Utterance Selection Techniques for TTS Systems Using Found Speech

Pallavi Baljekar, Alan W. Black

Language Technologies Institute
Carnegie Mellon University
pbaljeka@cs.cmu.edu, awb@cs.cmu.edu

Abstract

The goal in this paper is to investigate data selection techniques for found speech. Found speech unlike clean, phonetically-balanced datasets recorded specifically for synthesis contain a lot of noise which might not get labeled well and it might contain utterances with varying channel conditions. These channel variations and other noise distortions might sometimes be useful in terms of adding diverse data to our training set, however in other cases it might be detrimental to the system. The approach outlined in this work investigates various metrics to detect noisy data which degrade the performance of the system on a held-out test set. We assume a seed set of 100 utterances to which we then incrementally add in a fixed set of utterances and find which metrics can capture the misaligned and noisy data. We report results 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. All of our experiments are carried out on male speakers. We also show comparable results are obtained on a female multi-speaker corpus.
Index Terms: Data selection, Found data, TTS, Telephone speech.

1. Introduction

Conventional methods in speech synthesis use a cleanly recorded, phonetically balanced dataset of recordings from a single speaker. However, if we want to build a text-to-speech (TTS) system as part of a speech-to-speech translation system to be used in a disaster relief effort in areas such as Nepal, Orissa or Pakistan where the languages spoken are dialects of Nepali, Oriya and Pashto, then we are severely limited in our ability to build such systems because we do not have access to a native speaker. On the other hand, we might want to build a personalized voice for a target speaker who has a limited ability to produce speech such as people afflicted with motor neuron disease or vocal cord paralysis. Such a voice would need to match the physical and dialectical characteristics of the speaker. Both of these scenarios are examples where we would like to build TTS systems but are limited by an access to a large, clean, phonetically balanced dataset of recordings from a single speaker, that matches our requirements.

To alleviate this problem, we propose to build systems from the abundant but noisy data available on the web. Speech data, especially from low resource languages, is becoming increasingly available on YouTube and other streaming services in the form of news and radio broadcasts, speeches, demo-videos, etc. In addition, with the growing popularity of neural networks and other big machine learning models, large multi-speaker datasets mainly for the purposes of speech recognition, have become available. Even if such data is unavailable on the web it can be crowd sourced via the web [1],[2]. If we can properly use this large variety of data available to us on the web, it would allow us to build voices matching the dialectical and physical characteristics of a particular speaker and also give us access to an array of languages allowing us to build TTS systems for low-resource languages.

Any data that is available readily in the public domain is called *found data*. This includes data from audiobooks, public speeches, news and radio broadcasts, Youtube data and telephone conversations. This data has large variations in its type and characteristics. Single speaker databases such as audiobooks, public speeches, channel broadcasts on Youtube, *etc.*, have the advantage that they are spoken by the same speaker. However in terms of diversity of the data they have their own challenges. Audiobooks for instance, have large variations in the prosody and intonation as well as the speech rate. Even though, these variations in prosody make for interesting listening, they are difficult to model in current speech synthesis systems. Moreover, since current systems are built using single isolated utterances, they do not account for long range dependencies as seen in audiobooks over paragraphs and sections. On the other hand public speeches and YouTube broadcasts have a lot of channel noise due to variations in recording conditions such as variations in microphone characteristics, room acoustics, distance from microphone, *etc*. These types of databases require better data pre-processing and feature adaptation and normalization techniques which will be robust to channel conditions.

Multi-speaker databases include news and radio broadcasts, telephone conversations and voice search data. These types of utterances are generally short, have a lot of fillers and non-speech sounds and distortions. News broadcasts generally have each segment consisting of multiple speakers and having many queuing audio sounds like sequences of music and other data. In terms of building TTS systems with such relatively clean, multi-speaker corpora we need to look at various voice averaging and voice conversion techniques to average over the different speakers in the database in addition to data pre-processing techniques in order to select an optimal subset for synthesis and remove speakers who are very different and not representative of our target speaker. Telephone conversations on the other hand provide a very rich set of data with multiple issues. They have low sampling frequency and so are of low quality. Furthermore, they contain lots of channel distortion, background distortions and noisy utterances containing long periods of silences between utterances, which causes problems with labeling and alignment. The advantage of using this type of data is it can be easily obtained. This kind of data needs good pre-processing methods in order to be useful to build TTS systems which do not fail during training.

