Data Selection and Adaptation for Naturalness in HMM-based Speech Synthesis

Erica Cooper, Alison Chang, Yocheved Levitan, Julia Hirschberg

Columbia University

ecooper@cs.columbia.edu, alisonychang6@gmail.com,
yll2109@columbia.edu, julia@cs.columbia.edu

Abstract

We describe experiments in building HMM text-to-speech voices on professional broadcast news data from multiple speakers. We build on earlier work comparing techniques for selecting utterances from the corpus and voice adaptation to produce the most natural-sounding voices. While our ultimate goal is to develop intelligible and natural-sounding synthetic voices in low-resource languages rapidly and without the expense of collecting and annotating data specifically for text-to-speech, we focus on English initially, in order to develop and evaluate our methods. We evaluate our approaches using crowdsourced listening tests for naturalness. We have found that removing utterances that are outliers with respect to hyper-articulation, as well as combining the selection of hypo-articulated utterances and low mean f0 utterances, produce the most natural-sounding voices.

Index Terms: speech synthesis, parametric synthesis, data selection, naturalness, HMM, HTS, BURNC.

1. Introduction

Speech technology has seen a rapid proliferation in recent years, gaining mainstream acceptance by consumers around the world. This is especially the case for mobile Spoken Dialogue Systems (SDS) and virtual personal assistants like Siri for the iPhone, Microsoft Cortana, and Amazon's Alexa. This progress has led to very intelligible and more natural-sounding Text-to-Speech (TTS) synthesis for languages such as English, French, German, Japanese, and Mandarin. These high-resource languages (HRLs) have been studied extensively by speech researchers, who have built and made available such resources as pronunciation rules and dictionaries, created tools such as part-of-speech (POS) taggers and language models, and collected and annotated large corpora of high-quality speech from professional speakers, all in order to create high-quality synthesizers and other related speech technology. However, there are thousands of languages (about 6500) in the world, many of which are spoken by millions of people, which do not have such resources. These low-resource languages (LRLs) like Telugu, Tok Pisin, Vietnamese, and Pashto, for example, have few natural language resources available and no carefully recorded and annotated corpora which can be used for conventional TTS systems. Thus, speakers of these languages do not have the same access to speech-related technologies that allow communication across language barriers, such as SDS or speech-to-speech translation, of which TTS is a crucial component.

In the LRL setting, we do not have access to a large corpus of high-quality annotated data from a single professional speaker, due to the expense of data collection and annotation. Nevertheless, we do often have access to "found" data – data which has been collected for another purpose, such as automatic speech recognition (ASR), or web data. This data is typically recorded in more natural and thus noisy situations, such as broadcast news or telephone speech, and will also differ in speaking style from conventional TTS recordings. Such data will also most likely contain speech from multiple speakers. However, the development of statistical parametric speech synthesis, and in particular Hidden Markov Model (HMM) based speech synthesis [1], has made it possible to train TTS systems on data from multiple speakers and heterogeneous recording conditions and speaking styles. While there has been some prior work on training HMM voices on "found" data, the use of speech from multiple broadcast news speakers or conversational telephone speech has not been extensively studied. In particular, methods for selecting the best utterances and training procedures to optimize for naturalness have not been extensively identified and evaluated.

This paper describes research in producing natural-sounding HMM TTS voices from broadcast news data. While the ultimate goal of this work is to enable the rapid development of intelligible and natural-sounding TTS voices in LRLs without the expense of collecting and annotating a conventional TTS corpus, we first evaluate our methods on an American English corpus in order to identify successful methods and evaluate them quickly.

2. Related Work

The Simple4All project [2] aims to create front-end tools for building voices in LRLs automatically and with minimal human intervention or linguistic knowledge. The only input to the system is the training data; their front-end tools obtain unsupervised word tokenization, syllabification, and phonetic categories derived from the data. The project has seen success at building voices in arbitrary languages for the Blizzard Challenge using data recorded by professional speakers in high-quality studio environments.

