Characteristics of Text-to-Speech and Other Corpora

Erica Cooper, Emily Li, Julia Hirschberg

Columbia University, USA

ecooper@cs.columbia.edu, el2857@columbia.edu, julia@cs.columbia.edu

Abstract

Extensive TTS corpora exist for commercial systems created for high-resource languages such as Mandarin, English, and Japanese. Speakers recorded for these corpora are typically instructed to maintain constant f0, energy, and speaking rate and are recorded in ideal acoustic environments, producing clean, consistent audio. We have been developing TTS systems from "found" data collected for other purposes (e.g. training ASR systems) or available on the web (e.g. news broadcasts, audiobooks) to produce TTS systems for low-resource languages (LRLs) which do not currently have expensive, commercial systems. This study investigates whether traditional TTS speakers do exhibit significantly less variation and better speaking characteristics than speakers in found genres. By examining characteristics of f0, energy, speaking rate, articulation, NHR, jitter, and shimmer in found genres and comparing these to traditional TTS corpora, We have found that TTS recordings are indeed characterized by low mean pitch, standard deviation of energy, speaking rate, and level of articulation, and low mean and standard deviations of shimmer and NHR; in a number of respects these are quite similar to some found genres. By identifying similarities and differences, we are able to identify objective methods for selecting found data to build TTS systems for LRLs.

Index Terms: speech synthesis, found data, corpus statistics

1. Introduction

In recent years, text-to-speech synthesis (TTS) has become widespread in the form of mainstream consumer products such as mobile virtual personal assistants (Siri, Google Assistant), in-home devices (Amazon Echo), and other applications such as speech-to-speech translation. However, collecting the type of data required to build a high-quality TTS voice is typically very costly, and thus only undertaken with a major economic motivation. Typically, a professional voice talent reads dozens of hours of text with good coverage of the target domain in a soundproof room with a high-quality microphone and in as neutral and even a style as possible. They are typically instructed to maintain constant f0, energy, speaking rate, and articulation throughout. Without the resources to collect such data, is it still possible to create a high-quality voice? With the advent of statistical parametric speech synthesis (SPSS) such as Hidden Markov Model (HMM) based synthesis and neural network based synthesis, it is possible to create voices without necessarily having to collect large amounts of high-quality, single-speaker, in-domain speech. Furthermore, large amounts of available speech such as audiobooks and radio broadcast news present a promising source of data for building new voices. In this paper, we examine a number of corpora in different genres and collected for different purposes in order to compare their similarities and differences with respect to various acoustic and prosodic features. We aim to determine whether TTS corpora do in fact follow the "standard" TTS speaking style, whether other forms of professional and non-professional speech differ substantially from the TTS style, and which features are most salient in differentiating the speech genres.

2. Related Work

TTS speakers are typically instructed to speak as consistently as possible, without varying their voice quality, speaking style, pitch, volume, or tempo significantly [1]. This is different from other forms of professional speech in that even with the relatively neutral content of broadcast news, anchors will still have some variance in their speech. Audiobooks present an even greater challenge, with a more expressive reading style and different character voices. Nevertheless, [2, 3, 4] have had success in building voices from audiobook data by identifying and using the most neutral and highest-quality utterances. Furthermore, in our own prior work [5, 6, 7], we have created more natural-sounding voices out of radio broadcast news speech and data collected for automatic speech recognition (ASR) by selecting training utterances based on acoustic and prosodic criteria motivated by knowledge of what makes a "good" TTS voice. In the current work we will validate these assumptions about TTS voices empirically and identify similarities and differences when we compare them to other genres, for the purpose not only of identifying genres which may be most suitable for building TTS voices, but utterances within those genres which should be selected or discarded in the process.

3. Corpora

We examine statistical similarities and differences in various acoustic and prosodic features in a number of different corpora. Such corpora include TTS recordings, audiobook speech, radio broadcast news, and telephone conversations recorded to train ASR systems in a variety of languages.

3.1. TTS Corpora

The CMU ARCTIC databases [8] were collected in studio conditions for unit selection synthesis research and consist of approximately one hour per speaker of phonetically-balanced sentences collected from out-of-copyright texts. Currently, the database consists of two male and two female US English speakers, as well as Canadian, Scottish, and Indian English male speakers.

The SWARA corpus [9] contains studio-quality recordings from 17 volunteer Romanian speakers (9 female, 8 male) reading isolated sentences from newspaper articles. 880 utterances were common to all speakers.

The IIIT-H Indic databases [10] were collected for speech synthesis in Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil, and Telugu. One volunteer speaker per language read 1000 Wikipedia sentences selected for phonetic balance, result-

ing in about an hour and a half of speech per database. Recordings are studio-quality.

All these TTS corpora were collected for research purposes and made publicly-available data. In future, it would be interesting to examine commercial-quality TTS data, although these are of course proprietary.

3.2. Other Professional Speech

The Simple4All Tundra Corpus [11] consists of approximately 60 hours of speech from 14 audiobooks, each in a different language, and each read by a single speaker (8 male, 6 female). It was collected for the purpose of providing found data in many languages for text-to-speech research. Hour-long subsets of the data in each language have also been released, which have been selected for neutral style using an active learning based approach [12]. We consider both the full corpus as well as the 1-hour subsets.

