A Comparison of Speaker-based and Utterance-based Data Selection for Text-to-Speech Synthesis

Kai-Zhan Lee, Erica Cooper, Julia Hirschberg

Columbia University, USA

k12792@columbia.edu, ecooper@cs.columbia.edu, julia@cs.columbia.edu

Abstract

Building on previous work in subset selection of training data for text-to-speech (TTS), this work compares speaker-level and utterance-level selection of TTS training data, using acoustic features to guide selection. We find that speaker-based selection is more effective than utterance-based selection, regardless of whether selection is guided by a single feature or a combination of features. We use US English telephone data collected for automatic speech recognition to simulate the conditions of TTS training on low-resource languages. Our best voice achieves a human-evaluated WER of 29.0% on semantically-unpredictable sentences. This constitutes a significant improvement over our baseline voice trained on the same amount of randomly selected utterances, which performed at 42.4% WER. In addition to subjective human voice evaluations with Amazon Mechanical Turk, we also explored objective voice evaluation using mel-cepstral distortion. We found that this measure correlates strongly with human evaluations of intelligibility, indicating that it may be a useful method to evaluate or pre-select voices in future work.

Index Terms: speech synthesis, deep learning, parametric synthesis, data selection, intelligibility, found data, crowdsourcing

1. Introduction

Text-to-speech (TTS) synthesis is a crucial component in modern technology such as virtual personal assistants and home devices, GPS navigation, and speech-to-speech translation. There are close to 7,000 languages in the world, however only a few have received the research attention and data collection efforts required to create a TTS system. Data collection for TTS typically entails recording a single professional speaker reading for many hours in as even and neutral a style as possible using a high-quality microphone in an anechoic chamber. Building a high-quality, commercial-grade TTS voice for a new language costs around one million dollars, and thus this is usually only an effort that is undertaken with major economic motivation. However, with recent advances in statistical parametric speech synthesis (SPSS) such as neural network based synthesis, it is possible to build voices using more heterogeneous data. Thus, it may be possible to make use of other sources of data which have already been created or collected for other purposes to build voices in new languages.

In our prior work [1], we explored the use of speech in the form of short utterances read by many speakers over the telephone, a type of data which is typically collected for training automatic speech recognition (ASR) systems and which is readily available in many languages that do not necessarily have a professional TTS corpus. We experimented with a data selection approach in which we attempted to identify a subset of utterances out of the entire corpus which would be the most suitable for building TTS voice models. In this work, we extend those experiments by investigating data selection at the *speaker* level in addition to the utterance level; we want to determine whether speech from certain speakers in the corpus is better suited to building voice models, and also to test whether the consistency gained by training on more data from each speaker and using fewer speakers improves synthesis output.

2. Related Work

Previous work on selecting the best data from a noisy or inhomogeneous corpus has typically involved removing the noisiest utterances and choosing the most neutrally-spoken portions of the data. Audiobooks and radio broadcast news have been popular sources of found data [2, 3, 4, 5] due to their relatively clean recording conditions and the fact that they typically contain large amounts of speech from a single speaker. Corpora designed for automatic speech recognition have also been explored for building HMM-based TTS voices; in particular, [6] built TTS voices on various ASR corpora containing cleanly-recorded read speech, as well as some corpora containing speech in a noisy environment, with the goal of being able to create "thousands of voices" from the many speakers in each corpus. They examined the tradeoffs between amount of data and voice quality, finding that it is better to train on data from multiple speakers and adapt to target speaker data when less than an hour of target speaker is available, and it is better to train a speaker-specific voice if more than two hours of speaker data is available. [7] identified noisy and misaligned utterances in a corpus of conversational telephone speech by measuring mel cepstral distance (MCD) between original utterances and utterances synthesized by a model trained on all of the data, in order to find utterances that are outliers with respect to the overall data.

