Data Selection for Naturalness in HMM-based Speech Synthesis

Erica Cooper, Yocheved Levitan, Julia Hirschberg

Columbia University

ecooper@cs.columbia.edu, yll2109@columbia.edu, julia@cs.columbia.edu

Abstract

We describe experiments in training HMM text-to-speech voices on professional broadcast news data from multiple speakers. We compare data selection techniques designed to identify the best utterances for voice training in a corpus not explicitly recorded for synthesis, aiming to select utterances from the corpus which will produce the most natural-sounding voices. We also explore different methods for voice training and utterance synthesis that can improve naturalness. While the ultimate goal of this work is to develop intelligible and natural-sounding synthetic voices in Low Resource Languages rapidly, without the expense of collecting and annotating professional data specifically for text-to-speech, we focus on English first, in order to develop our methods. We also describe results of crowdsourced listening tests which identify the strengths and weakness of different data selection and voice training methods when rated by listeners in terms of naturalness.

Index Terms: speech synthesis, data selection, naturalness

1. Introduction

The rapid improvement of speech technology in recent years has resulted in its widespread adoption by consumers, especially in mobile applications such as Spoken Dialogue Systems (SDS) like Siri for the iPhone and Voice Search on Android phones. This progress has led to very intelligible and more natural-sounding Text-to-Speech (TTS) synthesis for languages such as English, French, German, Cantonese, Mandarin, Japanese, Italian, Spanish, and Korean. These High Resource Languages (HRLs) have been studied extensively by speech researchers, who have built pronunciation rules and dictionaries, part-of-speech (POS) taggers, and language models, and have collected and annotated large amounts of high-quality data from professional speakers in order to provide high-quality synthesis. 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. Low Resource Languages (LRLs) like Telugu, Tok Pisin, Tamil, Vietnamese, Tagalog, Cebuano, and Pashto, for example, have few natural language processing and data resources available to TTS researchers 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.

While, in the LRL setting, we will 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, we do often have access to "found" data, created for other purposes such as Automatic Speech Recognition which requires data collected in more natural and thus noisy situations, such as news broadcasts or telephone speech. The speaking style of such data will differ from that of data recorded specifically for TTS, and the presence of multiple speakers is also unconventional for a TTS corpus. However, the development of Hidden Markov Model (HMM) speech synthesis [1] has made it possible to train TTS systems on non-traditional data from multiple speakers and heterogeneous recording conditions. While there has been some prior work on training HMM voices on "found" data, the use of speech from multiple broadcast news speakers has not been extensively studied. In particular, methods for selecting the best utterances and training procedures to optimize for naturalness have not been identified and evaluated.

This paper describes research investigating the training of natural-sounding HMM TTS voices on broadcast news data. While the ultimate goal of this work is to facilitate the rapid development of intelligible and natural-sounding TTS voices in LRLs without the expense of collecting and annotating data specifically for TTS, we first evaluate our methods on an English corpus in order to identify successful methods and evaluate them quickly. In Section 2 we describe related work on synthesis using non-traditional corpora. In Section 3 we describe the corpus we have used to develop our methods, and in Section 4 we describe the selection methods and training procedures we have used to create voices for subjective evaluation. We describe results of our evaluations in Section 5.

2. Related Work

In [2], "found" data from political speeches was investigated for voice adaptation, after training an average HMM voice model (AVM) on many source speakers. The researchers were able to obtain a robust, natural-sounding voice using this method, with performance minimally degraded by the inclusion of the found data. They also discovered that, by using recording-condition-adaptive training, they could produce more stable synthetic speech. [3] used radio broadcast news recordings to train synthetic voices, investigating different speaker diarization and noise detection techniques to remove unsuitable utterances automatically. Corpora designed for Automatic Speech Recognition (ASR) have also been explored for building HMM-based TTS voices; in particular, [4] 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 in the case where less than an hour of data from a single speaker is available, it is better to train on data from multiple speakers, whereas if more than two hours of data for that speaker is available, training a voice for that individual speaker produces a better voice.