In [3], "found" data from political speeches was investigated for adapting average HMM voices. The researchers produced a robust, natural-sounding voice using this method. They also discovered that recording-condition-adaptive training produced more stable synthetic speech. [4] used radio broadcast news to train voices, investigating different speaker diarization and noise detection techniques to remove unsuitable utterances from the training data automatically. Corpora designed for automatic speech recognition (ASR) have also been explored for building HMM-based voices; in particular, [5] built TTS voices on various ASR corpora containing cleanly-recorded

read speech, as well as some speech from a noisy environment, with the goal of being able to create "thousands of voices" from the many speakers in their data. They examined the tradeoffs between amount of data and voice quality, finding that when there is less than an hour of data from a single speaker available, it is better to speaker-adaptively train on data from multiple speakers, whereas if more than two hours of data for a target speaker is available, training a voice on just that individual speaker's data produces a better voice.

Audiobooks have also been a popular source of "found" data for building TTS voices. In particular, [6] used audiobook speech to build unit selection voices. They controlled for recording condition by producing a recording-condition-based clustering and only using utterances from one cluster; they also controlled for variance in speaking style by removing outliers of mean and standard deviation of pitch. Furthermore, they removed sentences with a low alignment score in order to remove both poorly-aligned sentences as well as sentences where the speaker did not accurately read the text. They found that the combination of these approaches produced a better voice. Similarly, [7] built a corpus of 60 hours of speech from audiobooks in 14 different languages, also including only utterances with high automatic alignment confidence scores. They also created a module for selecting utterances with uniform speaking style (given the often very expressive nature of audiobook speech) using a lightly supervised active learning-based approach, for the purpose of building HMM-based voices for these languages. [8] also discarded low-confidence utterances based on ASR confidence rather than alignment; they also removed utterances that were not neutral or suitable for a TTS corpus, as judged by a human. They developed an automatic method for determining utterance naturalness as well, based on discarding utterances outside of manually-chosen thresholds for different acoustic features such as silences, utterance duration, f0, root mean square amplitude, and voicing, as well as text-based features such as punctuation and numbers which might result in front-end text normalization errors. Despite discarding nearly half the original data in both the manual and the automatic approach, 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; the manual approach also did significantly better than the automatic one. These results all show promise for data selection methods on nontraditional TTS training data for producing high-quality voices. Nevertheless, there are many additional features for data selection which have not yet been explored; these will be of particular interest when making use of heterogeneous data sources.

In [9], we trained HMM voices on broadcast news from the Boston University Radio News Corpus (BURNC) [10]. We selected subsets of the male and female utterances based on a number of different factors we hypothesized might be useful for utterance selection, such as speaking rate, f0 and energy mean and standard deviation, and level of articulation. Many of our chosen features were guided by our prior work on the acoustic features that correlate with charisma in American English, as well as in other languages such as Arabic and Swedish [11] [12] [13] [14], which found that in American English, louder utterances, utterances higher in the speaker's pitch range, and utterances with a faster speaking rate were rated by listeners as more charismatic. Furthermore, high mean pitch and high standard deviation of root mean square intensity correlated with charisma cross-culturally. Since these features are informative of charismatic speech, they may also play a role in percieved naturalness of synthesized speech. We were able to identify approaches that performed consistently poorly – choosing hyperarticulated and slow speaking rate utterances were some of the worst approaches for both male and female data. Data selection with the male data proved to be less successful than with female data, suggesting that the male baseline, using all of the male data, is already quite natural-sounding due to the similarity of the four male speakers to each other, and thus there is less room for improvement. We also found that, although speaker-adaptive training (SAT) can factor out speaker differences to produce more consistent models, our SAT-trained voices were not rated as significantly more natural-sounding than speaker-independently trained (SI) voices. While we found no data selection approach that consistently gave a significant improvement, we did identify some methods that showed promising tendencies, such as selecting low mean f0 utterances, which guide our further exploration in this paper.

3. Corpus and Tools

We use the English Boston University Radio News Corpus (BURNC), collected by Mari Ostendorf, Patti Price, and Stefanie Shattuck-Hufnagel and distributed by the Linguistic Data Consortium (LDC96S36) [10]. This corpus consists of professionally read radio news from four male and three female FM radio news announcers associated with the public radio station WBUR. The main corpus consists of news recorded in the station's studio during broadcasts over a two year period. In addition, the same announcers were recorded in a laboratory at Boston University in both non-radio and radio speaking styles. We used the broadcast radio news portion of the corpus for our experiments, which consists of 5 hours and 15 minutes of speech from male speakers, and 4 hours and 22 minutes of speech from female speakers. The original corpus was digitized at 16 kHz, orthographically transcribed and (partly) prosodically annotated manually, using the ToBI conventions [15]; it was phonetically aligned and part-of-speech tagged automatically and hand corrected. To date, we have not made use of any annotations except the orthographic transcripts. We trained only all-male or all-female voices to produce more consistent models.