Most work on building TTS from found data focuses on utterance-level selection. There has however been some work in selecting the best *speakers* for building a voice. [8] identified speakers that were acoustically similar for building a Bangla voice. They auditioned 15 speakers, did a crowdsourced evaluation to identify the speaker most preferred by listeners, and then picked 5 additional speakers from the original 15 who had similar vocal characteristics. [9] used human judgments of perceptually-similar source speakers, as well as objective measures, for building an average voice model (AVM) to build a better adapted voice than one based on an AVM trained on all the source speakers. Similarly, [10] aimed to select the best adult source speakers for building an AVM to adapt to child speech based on the objective measures of MCD and RMSE of log f0, finding that these measures also correlated well with human judgments of intelligibility for the adapted voices. In our work, we aim to identify speakers in ASR data that are both similar to each other and suitable for TTS, by selecting speakers based on a number of novel acoustic and prosodic features.

We also aim to compare speaker-level training data selection to utterance-level training data selection to determine which produces better voices, with the hypothesis that speaker-level selection, by using more data from fewer speakers, will produce more consistent and thus more intelligible-sounding models.

3. Tools and Corpora

We used the 160-hour Macrophone corpus of "approximately 200000 [transcribed telephone speech] utterances by 5000 [American English] speakers." [11] The majority demographic of Macrophone are adult, female speakers, comprising 83.5 hours of speech. The Macrophone corpus was designed for the development of ASR for telephone-based dialogue systems such as travel booking and other database-related tasks, and consists of short phrases and queries related to these types of tasks. There are 4005 female speakers in the corpus. The mean number of utterances per speaker is 40.7, with a standard deviation of 4.27. The minimum number of utterances per speaker is 1, the maximum is 44, and the median is 42.

We used the University of Edinburgh's deep-learning-based speech synthesis model ("voice") toolkit, Merlin [12], to train all of our voices. Each voice has two associated deep-learning-based models, both trained on time-aligned context-dependent label files containing phonetic and linguistic features generated from raw text using Festival [13] with EHMM alignment [14], and acoustic features (frame-level log f0, mel-generalized cepstra, and band aperiodicity features) extracted by the WORLD vocoder from the original audio. With a final goal of mapping phoneme sequences to acoustic features, the trained *duration model* predicts the duration of each phoneme for synthesis and the *acoustic model* generates acoustic features for each phoneme, which are ultimately converted to audio. We based our training on the "build your own voice" recipe from Merlin. These models consist of 6 TANH layers each of size 1024, with a linear activation function at the output layer, and a batch size of 64 for the duration model and 256 for the acoustic model. Learning rate was fixed at 0.002, momentum was 0.3, and number of training epochs was 25. We performed a standard 10:1:1 train/development/test set split for all voices trained.

Praat [15], a toolkit for phonetic and acoustic analysis and labeling of audio files, was used to extract standard acoustic features from each utterance for the purpose of intelligent training data subset selection. The set of features we extracted for each utterance and for each speaker's combined data were f0 (min, max, mean, median, standard deviation), mean absolute f0 slope (MAS), energy (min, max, mean, and standard deviation), and ratio of voiced to total frames.

For preliminary objective evaluation, we used the IBM Watson Speech to Text service [16] to achieve a rough measure of a voice's intelligibility in terms of word error rate (WER). Synthesized utterances were sent to Watson ASR servers via the API for transcription, from which text transcriptions were returned. For subjective evaluation, we used Amazon Mechanical Turk (MTurk), a popular crowdsourcing platform, to evaluate a voice's intelligibility with human listeners. We used the Python library `mcd` [17] to explore the robustness of mel-cepstral distance (MCD) as an objective voice evaluation metric as well. The `editdistance` package [18] was used to efficiently compute Levenshtein distance for WER.

GNU Parallel [19] was used throughout our work to parallelize computation when possible. SoX [20] was used throughout for direct manipulation of audio files such as concatenating all utterances from a single speaker.