Audiobooks have also been a popular source of "found" data for building TTS voices. In particular, [5] used a corpus of

audiobook speech to build unit selection voices. To handle the different recording conditions of the various audiobooks, they first did a recording-condition-based clustering, and kept only utterances from one cluster. Since audiobook speech contains a great deal of expressivity, such as emotion and character voices, they selected the most neutral utterances by plotting the mean and standard deviation of pitch, and keeping only the 90% of the data closest to the centroid. Furthermore, since the 140 hours of speech in their corpus had to be aligned with the text automatically, 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 read the text exactly as it was written. They found that the combination of these approaches did produce a better voice. Similarly, [6] built a corpus of 60 hours of speech from audiobooks in 14 languages, one speaker per language, also including only utterances with high automatic alignment confidence scores. They also created a module for selecting utterances with uniform speaking style using a lightly supervised active learning-based approach, specifically for the purpose of building HMM-based voices in different languages. Finally, [7] also discarded low-confidence utterances, but based on ASR confidence rather than alignment confidence, and discarded utterances that were not neutral or suitable for a TTS corpus, as judged by a human. They also developed an automatic method for deciding utterance naturalness, based on discarding utterances outside of manually-chosen thresholds for 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 text normalization errors. Despite discarding nearly half the original data with each of these approaches, they found that the HMM voices they trained using both of these methods were judged as significantly better than using all of the data in a preference test, and 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. Although these methods were developed primarily for audiobooks and for data from a single speaker, some of them may also prove to be applicable for building HMM voices from other types of found data from multiple speakers.

3. Corpus

For our data selection experiments 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) [8]. This corpus consists of professionally read radio news data and includes speech from seven (four male, three female) FM radio news announcers associated with the public radio station WBUR. The main corpus consists of over seven hours of news stories 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, where they read 24 stories from the radio news portion, first in a normal, non-radio style and then, 30 minutes later, in their radio style. We used the broadcast radio news part of the corpus for our experiments. The original corpus was digitized at 16 kHz, orthographically transcribed and (partly) prosodically annotated manually, using the ToBI conventions [9]; they were phonetically aligned and part-of-speech tagged automatically and hand corrected. To date we have not made use of annotations except the orthographic transcripts. Initially, we looked only at data from the three female speakers; we subsequently repeated our experiments on speech from the four male speakers. We trained only all-male or all-female voices in order to produce more consistent models.

4. Data Selection and Training Approaches

To provide material for our experiments, we created subsets of the BURNC corpus by selecting utterances based on a number of different criteria. We compared these subsets to two baselines, one (for the female voices) trained on all of the female data (4h 40m) and one (for the male voices) trained on all of the male speech (5h 15m). For the baselines and our selected subsets, utterances were defined as sentences in the transcript text, and the audio was segmented accordingly.

We selected subsets of the male and female utterances based on a number of different criteria, based upon previous work in our lab and other factors we hypothesized might be useful for utterance selection. We examined mean and standard deviation of energy and fundamental frequency (f0), selecting subsets of utterances totalling one hour in duration for the highest, lowest, and middle of each of these values, computed using the Praat speech analysis software [10]. We did the same for speaking rate, defined by syllables per second. We also considered that hypo- and hyper-articulation of training utterances might have an effect on the naturalness of the resultant voice, so we selected subsets of low- and high-articulation utterances, computed by mean energy divided by speaking rate, also each one hour. We also hypothesized that there may be some optimal utterance length for TTS training data, so we selected subsets of the longest, shortest, and median utterances based on the length of the audio file. The time ranges of these subsets can be found in Table 1.