For the baselines and our selected subsets, utterances were defined as sentences in the transcript text, and both the text and audio were segmented accordingly. We trained our TTS voices using the Hidden Markov Model Based Speech Synthesis System (HTS) [16]. For our training recipes, we used the speaker-independent and speaker-adaptive demo recipes for HTS version 2.2. We obtained the standard set of full-context phonetic labels for each utterance of the BURNC data using the Festival Speech Synthesis System front-end [17]. Synthesis and vocoding from trained models were done using hts-engine.

4. Experiments and Results

4.1. Features

In our prior work [9], we produced subsets of utterances based on a number of different criteria. Our features included mean and standard deviation of energy and fundamental frequency (f0), computed using the Praat software [18]; speaking rate, defined by syllables per second; and utterance length based on the duration of the audio. We also hypothesized that hypo- and hyper-articulation of training utterances might have an effect on the naturalness of the resultant voice. [19] notes that slow speaking rate and louder speech are associated with hyper-

articulation. We therefore computed articulation as mean energy divided by speaking rate, so that a high articulation value would be associated with high mean energy and slow speaking rate. We selected voice training subsets of one hour for high, middle (median), and low values for each of these features as follows: We sorted all of the utterances by feature value, and then took the top, middle, or bottom hour of data from that sorted list to create our subsets. We compared our test voices to baselines that used all of the data for each gender. In this work we continue to explore these same features, looking beyond 1-hour subsets, towards different amounts of data, trimming outliers, and combinations of the best features.

4.2. Evaluation

To evaluate naturalness, we published listening tests online, using Amazon Mechanical Turk (AMT), a popular crowdsourcing platform. To restrict our task to native speakers of English, we required workers to complete a qualification test first, in which workers chose the languages they spoke since birth from a list of options. We only allowed workers who selected English and no more than two other languages to participate in our evaluation, in order to exclude those who might select, e.g., all of the languages, in an attempt to cheat the system. We also restricted our tasks' visibility to workers within the United States.

Our task consisted of a pairwise comparison between the baseline and a test voice. Voices trained on subsets of the male utterances were always paired with the male baseline, and voices trained on female utterances were paired with the female baseline. Each task thus contained only two audio files, the same sentence spoken by the baseline voice and one of our test voices. Workers could rate as many or as few pairs of utterances as they wished. Half of the sentences were presented in A/B order and the other half in B/A order, to avoid possible order effects. We ensured that raters played both audio files entirely before they were allowed to submit their preference. Raters were given a forced choice, i.e. there was not a "no preference" option. We chose 12 lexically neutral sentences of varying length from the fable "Jack and the Beanstalk" and synthesized them with each of our voices. Each task was completed by 5 different workers, for a total of 60 comparison ratings for each voice.

4.3. Varying Subset Sizes

In our prior work [9], we trained voices on a constant subset size of one hour. In the current work, we wanted to explore additional subset sizes. We took our two most promising features, hypo-articulated and low mean f0 female utterances, and created 30-minute and 2-hour subsets of the utterances to complement our 1-hour subsets. We trained a voice on each subset, using the HTS 2.2 speaker-independent demo scripts as in our prior work, since we have also previously observed that, while speaker-adaptive training takes substantially more time and computational resources, the average voice model it produces is not ultimately rated as more natural-sounding. Results for pairwise naturalness comparisons are presented in Table 1.

We see increasing preference for larger data sets, indicating that, in this case, more data is better. Furthermore, none of these subsets produced voices rated to be significantly better than the baseline.

4.4. Combination of Best Approaches

We next hypothesized that combining our best approaches might produce a better voice. We tried a number of different combinations of hypo-articulation and low mean f0: 1) We took a 2-hour subset of the most hypo-articulated utterances and intersected it with a 2-hour subset of the lowest mean f0 utterances to produce a 54-minute training set of female data; 2) we combined 30-minute subsets of each by set union into a 56-minute subset (not a full hour because some utterances appear in both sets); 3) we combined 1-hour subsets of each into a 1 hour and 46 minute subset; 4) we multiplied the mean f0 values for every female utterance by their articulation values, and selected a 1-hour subset of the utterances with the lowest resulting values; 5) same as (4) except we selected 2 hours; 6) same except 3 hours; 7) same except 4 hours. Results are shown in Table 2. We see that subsets (3) and (4) perform well, with (5) performing significantly better than the baseline ($p \leq 0.05$). This motivated us to try (6) and (7), using the same filtering method but with increasingly more data, however in this case more data did not help.