4. Experiments and Results

4.1. Single-feature Speaker and Utterance Selection

In our first set of experiments, we examined one feature at a time as selectors for choosing training speakers and utterances. We extracted the 11 aforementioned Praat acoustic features for all of the data from each speaker. We also calculated speaking rate in syllables per second based on the syllable information in the label files obtained with Festival, and level of articulation, which we defined as mean energy divided by speaking rate, to encode the loud and slow speech that characterizes hyperarticulation. Furthermore, we obtained an approximate measure of which speakers were most intelligible, by running their speech through the Watson ASR API and calculating a WER for each speaker. We also computed the same features at the utterance level as well, in order to compare speaker- vs. utterance-level data selection. Our baseline voice was the first 10 hours of female Macrophone data, with no feature-based data selection. Our test voices were trained on 2-, 4-, and 10-hour subsets chosen by these features and selected out of the entire 83.5 hours of female data. For example, for the 4-hour high-clustered mean f0 speaker-selected voice, we sorted all female speakers by their mean f0 and accumulated data from the speakers by their distance from the highest mean f0 speaker one by one until we had a training set containing 4 hours of data. We trained voices for low, median, mean, and high-clustered values of each of our 14 features for both speaker-selected and utterance-selected subsets. After observing that our best speaker-selected voices were all trained on 10 hours of data as opposed to smaller amounts, we then only trained 10-hour voices for utterance selection.

Due to the slow turnaround time for evaluating synthetic voices for intelligibility by human transcription, we did an initial automatic evaluation of all of our voices using the Watson ASR API, which we have found in our prior work to correlate well with human judgments for intelligibility of synthetic voices [1]. Synthesized test utterances consisted of eleven 7-word semantically-unpredictable sentences (SUS) of the standard form `det adj noun verb det adj noun`, in order to prevent contextual recognition of words. After obtaining a set of hypothesized transcriptions for each voice, we computed the Levenshtein distance between each utterance's highest-confidence transcription and the corresponding true sentence; taking the average of the Levenshtein distances yielded a WER for each voice. Our baseline voice had a WER of **82.9%**. 74 out of the 168 total speaker-selected voices, and 28 out of the 56 total utterance-selected voices, had smaller WERs than the baseline voice. For brevity, the top 5 voices for each selection method, with their WERs from Watson, are in Tables 1 and 2. All of these voices were trained on subsets of 10 hours of data.

Table 1: *Watson WERs for 5 Best Speaker-Selected Voices*

Speaker feature	Cluster	WER
WER	Low	58.4%
Voiced vs. total	High	59.7%
Mean energy	Low	61.0%
F0 MAS	Avg	61.0%
Min energy	Low	62.3%

Next, it was necessary to corroborate our automatic evaluation with human judgment. This is especially important for the voices trained on subsets selected based on low ASR WER;

Table 2: *Watson WERs for 5 Best Utterance-Selected Voices*

Utterance feature	Cluster	WER
Mean energy	High	67.5%
Mean F0	Median	68.8%
Mean energy	Median	71.4%
Max F0	High	74.0%
Speaking rate	Low	74.0%

we would expect those voices to do well when evaluated by the same ASR, as they did, but we need to verify whether humans perceived the best automatically-rated voices as being high quality as well. We created an MTurk HIT (human intelligence task) using our baseline voice, our 5 best speaker-selected voices, and our 5 best utterance-selected voices, as well as one semantically-predictable sentence spoken clearly by one of the authors, intended as an attention check for MTurk workers. We synthesized the same set of 11 SUS with each of these 11 voices, and we presented them to workers in a Latin-square configuration so that each worker would be presented with each sentence once, each spoken by a different voice, so that any listener differences or bias would be averaged out over all of the voices. We restricted our task visibility to the United States, and included a qualifier that asked which languages each participant had spoken since birth, to select for native English speakers. Workers were permitted to play each audio file only twice, and were requested to transcribe what they heard. Our 11 tasks were completed by 5 workers each, and no worker was permitted to do more than one task, since they all used the same sentences and we wanted to eliminate bias that would arise from having previously heard the sentences. We compared the transcriptions to the actual text of the sentences to obtain a WER for each voice, averaging over the 5 transcriptions for each sentences. Results as well as p-values from a two-tailed t-test in comparison with the baseline are reported in Table 3.