In this paper we broadly categorize the error types found in such data, into two main types, misalignment errors and er-

rors due to variation in channel conditions. Misalignment errors occur because certain types of sounds are not described in the transcript such as claps, laughs, coughs, breaths, etc. The second type of error is caused due to varying microphone conditions, channel noise *etc*. Errors of type 1 are never good for the system and we would like to detect and remove utterances containing such errors from our training data . However, errors of type 2 caused due to channel variation and noise, might in some cases be good for training our models, adding to the diversity. Thus our goal here is to find good measures which detect for the misalignment errors and bad utterances which will be detrimental to the system, while retaining the sentences which even though they might not be representative of the training set, might still provide valuable and diverse characteristics to the training data.

This paper is organized as follows, In Section 2, we discuss how our work fits into work that has been done so far in this domain. In Section 3, we describe the datasets that we used and in Section 4 describe the metrics we used for utterance selection and our motivation in using these features. In Section 5, we report our results on the artificial data we created and the natural single and multi-speaker datasets. Finally, Section 6 concludes our findings and discusses our future work in this direction.

2. Related Work

Previous methods have looked at building speech synthesis systems from audio books [3],[4] and ASR corpora [5]. Most of these previous techniques have concentrated on techniques for building average voices using a variety of speaker adaptation techniques[6],[7],[8]. Even though results in this paper are reported on English corpora, our experimental design has been guided by the over-arching goal of building a system for low-resource languages, *i.e.*, languages where there is mostly noisy data available in the public domain such as public speeches and speech recorded over the telephone.

Furthermore, as far as we know there has been only one previous study so far on data subset selection tasks for statistical parametric synthesis from found speech [9]. This work investigates low-level acoustic descriptors as indicators of utterance naturalness and quality for selecting utterances. Similar earlier experiments on data selection techniques have been investigated for unit-selection synthesis, optimizing for coverage of linguistic units over utterances [10] and pruning out outliers post-hoc with respect to duration at the phoneme level [11].

In this paper, we show that using a small subset of data which is both *representative* of the target domain, but also diverse enough to be *informative* can produce better synthesis systems than using all of the noisy data available to us. We investigate some strategies to pick appropriate metrics to select good utterances. We show results on experiments carried out on an artificially degraded dataset to address the issues of misalignment errors and errors caused due to large non-representative channel conditions. We also evaluate the best metric found on the artificially degraded corpus on two corpora of found data, a single speaker corpus of public speeches and multi-speaker dataset of telephone speech.

3. Data

3.1. Artificially Degraded Clean Corpus

Since the goal is to evaluate metrics to detect misaligned data and data that is very different from the majority of the target domain such as data recorded in very noisy environments we created this dataset to act as ground truth and provide us an insight into how well different metrics perform.

To simulate misaligned data we shifted 100 utterances in the Arctic dataset for speaker RMS. For instance, the acoustics of utterance 1 corresponded to the transcription for utterance 10. To simulate varying channel conditions, we convolved another set of 100 utterances with an impulse response recorded inside a chamber. We then mixed the two sets to simulate having both misaligned data and channel noise in the data. The total duration of this data was about an hour of speech (1000 utterances) out of which ten minutes (100 utterances) were misaligned and the other 10 minutes had very different channel characteristics.