	Hypo-articulation		Low Mean F0	
Amount	Preferred	P-value	Preferred	P-value
30min	31.7%	4.51e-3	36.7%	0.04
1hr	43.3%	0.30	53.3%	0.61
2hr	58.3%	0.20	56.7%	0.30

Table 1: *Pairwise comparison preferences for female voices trained on subset sizes of 30 minutes, 1 hour, and 2 hours over the baseline.*

Combination	Preferred	P-value
1. Intersection (54min)	53.3%	0.61
2. Union (56min)	58.3%	0.20
3. Union (1hr46min)	61.7%	0.07
4. Multiplication (1hr)	61.7%	0.07
5. Multiplication (2hr)	68.3%	0.005
6. Multiplication (3hr)	51.7%	0.80
7. Multiplication (4hr)	45.0%	0.44

Table 2: *Pairwise comparison preferences for female voices trained on different subsets of combinations of hypo-articulation and low mean F0.*

4.5. Subset Adaptation

Our results so far indicate that limiting the size of the training data is not always beneficial. With this in mind, we trained voices using all of the data for one gender, adapting to each of our subsets, using the HTS speaker-adaptive training recipe. By labeling the selected subset as one "speaker" and the rest of the data as another "speaker," we hope to obtain the benefits both of using all the data and also of identifying subsets of utterances that may improve voice naturalness. The adapted model does not correspond to any specific speaker, but rather to the feature for which we are selecting. We trained adapted voices for each of the 1-hour subsets we used in our prior work for the female data. Results are shown in Table 3.

Our two best voices were those adapted to 1-hour subsets of hypo-articulated utterances and middle mean energy utterances, with hypo-articulation approaching significance. As with our speaker-independently trained subsets, we examined additional variations of 30-minute and 2-hour subsets for subset adaptation based on these two features. We also tried combining these two best approaches by intersecting 2-hour sets of each to produce a 59-minute subset. Results are presented in Table 4.

Adaptation Subset	Preferred	P-value
High mean energy	35.0%	0.02
High mean F0	40.0%	0.12
Hyper-articulation	41.7%	0.20
Low mean energy	43.3%	0.30
Middle mean F0	43.3%	0.30
Slow speaking rate	45.0%	0.44
High std.dev energy	46.7%	0.61
Middle std.dev F0	46.7%	0.61
Short duration	48.3%	0.80
High std.dev F0	48.3%	0.80
Fast speaking rate	50.0%	1.0
Low mean F0	50.0%	1.0
Medium duration	50.0%	1.0
Low std.dev energy	50.0%	1.0
Long duration	51.7%	0.80
Middle std.dev energy	53.3%	0.61
Medium speaking rate	56.7%	0.30
Low std.dev F0	56.7%	0.30
Middle mean energy	60.0%	0.12
Hypo-articulation	61.7%	0.07

Table 3: *Pairwise preferences for female voices adapted to 1-hour subsets.*

Adaptation Subset	Preferred	P-value
Hypo-articulation - 30min	56.7%	0.30
Hypo-articulation - 2hr	48.3%	0.80
Middle mean energy - 30min	50.0%	1.0
Middle mean energy - 2hr	53.3%	0.61
Intersection - 59min	48.3%	0.80

Table 4: *Pairwise preferences for female voices adapted to 30-minute, 2-hour, and intersected sets of hypo-articulated and middle mean energy utterances.*

Adapting to subsets of 30 minutes and 2 hours does worse than adapting to 1-hour subsets, for both hypo-articulation and middle mean energy. Intersecting 2-hour sets of these best approaches to get an approximately 1-hour set for adaptation did not turn out to be useful. Adaptation to other types of "combination" subsets should be investigated in future work as well.