Table 3: *MTurk Results for Single-Feature Utterance- and Speaker-Selected Voices*

Unit	Selection Feature	Cluster	WER	p-value
Spkr	WER	Low	**41.8%**	**0.0145**
Spkr	Voiced vs. total	High	43.4%	0.0175
Spkr	Mean energy	Low	43.4%	0.0216
Spkr	Minimum energy	Low	43.6%	0.0275
Spkr	F0 MAS	Average	46.2%	0.1037
Utt	Mean F0	Median	54.3%	0.9618
	[Baseline]		54.5%	
Utt	Mean energy	Median	55.3%	0.8793
Utt	Mean energy	High	57.7%	0.5354
Utt	Speaking rate	Low	62.9%	0.0882
Utt	Max F0	High	63.1%	0.0971

We observed that our 5 best speaker-selected voices were all rated as more intelligible than all 5 of our best utterance-selected voices. Furthermore, all speaker-selected voices were rated as more intelligible than the baseline, with the top four obtaining significance at $p < 0.05$. The best performing voice was trained on the 10 hours of data from the speakers with lowest WER as determined automatically by Watson ASR, indicating that this ASR matches well with human perception of speech for intelligibility, and that selecting more intelligible speakers for training data does produce more intelligible voices. Training on speech with a greater proportion of voiced data also produced a more intelligible voice, and lower energy levels appeared to be a useful selector as well.

Comparing the features that did well for utterance selection to features that did well for speaker selection, it is interesting to note that they are different. For instance, Watson WER was not in the top 5 features for utterance selection. This indicates that selecting training data based on which *speakers* are most intelligible is a good approach, but training on only the most intelligible *utterances* regardless of speaker perhaps does not result in as cohesive a training set. Looking at which best features are common across utterance and speaker selection, mean energy stands out as appearing in both sets, with low mean energy being a good selector for speakers, and both middle and high mean energy appearing to be good selectors for utterances. This indicates that perhaps the actual energy level is not so important, but that having a similar energy level across your training data will produce a better voice.

4.2. Joint-feature Speaker and Utterance Selection

Next, we investigated whether combining features would be better than using individual features. We tested combinations of the top 2 to 7 features, using them to assign each utterance a heuristic score and selecting 10 hours worth of highest-scoring data (utterances or speakers) for training. We took z-scores of the negated absolute difference between each feature and its cluster statistic (maximum, median, mean, or minimum), then tested 5 different methods to combine these z-scores in a heuristic: sum, product (after normalizing the minimum value to 0), sigmoid product, log sum, log product. These heuristics were also computed at both the speaker and utterance level. After voice training, we used IBM Watson once again to evaluate them, identifying the top 5 speaker-selected and top 5 utterance-selected voices. We then posted another MTurk HIT in the same format using the same check question and newly-generated SUS utterances for evaluating 13 selected voices in total: the aforementioned 10 voices, the baseline voice, the best voice from the previous HIT, and a human voice reading the same sentences. We added a human voice to this iteration to determine a lower bound for WER, relative to the other voices. Results are located in Table 4.

Table 4: *MTurk Results for Joint-Feature-Based Voices*

Unit	N ftrs	Combination	WER	p-value
		[Human]	*5.3%*	*$P < 0.0001$*
Spkr	4	Sum	**29.0%**	**0.0004**
Spkr	2	Sum	31.9%	0.0030
Spkr	4	Sigmoid	32.7%	0.0079
Spkr	2	Sigmoid	33.8%	0.0153
Utt	2	Product	38.2%	0.1249
Utt	2	Sum	40.4%	0.3176
Utt	3	Sigmoid	40.7%	0.3466
Spkr	3	Log Product	41.3%	0.3751
		[Previous Best]	42.4%	0.5624
Utt	7	Log Product	44.4%	0.8785
		[Baseline]	45.1%	
Utt	5	Log Product	46.8%	0.7084

We found that selecting a subset based on the top four features combined by summation of z-scores led to the most in-

telligible voice. Four out of five of the speaker-selected voices outperformed all of the utterance-selected voices. We also generally find that it is possible to substantially improve intelligibility by selecting based on a combination of our features, rather than just a single feature – most of our new best joint-feature-based voices outperformed our best single-feature voice from the previous set of experiments.