Female		Male	
Subset	Range (s)	Subset	Range (s)
Short	0.25 - 5.50	Short	0.71 - 5.24
Middle	5.25 - 6.98	Middle	6.40 - 7.22
Long	10.05 - 22.19	Long	11.22 - 25.39

Table 1: *Range of utterance durations for hour-long subsets of short, middle, and long utterances.*

Many of our selection features were guided by our prior work on the prosodic characteristics that correlate with charisma in American English, as well as other languages such as Arabic and Swedish [11] [12] [13] [14]. These studies investigated pitch, intensity, speaking rate, and durational features, finding that in American English, louder utterances and utterances higher in the speaker's pitch range were rated by listeners as more charismatic, as well as those with a faster speaking rate. Furthermore, a high mean pitch and high standard deviation of rms intensity correlated with charisma cross-culturally. We believe that since these features are informative of charismatic speech, they may also play a role in percieved naturalness of synthesized speech.

4.1. Voice Training Methods

We trained our TTS voices using the Hidden Markov Model Based Speech Synthesis System (HTS) [15]. We based our training recipe for the baselines and for the data selection subsets described above on the speaker-independent training demo recipe. We treat all of the data in each subset as if it were

from one speaker, primarily for computational efficiency, as we wanted to be able to create many voices as rapidly as possible. We obtained full-context phonetic labels for the BURNC data using the Festival Speech Synthesis System [16]. Since speaker adaptive training (SAT) is known to produce more stable and better-sounding voices [17], we also speaker-adaptively trained one voice using all of the male data, and one voice using all of the female data, based on the HTS SAT demo. We synthesized our test utterances from the resulting average voice model (AVM) in order to determine whether this more computationally-intensive training approach actually produces more natural-sounding voices from our data.

It was shown in [4] that training a speaker-dependent voice when at least 2 hours of target speaker data is available produces a better voice than training a voice on mixed data, and that if one has less than an hour of target-speaker data, one is better off using SAT with data from multiple speakers. Indeed, we would expect speaker-dependent voices to be more stable than one trained on multiple speakers when sufficient data is available. In an effort to determine the extent to which these findings generalize to the type of data we are using, we also trained speaker-dependent voices for each of the seven speakers in the BURNC corpus; Table 2 shows the amount of data available for each of the speakers.

Female		Male	
Speaker	Amount	Speaker	Amount
f1a	01:06:00.98	m1b	00:54:59.87
f2b	00:55:36.97	m2b	01:09:34.02
f3a	02:20:20.92	m3b	00:49:00.22
		m4b	02:21:45.75

Table 2: *Amount of data per speaker in the Boston University Radio News Corpus.*

Finally, we noticed that many of the voices we produced had a very choppy-sounding f0 contour, despite our data selection. We hypothesized that a simple way to remedy this would be to set the f0 values to a constant in the training data, train a monotone voice with that training data, and then interpolate the monotone voice with the baseline voice in equal proportions. We therefore created one male and one female monotone voice, as well as one male and one female interpolated voice.

5. Experiments and Results

5.1. Evaluation

To evaluate the naturalness of the voices resulting from the selection methods described in Section 4, we published crowdsourced listening tests online, using Amazon Mechanical Turk (AMT), a popular crowdsourcing platform. To restrict our listeners to native English speakers, we required workers to complete a qualification test before completing any of our tasks, in which we asked which languages they have spoken since birth, from a list of languages. We only allowed Turkers who selected English as one of these language and no more than two other languages, in order to exclude those who might select, e.g., all of the languages. We also restricted our tasks' visibility to workers within the United States.

