. We used the Method of Paired comparisons [30, pp. 10–11] to rank these selection criteria on different experimental conditions. Listeners selected their preferred clip in a pair of audio that came from two of the three kinds of TTS systems: systems trained on data selected by metric set v1, v2, or v3. Figure 2(d) shows the perceptual score for each metric set given by the Method of Paired Comparisons. The results suggest that the new sets of selection metrics proposed in this paper (in addition to those from our previous work in [11]) did not help improve the quality. Further investigation is needed to determine the best combination of selection metrics.

[2] v3>v1 (z=2.439, p=0.0147); v2>v1 (z=2.559, p=0.0105); v3=v2 (z=0.144, p=0.8855)

4. Conclusion and Future Work

We investigated different selection and training schemes to improve the naturalness of synthesized text-to-speech voices built on found data. The objective and subjective experiments ran on two 24-hour single-speaker corpuses of found data collected from diverse sources showed that using a carefully selected set of data using a combination of metrics can lead to a text-to-speech system able to generate more natural speech than a system trained on the complete dataset. Moreover, we showed that training the system by fine-tuning from the system trained on the whole dataset leads to additional improvement in naturalness by allowing a more aggressive selection of the training data compared to systems trained from scratch on the same selection. In future work we plan to focus on finding an optimal combination of the metrics by running more experiments on the same datasets. We are also interested in investigating the benefits of our proposed data selection strategies in more recent TTS systems based on sequence-to-sequence mapping such as Char2Wav [31], Tacotron [32] etc.

5. Acknowledgements

The authors would like to thank Dr. Hsiao-Chi Ho and students at Graduate Institute of Education at Providence University, Taiwan for their help and participation in the subjective evaluation.

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

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

[4] 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.

[5] 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.

[6] 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.

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

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

[9] 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.

[10] K.-Z. Lee, E. Cooper, and J. Hirschberg, "A comparison of speaker-based and utterance-based data selection for text-to-speech synthesis," in *Proc. Interspeech 2018*, Hyderabad, India, 2018.

[11] F.-Y. Kuo, S. Aryal, G. Degottex, S. Kang, P. Lanchantin, and I. Ouyang, "Data selection for improving naturalness of tts voices trained on small found corpuses," in *Proc. IEEE Spoken Language Technology Workshop*, 2018, pp. 319–324.

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

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] "ITU-T recommendation p.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO," Tech. Rep., 2003.

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

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

[28] 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.

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

[30] Thomas C. Brown and George L. Peterson, *An Enquiry Into the Method of Paired Comparison: Reliability, Scaling, and Thurstones Law of Comparative Judgment*, Rocky Mountain Research Station Publishing Services, 2019.

[31] J. Sotelo, et al., "Char2Wav: End-to-end speech synthesis," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

[32] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of Interspeech*, 2017, pp. 4006–4010.