3.2. Single-Speaker Found Data

The single speaker dataset was created using speeches from the American President. We used subsets of three different speeches. The third speech was chosen to be very noisy with a lot of extraneous sounds such as claps and containing a lot of reverberation. The other two speech subsets were much cleaner in comparison. The total number of utterances in the dataset was 495 utterances totaling about 1 hour of training data. This subset of utterances was selected after running interslice [3] on each of the speeches and selecting utterances that had a minimum of four words. The utterances in this dataset tended to be longer on an average than the Arctic dataset.

3.3. Multi-Speaker Found Data

For the multi speaker found corpus we used the CallHome dataset and used all of the Males from this corpus. The CallHome corpus consists of 30 minutes of unscripted telephone conversations. For these experiments we used only the primary speakers from each of the conversations. We first created a clean set of about 400 utterances by removing sentences transcribed as containing laughs, breaths, fillers and other channel distortions. The noisy version consisted of about 900 utterances. The age range varied from 8 to 70 years for these conversations. The data was further pre-processed to remove too short or too long utterances. The same process was carried out for the Female CallHome corpus.

4. Rank and Select Methods

In this paper, we would like to answer the question as to whether all data is good data or can we do some pre-processing of the data in order to remove the really bad utterances and obtain improvements as measured in terms of Mean Cepstral Distortion (MCD) [12] between the original data of a held-out dataset and the same data synthesized from the model trained on a subset of the selected utterances. We therefore investigate various metrics for selecting the best subset that will help us build a good TTS system. The selection of an appropriate subset of training data can be done at various levels. It can be done at the utterance level, the phoneme level, the HMM state level as well as the frame level. In this paper, we have only investigated utterance level selection. Thus, the main aim in this paper is to find a *measure of goodness* of an utterance to be selected for training. We investigate two scenarios with respect to our experiments. Scenario 1 is where there is a seed set of utterances available, while scenarios 2 involves building a model out assuming no seed set is available.

4.1. Seed data selection

For the artificially degraded corpus we assumed the phonetically balanced subset of 100 utterances of the Arctic RMS utterances to be the seed. For the noisy datasets, since we did not have a seed dataset to begin with, we investigated different methods of finding this small seed subset of good utterances. We investigated using only acoustic measures such as the 100 best performing utterances in terms of MCD, and on the other hand purely optimizing for coverage of linguistic sub-units such as obtaining phonetically balanced subset by counting the frequency of occurrences for each HMM state. We also tried a combination of the two, however, we found that optimizing for linguistic coverage gave the best results in terms of MCD on a held out test set.

4.2. Voice Building

For all of the experiments in this paper, we have used CLUSTERGEN [13] for the parametric speech synthesis. In addition, we have only used the base voice building tools without using Move-Label [14] and Random Forests [15]. Given the noisy nature of the data we were not sure how robust it would be to use Random Forests, which might have given some performance improvements.

4.3. Metrics

We evaluated various metrics such as duration, spectral measures and other cross-correlation based measures that could be directly calculated from the synthesized wavefiles.

4.3.1. Mean Cepstral Distortion (MCD)

The mean cepstral distortion [12] is a weighted Euclidean distance between the true and the predicted Mceps and is evaluated for each predicted frame. We score each utterance by the frame-wise MCD averaged across the utterance. The main intuition in using the MCD was because this was a direct measure of the frequency content of the signal and thus a higher MCD would imply the synthesized wavefile is further away from the true wavefile in terms of predicted Mceps on the trained model. Thus, a really high MCD would imply that the acoustics in the utterance are not being modeled well and so would indicate misalignment errors and errors that are not representative of the majority of the training data.

4.3.2. Duration

For durations, we used the root mean square error between the predicted duration for each senone in the utterance compared to the *true* label given to an utterance after training 30 iterations using Baum Welch. The predicted durations were obtained from two models, one was from the entire noisy dataset and the other was on a model trained with a small seed set of a 100 utterances. The main goal here was to eliminate the really bad utterances which would in turn have bad labeling. This measure was expected to give good results on the misaligned data.