4.3. Subset Characteristics

We were interested in exploring the characteristics of our different training data subsets, in addition to their original selection features; thus, we examined subsets' utterance counts. Two subsets that both add up to the same total amount of time may have a different number of utterances. Fewer utterances in a subset means that the average utterance length is larger than the average utterance length of another subset with the same total amount of time and a higher utterance count.

Table 5: *Subset Statistics: Utterance Counts*

Set	Mean	Std	Min	Med	Max
2hr utt	2000.9	519.1	1084	1931.5	3527
2hr spkr	2034.8	185.8	1441	2066.5	2376
4hr utt	4007.6	835.8	2736	3895	6361
4hr spkr	4068	329.1	2880	4112.5	4766
10hr utt	9985.5	1620.7	7448	9706.5	14238
10hr spkr	10202.5	668.6	7673	10294.5	11640
Joint utt	8520.5	1162.8	6862	8547.5	10568
Joint spkr	10927.2	632	9484	11071.5	11532

We observed that utterance-selected subsets tend to have a comparatively higher variance and wider range in the number of utterances per subset compared to speaker-selected subsets, for the same total amount of audio. On average, utterance-selected subsets have a 140% greater standard deviation than the comparable speaker-selected subsets (e.g. 2-hour utterance- vs. speaker-selected subsets). We also observe that the mean number of utterances per subset is consistently lower for utterance vs. speaker selection, most notably for the joint-feature-based voices, which sees an increase of 22% in mean utterance count from speaker- to utterance-selected subsets. When computing correlations between utterance count and Watson WER for the voice trained on that subset, we found strong -64.5% correlation among all speaker-selected voices and -17.2% correlation among utterance-selected voices; with joint selection, correlations strengthen to -78.2% and -28.1%, respectively. These results suggest that, especially in the case of speaker selection, that training on more short utterances is better than training on fewer, longer utterances, and that utterance length may be a selection criterion that we should explore in the future.

4.4. MCD for Evaluation

Because low-resource languages do not necessarily have high-quality or reliable ASR, we experimented with the mel-cepstral distortion (MCD) objective function [21]. This function measures the difference between two time-aligned mel-cepstral sequences and is commonly used as an objective evaluation metric for TTS. For unaligned sequences, dynamic time warping may be performed to align the sequences before comparing them.

We computed average MCD on the validation set of each voice and found a 70.1% correlation between MCD and IBM Watson WER among all voices; correlation increased to 75.6% when considering only the top ten MCD-ranked voices.

Among voices selected for MTurk HITs, there was strong correlation of 80% to 90% between MTurk WER and Watson WER. MCD and MTurk WER had a moderate correlation of 43.0% for single-feature voices; however, among the higher quality joint-feature-based voices, correlation increased to 91.2%, which is quite promising for future objective voice evaluation. Results are located in Table 6.

Table 6: *Evaluation Method Correlations*

Comparison	Single-Feature	Joint-Feature
Watson, MCD	0.312	0.748
Watson, MTurk	**0.814**	0.899
MCD, MTurk	0.430	**0.912**