5.2. MOS Test - Female Voices

Our first experiment was modeled after the Mean Opinion Score (MOS) AMT listening test outlined in [18] in which Human Intelligence Tasks (HIT) asked Turkers to evaluate the naturalness of 12 spoken utterances by selecting from a 5-point Likert scale, from 1 = very unnatural, 2 = somewhat unnatural, 3 = neither natural nor unnatural, 4 = somewhat natural, and 5 = very natural. We chose lexically neutral sentences of varying length from the fable "Jack and the Beanstalk" and synthesized them with our baseline and subset-trained voices. Voices were given coded names so that participants would not be able to infer how they were created. We included 23 voices in this original AMT experiment - 20 synthesized subset voices, 1 human voice (resynthesized), 1 robotic voice (generated using Mac OSX's 'say' command, Zarvox speaker), and 1 baseline voice (trained on the entire female speaker dataset). The human and robotic voices were used as references to determine whether to accept the submitted work. If they received incorrect ratings (anything but 4 or 5 for the human voice, or 1 or 2 for the robotic voice) the submission was not used. We randomized the order of the test audio files but ensured that the reference voices were always in the last half of the playlist to avoid skewing listeners' opinions. In addition, Turkers were not allowed to rate a voice until they had listened to the entire audio file being rated. Each HIT contained 23 audio files, each created from the same sentence, but spoken by a different voice. Every voice was rated by 5 unique individuals. Table 3 presents MOS scores of naturalness ratings resulting from this experiment.

Voice	Rating	Voice	Rating
Robotic	1.03	Low mean energy	2.41
High mean f0	1.97	Mid mean energy	2.41
Hyper-articulated	2.08	Longest utts	2.5
High mean energy	2.08	Fast rate	2.55
Mid length utts	2.08	Mid mean f0	2.55
Slow rate	2.13	Mid sdev f0	2.6
High sdev energy	2.13	Low sdev f0	2.6
Mid sdev energy	2.28	Baseline	2.68
Shortest utts	2.33	Hypo-articulated	2.7
High sdev f0	2.37	Low mean f0	2.7
Low sdev energy	2.37	Natural speech	4.95
Mid rate	2.4		

Table 3: *Average naturalness rating for each voice (low to high), MOS experiment*

Although none of our test voices did significantly better than our baseline, we did observe some trends. Voices trained on utterances selected for hypo-articulation and for low mean f0 did slightly but not significantly better than the baseline.

5.3. Pairwise Comparisons - Female Voices

Since the MOS test required that each rater listen to 23 different voices, and since we failed to find substantial differences between the ratings of the test voices, we suspected that a MOS test in which raters were asked to compare 23 voices might have been overwhelming, in effect precluding meaningful comparisons. We therefore designed and posted a second task, a pairwise comparison between the baseline voice and each test voice. Each HIT thus contained only two audio files, the same sentence spoken by the baseline voice and one of our synthesized 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 listened to both audio files entirely before they were allowed to submit their prefer-

ence. Raters were given a forced choice, i.e. there was not a "no preference" option. We used the same 12 sentences as in our MOS test. However, we only asked Turkers to rate the two voices that had scored better than the baseline in the MOS test - the hypo-articulated and low mean pitch voices - as well as the two voices that had scored just below the baseline, low standard deviation for pitch, and middle standard deviation for pitch. We also posted HITs for the AVM voice, the monotone voice, the interpolated voice, and middle-range f0 standard deviation, as these voices had been produced after our MOS test. We compared these to the baseline as well. Table 4 presents the results:

Female Voice	Preferred	P-value
Monotone	1.7%	6.99e-14
f2b	35.0%	0.02
Mid sdev f0	38.3%	0.07
f1a	40.0%	0.12
f3a	41.7%	0.20
Hypo-articulated	43.3%	0.30
Low sdev f0	50%	1
Low mean f0	53.3%	0.61
SAT-trained AVM	56.7%	0.30
Interpolated	63.3%	0.04

Table 4: *Percent of votes for the test voice over the female baseline (out of 60 ratings), from low to high, and p-value.*

The MOS task showed that the voice trained on low mean f0 utterances was rated slightly but not significantly better than the baseline, and we see this tendency in the pairwise comparison test as well. The SAT-trained AVM voice in this task was also preferred slightly but not significantly to the baseline. Voices trained on individual speakers tended to do worse than the baseline (using all of the data). The voice that was created by interpolating the baseline with the monotone voice was the only one which was significantly ($p<0.05$) preferred to the baseline.