4.3.3. Modulation Spectrum

Since the modulation spectral trajectories capture the temporal dynamics of components of the spectral envelopes [16], we decided to investigate it as a global indicator of differences in spectral dynamics between the true and synthesized wavefiles. Moreover, using the Modulation Spectrum as a postfilter has shown gains in synthesis [17]. Thus, we expected this metric to be an informative measure in capturing large channel differences between recordings. We scored each utterance with the mean error between the modulation spectrum of the true and the synthesized wavefile in order to obtain a score per utterance.

4.3.4. Global Variance

This was another global spectral measure we tried since it improves results when used as a post-filter [18]. The idea here was to investigate whether differences in global variance of the predicted Mceps as compared to the true ones are correlated to the noisiness in the data.

4.3.5. Cross- correlation based measures

The main intuition in using this metric was to detect the misaligned data. If the two sentences are similar they should have a high peak when the two wavefiles, the true training wavefile and the synthesized wavefile are cross-correlated. We experimented with three cross-correlation based measures. We simply cross-correlated the two wavefiles and used the maximum of the resulting cross-correlated sequence as the measure. We also experimented by cross correlating the Teager Energy operator of the two wavefiles and the Hilbert envelope. The Teager energy operator gives a running energy estimate [19] of the wavefile and is supposed to capture the energy of the system that produced the speech rather that the energy in the speech itself. Thus, we hoped that this metric would also be helpful in capturing channel distortions. The Hilbert envelope on the other hand, computes the discrete-time analytic signal of the real part of a complex signal. The intuition in using this measure was to make it easier to detect the differences in misaligned data, since it can be used as a correlate to the envelope of the speech signal.

4.3.6. Instantaneous Frequency

The instantaneous frequency is supposed to capture the the average of the sinusoids at each point in time in a signal. The utterances were scored by taking the mean of the absolute difference between the instantaneous frequencies in the synthesized and the true wavefiles. So we would expect that signals that are similar acoustically will have smaller differences in error between the true and synthesized waveforms while signals differing in acoustics will have higher errors.

5. Empirical Evaluation

This section discusses the results on artificially degraded speech as well as noisy speech. We first describe the evaluation of various metrics on the artificially degraded data for both types of errors and a combination of the two. These metrics are evaluated on how accurately they can detect the artificially misaligned sentences and sentences that have been degraded with noise. The best performing metric is then used to iteratively select 10% of the best utterances which are then used to re-train the models. The models can be retrained at each iteration by re-clustering the state models while assuming fixed segmentation or the models can be re-trained by re-aligning the models each time using only the selected utterances and then re-clustering based on the new segmentation obtained. The baseline is calculated as the average MCD error on a held out test set from a model trained on all of the noisy data. In addition, we also investigate how this metric scales with larger amounts of noisy data.

5.1. Metric Evaluation

We carried out two sets of experiments. The results in Table 1 show results when using synthesized wavefiles obtained from a system trained with a small almost phonetically balanced seed dataset of about 100 utterances (approximately 10 minutes of speech). The results in Table 2 show results assuming the wavefiles are synthesized from a model trained on the entire dataset. Since the goal is to find the best metric which can detect the noisy utterances, we have reported the detection accuracy of each metric in detecting the artificially degraded utterances. Thus an accuracy of 90 implies that 90 out of the worst 100 sentences by some metric were the artificially degraded sentences.

Table 1: *Evaluation metrics on artificially degraded set assuming a model built from a* **small seed set** *of utterances. (% Accuracy of detection)*

Metric	Mis-alignment Noise	Channel Noise	Mixed Noise
Mean Cepstral Distortion	**99**	**88**	**95.0**
Duration	85	36	63.0
Modulation Spectrum	51	27	35.5
Global Variance	59	4	34.0
Cross-corr	87	1	32.5
TEO Cross-corr	20	43	51.5
Hilbert Env. Cross-corr	45	1	34.5
Instantaneous Freq.	53	2	17.5

Table 2: *Evaluation metrics on artificially degraded set assuming a model built from* **all** *utterances. (% Accuracy of detection)*