4.6. Removal of Outliers

Returning to our finding that removing too much data may be detrimental to naturalness, given that this is fairly high-quality data to begin with, we trained voices by removing a smaller portion of the data – outlier utterances based on the features that produced the worst voices in our prior work. Since we have seen in the past that utterances with speaking rates at the extremes and hyper-articulated utterances produced some of the worst voices, we created sets where we trimmed the upper and lower tail of the female utterances when sorted by speaking rate (two separate sets), and the upper tail of hyper-articulation. For high speaking rate, we found that the mean was 4.66 syllables per second and the standard deviation was 0.75, so we chose a cutoff of mean plus one standard deviation (5.40), producing a set of 3 hours and 52 minutes. For removing low speaking rate utterance outliers, we did a similar cutoff of mean minus one standard deviation, producing a subset of 4 hours and 6 minutes. For hyper-articulation, we found a mean of 13.91 and a standard deviation of 2.67; we again chose a cutoff of mean plus 1 standard deviation (16.57). This gave us 4 hours and 6 minutes of data. Pairwise comparison results for these two voices are in Table 5. We have thus obtained a significantly more preferred voice ($p \leq 0.05$) by removing the outlying hyper-articulated utterances from the training set. This is a promising direction for future exploration.

Outlier feature	Preferred	P-value
High speaking rate	56.7%	0.30
Low speaking rate	51.7%	0.80
Hyper-articulation	65.0%	0.02

Table 5: *Pairwise comparison preferences for female voices trained on data sets with outliers removed.*

5. Conclusions and Future Work

Level of articulation has shown to be a consistently useful feature for all of our approaches, especially in the case of removing hyper-articulated outlier utterances and when combining hypo-articulation with low mean f0, both of which produce female voices that were rated as sounding significantly more natural by Mechanical Turk workers than the baseline. Training on all of the data and adapting to subsets has shown some promising tendencies as well when adapting towards hypo-articulated utterances, which requires further exploration. Our future work will also continue to investigate the outlier-removal approach with more different features and combinations of features.

We would also like to explore machine learning approaches for identifying the best utterances to use for voice training, based on a small sample of labeled utterances, using the same kinds of features we are investigating here, as well as perhaps lower-level features such as frame-level acoustic features, and higher-level features such as speaker characteristics. We will also examine whether our approaches generalize to other types of data, such as conversational telephone speech, and to actual low-resource languages. We would also like to find methods that generalize to male speech, with which we have seen less success with our current approaches. We also plan to explore the use of other kinds of "found" data as well, such as audiobooks and video lectures, and to examine the feasibility of finding and collecting such data in different languages from the web.

6. Acknowledgements

This work was supported by NSF 1539087 "EAGER: Creating Speech Synthesizers for Low Resource Languages" and by Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

We also thank Luise Valentin Rygaard for her early work on this project, Heiga Zen for his helpful suggestions regarding adaptation to subsets, and Meredith Brown for her advice on setting up experiments and interpreting results with Mechanical Turk.

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] O. Watts, S. Ronanki, Z. Wu, T. Raitio, and A. Suni, "The NST-GlottHMM entry to the Blizzard Challenge 2015," *Proc. Blizzard Challenge Workshop*, 2015.

[3] J. Yamagishi, Z. Lin, and S. King, "Robustness of HMM-based speech synthesis," *INTERSPEECH*, 2008.

[4] 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*, 2004.

[5] 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.

[6] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, and S. Raptis, "Using audio books for training a text-to-speech system," *Proceedings of the 9th International Conference on Language Resources and Evaluation*, 2014.

[7] 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.

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

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

[10] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Tech. Rep.*, 1995.

[11] A. Rosenberg and J. Hirschberg, "Acoustic/prosodic and lexical correlates of charismatic speech," *Eurospeech*, 2005.

[12] F. Biadsy, J. Hirschberg, A. Rosenberg, and W. Dakka, "Comparing American and Palestinian perceptions of charisma using acoustic-prosodic and lexical analysis," *INTERSPEECH*, 2007.

[13] F. Biadsy, A. Rosenberg, R. Carlson, J. Hirschberg, and E. Strangert, "A cross-cultural comparison of American, Palestinian, and Swedish perception of charismatic speech," *Speech Prosody*, 2008.

[14] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Speech Communication*, 2008.

[15] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," *Proc. of the 1992 International Conference on Spoken Language Processing, vol. 2, 1992, pp. 12-16*, 1992.

[16] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," *6th ISCA Workshop on Speech Synthesis*, 2007.

[17] A. W. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system." [Online]. Available: http://www.festvox.org/festival/

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

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