5. Conclusions and Future Work

We have generally found that selecting training data based on speaker features rather than on separate utterance features leads to the creation of overall more intelligible voices. We have 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 future work, we plan to extend our methods to similar corpora in low-resource languages. We have begun experimenting with data from the IARPA BABEL project [22], which consists of read and conversational telephone speech in 25 different low-resource languages, collected for the purpose of developing spoken keyword search systems. Although we will not have the benefit of a high-quality ASR system like Watson to do preliminary voice evaluation in all of these languages, we have nevertheless been experimenting with using ASR systems trained on the Babel data to evaluate intelligibility of trained voices in these languages. While it may seem circular to evaluate TTS systems using an ASR trained on the same data, it may nevertheless tell us which voices best match the bulk of the available spoken data. Furthermore, since we have determined that MCD correlates well with human judgments of intelligibility for TTS voices, we plan to use this as an objective measure of intelligibility for TTS voices for low-resource languages as well. Also, now that we have a number of voices trained on different subsets of the Macrophone corpus, along with human judgments of which voices produced from those subsets are most intelligible, we would like to further explore what characteristics define the best training subsets, in addition to the feature(s) on which they were selected. In addition to exploring the tradeoffs around utterance number and length, we would also like to further examine the tradeoffs around using more data from fewer speakers or less data from more different speakers. Finally, having a sense of what characterizes a "good" TTS training set will enable us to develop more automatic, machine learning based methods of choosing those subsets from large, heterogeneous corpora.

6. Acknowledgements

This work was supported by the National Science Foundation under Grants IIS 1548092 and 1717680.

7. References

[1] E. Cooper, X. Wang, A. Chang, Y. Levitan, and J. Hirschberg, "Utterance selection for optimizing intelligibility of tts voices trained on asr data," *INTERSPEECH*, 2017.

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

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

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

[5] A. Gallardo-Antolín, J. Montero, and S. King, "A comparison of open-source segmentation architectures for dealing with imperfect data from the media in speech synthesis," *INTERSPEECH*, 2014.

[6] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y.-J. Wu, K. Tokuda, R. Karhila, and M. Kurimo, "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, 2010.

[7] P. Baljekar and A. W. Black, "Utterance selection techniques for tts systems using found speech," *9th ISCA Speech Synthesis Workshop*, 2016.

[8] A. Gutkin, L. Ha, M. Jansche, K. Pipatsrisawat, and R. Sproat, "TTS for low resource languages: A Bangla synthesizer," *10th edition of the Language Resources and Evaluation Conference*, 2016.

[9] R. Dall, C. Veaux, J. Yamagishi, and S. King, "Analysis of speaker clustering strategies for hmm-based speech synthesis," *INTERSPEECH*, 2012.

[10] A. Govender and F. de Wet, "Objective measures to improve the selection of training speakers in hmm-based child speech synthesis," *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference*, 2016.

[11] J. Bernstein, K. Taussig, and J. Godfrey, "Macrophone: An American English telephone speech corpus for the POLYPHONE project," *ICASSP*, 1994.

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

[13] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system, system documentation, edition 2.4, for festival version 2.4.0," http://www.festvox.org/docs/manual-2.4.0/festival_toc.html, 2014.

[14] K. Prahallad, A. W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," *ICASSP*, 2006.

[15] P. Boersma, "Praat, a system for doing phonetics by computer," *Clot International, vol. 5, no. 9-10, pp. 341345*, 2001.

[16] Watson developer cloud speech to text. [Online]. Available: https://www.ibm.com/watson/developercloud/speech-to-text.html.

[17] mcd 0.4: Python package index. [Online]. Available: https://pypi.python.org/pypi/mcd

[18] H. Tanaka. aflc/editdistance: Fast implementation of the edit distance(levenshtein distance). [Online]. Available: https://github.com/aflc/editdistance

[19] O. Tange, "Gnu parallel - the command-line power tool," *;login: The USENIX Magazine*, vol. 36, no. 1, pp. 42–47, Feb 2011. [Online]. Available: http://www.gnu.org/s/parallel

[20] Sox - sound exchange. [Online]. Available: http://sox.sourceforge.net/

[21] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on*, vol. 1. IEEE, 1993, pp. 125–128.

[22] M. Harper, "IARPA solicitation IARPA-BAA-11-02," 2011.