5.4. Pairwise Comparisons - Male Voices

Having obtained these ratings for voices trained on different subsets of the female BURNC data, we examined which, if any, of the selection methods would generalize to the male data. We trained voices using each of the data subset selection methods and training approaches on the male data as we had done for the female data. We decided to restrict this and future ratings to pairwise comparisons rather than the MOS approach since it is a simpler and clearer task. Results from rating of male voices are shown in Table 5.

We see that the voice trained on male utterances with a low mean f0 was one of the worst-rated voices, in contrast to the female low mean f0 voice, which was slightly preferred to the baseline. Voices trained on individual speakers also tended to do worse than the baseline, while there was approximately no preference for the SAT-trained AVM male voice. The voice created by interpolating the baseline voice with the monotone one was the only voice to get more votes than the baseline for its comparison, however this was not a significant preference.

6. Conclusions and Future Work

While none of our methods worked significantly better than the baseline for both male and female data, the fact that the voices that were created by interpolating the baselines with monotone voices did best indicates that more direct modeling of prosody might be the best future approach to improve naturalness and reduce the "choppiness" of the voice. Since the BURNC corpus contains prosodic annotations, we could include these in the TTS full-context labels, and also compare this to an automatic approach for obtaining prosodic annotations, such as AuToBI [19], which would generalize to low-resource languages where such hand annotations are not readily available. Our experiments with different training approaches revealed that we do not have enough single-speaker data to produce a voice that is more natural than one trained on multiple speakers. Furthermore, we learned that, in our case, creating a SAT-trained AVM is not worthwhile, since this did not produce a substantially better voice. Finally, we learned which methods do consistently badly – choosing hyper-articulated and slow utterances were some of the worst approaches for both male and female data. Perhaps instead of selecting subsets that fit a certain criteria, we can do the opposite – simply filter out the utterances which are likely to hurt the naturalness of the voice. We would also like to try approaches such as removal of outliers and use of alignment confidence score; these have been popular and effective in producing voices from audiobook corpora. Finally, combinations of our data selection approaches may also yield better voices.

Male Voice	Preferred	P-value
Monotone	1.7%	6.99e-14
Low mean f0	16.7%	2.42e-7
Mid mean energy	20%	3.36e-6
Mid mean f0	20%	3.36e-6
Slow rate	21.7%	1.14e-5
Hyper-articulated	23.3%	3.61e-5
Low sdev f0	25%	1.08e-4
Fast rate	25%	1.08e-4
Medium length utts	25%	1.08e-4
High stdv energy	28.3%	7.89e-4
Mid stdv energy	28.3%	7.89e-4
High sdev f0	28.3%	7.89e-4
Mid articulated	28.3%	7.89e-4
Mid rate	28.3%	7.89e-4
High mean energy	30%	1.95e-3
High mean f0	30%	1.95e-3
Middle sdev f0	30%	1.95e-3
m1b	30%	1.95e-3
m4b	30%	1.95e-3
Shortest utts	31.7%	4.51e-3
Low mean energy	35%	0.02
Hypo-articulated	35%	0.02
m3b	38.3%	0.07
Longest utts	38.3%	0.07
Low stdv energy	41.7%	0.20
m2b	46.7%	0.61
SAT-trained AVM	48.3%	0.80
Interpolated	55%	0.44

Table 5: *Percent of votes for the test voice over the male baseline (out of 60 ratings; low to high), and p-value.*

7. Acknowledgements

This work was supported by NSF 1539087 "EAGER: Creating Speech Synthesizers for Low Resource Languages." We also thank Meredith Brown for her helpful advice regarding setting up experiments and interpreting results on Mechanical Turk.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," *Tech. Rep.*, 1995.

[9] 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.

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

[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] 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.

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

[17] J. Yamagishi, "Average-voice-based speech synthesis," *PhD Thesis, Tokyo Institute of Technology*, 2006.

[18] K. Georgila, A. W. Black, K. Sagae, and D. Traum, "Practical evaluation of human and synthesized speech for virtual human dialogue systems," *LREC*, 2012.

[19] A. Rosenberg, "Autobi - a tool for automatic tobi annotation," *INTERSPEECH*, 2010.