Metric	Mis-alignment Noise	Channel Noise	Mixed Noise
Mean Cepstral Distortion	**100**	**94**	**96.5**
Duration	87	43	67.0
Modulation Spectrum	21	25	31.5
Global Variance	86	4	47.0
Cross-corr	44	1	30.5
TEO Cross-corr	36	51	55.5
Hilbert Env. Cross-corr	39	1	33.5
Instantaneous Freq.	13	3	16.0

We see that the MCD and duration parameters outperform all of the other metrics on both misaligned data as well as channel noise. It makes sense given the fact that these are the two main parameters about the vocal production mechanism that we model. Moreover, we see that both of these measures do better on the model trained with the entire noisy dataset, because on the noisy dataset, the really bad utterances will be synthesized wrongly, while the good ones that fit the majority of the data will be synthesized nicely. However, it is not guaranteed that the model trained on the seed dataset encompasses all of the diversity in the data and so might even reject utterances which might provide it information and make it a better model.

The cross-correlation metrics do much better on the model trained with the clean seed subset of about 100 utterances. We find that it is much easier to detect misaligned data than it is to detect channel mismatch. The Teager energy operator is quite successful in detecting channel mismatch, as compared to all of the other correlation metrics.

5.2. Re-alignment vs. Re-clustering

The subset of utterances selected was used to either re-cluster and re-align the models to obtain labels suited to the iteratively changing data subset or fix the labels by training on the entire corpus of noisy data and then only re-cluster based on the utterances selected with the MCD metric. Figures 1-4 show plots of iteratively selecting the best 10% of utterances and using these utterances to either only re-cluster the state models (blue line) or re-align at each iteration and re-cluster based on new segmentation obtained (pink line) We find that realigning the labels each time helps and results in a much lower MCD on both of the single speaker datasets as illustrated in Figures 1 and 2. However, the trend is opposite in case of the multi-speaker corpora for both males and females as shown in Figures, 3 and 4. This might be because each subset has a different set of speakers and having labels trained from the entire data set might be better than having labels from a smaller set which does not encompass all of the diversity in the data. In addition, we find that in case of the noisy datasets as seen in Figures 2, 3 and 4, we see that the results with realignment are not monotonically decreasing upto a certain point and have a slightly erratic behavior. This might be because at every iteration, the alignment shifts based on the limited amount of training data, which in some cases might be representative of the test set , while in some others it might not be. In contrast, re-clustering the data, assuming fixed labels obtained by running Baum-Welch on the entire set always results in the MCD decreasing monotonically to a certain point and then increasing as bad data keeps getting added.

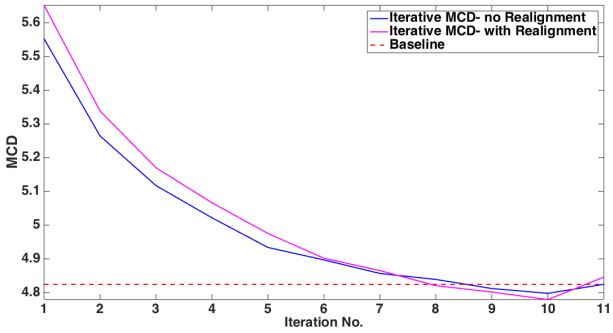

Figure 1: *Iterative MCD for artificially misaligned data*

Thus, we find that even though realigning data gives a higher increase in MCD, it is also more time consuming. In addition, there is no clear trend as to where we need to stop and ignore the data, unlike the case when only re-clustering on the data, in which case there is a clear point after which performance of the system degrades. However in all of the four plots, Figures 1 - 4 we see that the MCD metric with re-clustering and no re-alignment, is in fact doing better than the baseline. Moreover, the lowest MCD obtained on the artificially misaligned data is the same as obtained when testing on a model trained only with the clean data. This shows us that this metric is indeed rejecting the noisy data which is not helpful to the system.

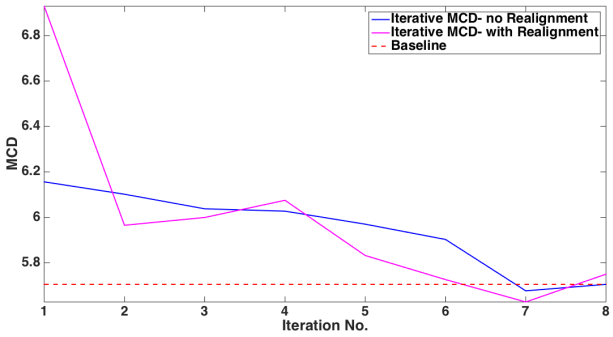

Figure 2: *Iterative MCD for Single speaker found data*

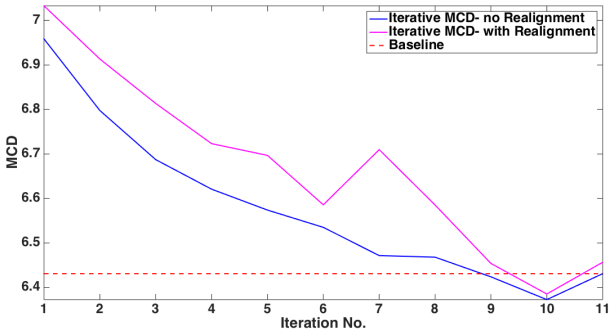

Figure 3: *Iterative MCD for multi-speaker found data-Male*

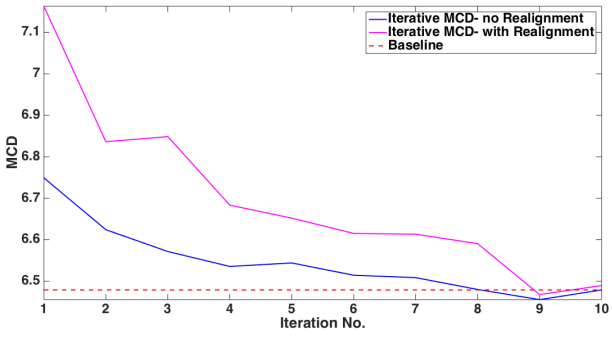

Figure 4: *Iterative MCD for multi-speaker found data-Female*

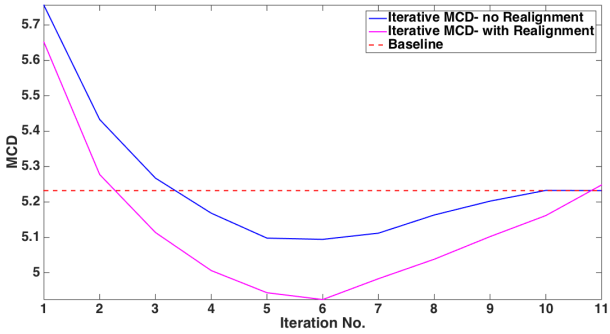

Figure 5: *Iterative MCD for Arctic RMS containing 50% misaligned data*

5.3. How does it scale?

The question then arises can this metric scale when larger amounts of data are misaligned and can this high performance in terms of accuracy hold up even when half the dataset has been artificially misaligned? From the figure, Fig.5, we see that yes indeed, this method based on pruning out misaligned data does work even when more than half the data is misaligned. In fact, the gain obtained over the baseline, a change of 0.3 MCD is significant. A 0.12 MCD change has been shown to be a significant, almost equal to the improvements obtained by doubling the amount of training data [12]. We also find that results with realignment which is almost 0.3 lower in MCD from the baselines is more beneficial in this case as compared to just re-clustering which is 0.14 MCD lower than the baseline.

6. Conclusions

In this paper, we show that our claim that not all data is good data holds. We see that selecting a smaller, cleaner subset for voice building is much better and less time consuming than building from the full noisy dataset.

We explore various utterance level metrics which would be indicators of the *measure of goodness* of an utterance. We show results considering both the availability of a small seed set of about 100 utterances and building from all of the noisy data without access to such a seed dataset. We find that there is no advantage in having access to a seed dataset and we can instead get similar if not better results by training our initial model on the entire dataset. From all of the measures we explored, we find that the mean cepstral distortion performs the best followed by the error in the duration prediction. We surprisingly find that the various cross-correlation based metrics are not good indicators of the presence of misaligned data. In addition, contrary to our expectations, we find the global spectral measures, *i.e.*, modulation spectrum and global variance perform poorly in detecting changes in channel conditions.

In terms of the errors we find that it is much easier to detect misaligned data than it is to detect noisy channel variations. We also find that the data selection does scale even when 50% of the dataset has been misaligned and yields a MCD which is 0.3 lower than the baseline. In addition, we find that re-clustering works better on the multi-speaker dataset while re-aligning works better and gives a lower MCD on the two single speaker corpora. However, when re-clustering on these datasets the MCD monotonically improves upto a certain point, while the behavior of the average MCD with each iteration is erratic when also realigning each time.

We show preliminary results in this paper on data selection. We show that it scales well on misaligned data and gives us significantly better results than using the entire subset of noisy data. In the future, we would like to test our methods on large non-English corpora such as the Babel[20] datasets, that contain speech recorded over telephones mainly for the purposes of keyword spotting.

Furthermore, in this paper we investigate various utterance-level metrics to detect and prune out utterances. In the future, we would like to investigate methods detailed in [21], which given pairwise dissimilarities between the source and target sets, tries to find a source set that best encodes the target set and can efficiently describe it. In addition, this paper details experiments only using one metric at a time, however in the future we would like to explore experiments with various metric combinations and see if we can improve the performance further.

7. References

[1] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," *Proc. SLTU*, pp. 194–200, 2014.

[2] A. Gutkin, L. Ha, M. Jansche, O. Kjartansson, K. Pipatsrisawat, and R. Sproat, "Building statistical parametric multi-speaker synthesis for bangladeshi bangla," *Procedia Computer Science*, vol. 81, pp. 194–200, 2016.

[3] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1444–1449, 2011.

[4] 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.

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

[6] J. Yamagishi, T. Kobayashi, S. Renals, S. King, H. Zen, T. Toda, and K. Tokuda, "Improved average-voice-based speech synthesis using gender-mixed modeling and a parameter generation algorithm considering GV," pp. 125–130, 2007.

[7] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[8] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 66–83, 2009.

[9] E. Cooper, Y. Levitan, and J. Hirschberg, "Data selection for naturalness in HMM-based speech synthesis," in *Speech Prosody*. The International Speech Communication Association, 2016.

[10] A. W. Black and K. A. Lenzo, "Optimal data selection for unit selection synthesis," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[11] J. Kominek and A. W. Black, "Impact of durational outlier removal from unit selection catalogs," in *5th ISCA Workshop on Speech Synthesis*, 2004.

[12] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion." in *SLTU*, 2008, pp. 63–68.

[13] A. W. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Ninth International Conference on Spoken Language Processing*, 2006.

[14] A. W. Black and J. Kominek, "Optimizing segment label boundaries for statistical speech synthesis," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3785–3788.

[15] A. W. Black and P. K. Muthukumar, "Random forests for statistical speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] H. Hermansky, "Modulation spectrum in speech processing," in *Signal Analysis and Prediction*. Springer, 1998, pp. 395–406.

[17] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, April 2016.

[18] T. Tomoki and K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[19] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990, pp. 381–384.

[20] "IARPA broad agency announcement IARPA-BAA-11-02," https://www.fbo.gov/utils/view?id=ba991564e4d781d75fd7ed54c9933599, 2011.

[21] E. Elhamifar, G. Sapiro, and S. Sastry, "Dissimilarity-based sparse subset selection," 2014.