The Boston University Radio News Corpus (BURNC) [13] is a corpus of professionally read radio broadcast 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 setting 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 examined the broadcast radio news part of the corpus for our experiments here.

3.3. ASR Corpora

The CALLHOME corpus [14] consists of spontaneous, orthographically transcribed telephone conversations between native speakers of US English. The data includes 6 hours and 45 minutes of utterances from 86 different female speakers, 1 hour and 43 minutes from 32 male speakers, and 8 hours and 32 minutes from speakers whose gender was not annotated in the corpus. For this paper we only examine speakers of known gender.

The MACROPHONE corpus [15] was designed for the development of telephone-based dialogue systems, such as travel booking and other database-related tasks. The utterances were read by 5,000 speakers over the phone. The data includes speech from male and female adults and children. We restricted our study to adult speaker of known gender (about 63 hours of male speech and 84 hours of female speech).

The IARPA BABEL program [16] focused on the rapid creation of spoken keyword search systems for a diverse set of languages which have historically not received a great deal of attention from the speech research community. While the goal of BABEL was primarily a speech recognition and spoken keyword search task, we are currently using some of this multi-speaker, conversational telephone data collected in 25 different languages for BABEL to build TTS voices for these languages. This data consists of both scripted and conversational telephone speech data from a variety of LRLs; in this work, we examine Telugu, Amharic, and Turkish. The unscripted speech was recorded from a variety of native speakers conversing over the telephone.

4. Features, Tools, and Methods

We extracted mean and standard deviation of f0, energy, speaking rate (measured in syllables per second), articulation, which we defined as (energy / speaking rate) * standard deviation of pitch (such that a high articulation value would correspond to loud energy, slow speaking rate, and large variation in pitch), NHR, jitter, and shimmer. Speaking rate was determined using either the syllable labels included in the data for which a Festival [17] frontend or existing labels were available; or a Praat script for approximating syllable nuclei otherwise [18]. Acoustic features were extracted using Praat [19]. Statistical comparisons were conducted using a sign test for a median, which tests the null hypothesis that two samples are from populations with the same median. We use a threshold of $p < 0.05$ for rejecting the null hypothesis. In addition, speakers were compared through bar graphs plotting mean with +/- 1 SD.

5. Acoustic Features

According to [1], TTS speakers are typically instructed to speak as consistently and with as little variation as possible. Interestingly we did not observe that TTS speakers had consistently lower standard deviation of f0 than did speakers in other genres; in particular, male TTS data from the ARCTIC and IIIT-H corpora had some of the highest SD of f0 out of all the male data we examined. Nevertheless, the female data matched expectations better – female TTS data tended to have lower SD of f0 than other genres. We compared the two ARCTIC female speakers ('slt' and 'clb') to each other in terms of SD of f0, and then compared them to the pool of female CALLHOME ASR data (Table 1). Comparing ARCTIC speakers to each other, the p-value is >0.05, and comparing ARCTIC to CALLHOME, the p-value is <0.05. Thus, female ARCTIC speakers are relatively similar in terms of standard deviation of f0, and significantly different from their conversational English ASR data counterparts by the same metric, suggesting that female TTS speakers can indeed be characterized by low SD of f0. While we do not observe consistently lower standard deviations for pitch of TTS data across genders, we do interestingly observe that all TTS corpora in each language show a lower *mean* pitch than other corpora in the same language and gender (Figure 1, although for this study not all genres were available for all languages and genders, so only available ones are shown; error bars show +/- 1 SD). This is consistent with anecdotal reports that listeners generally prefer TTS voices with lower pitch, and with our experimental findings [5] that training a voice on a subset of the lowest mean f0 utterances produces a voice that is preferred by listeners over the baseline.

Table 1: *Sign tests: ARCTIC female speakers compared to each other and to CALLHOME female data for SD of f0*

Comparison	***p*-value**
slt vs. clb (ARCTIC)	0.336
slt vs. CALLHOME	5.65E-245
clb vs. CALLHOME	1.01E-245

With respect to energy, we can see a clear separation between the conversational speech collected for ASR and corpora of professional and read speech (Figure 2). As expected, the TTS corpora, radio broadcast news, and audiobooks all have lower standard deviations of energy, whereas MACROPHONE, CALLHOME, and BABEL speech all have higher standard deviations.

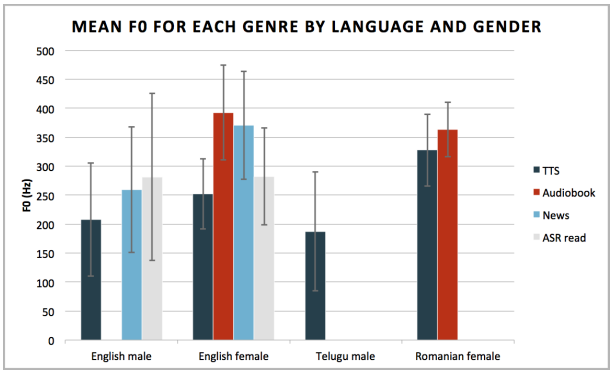

Figure 1: *TTS corpora exhibit lower mean pitch than other genres in the same language.*

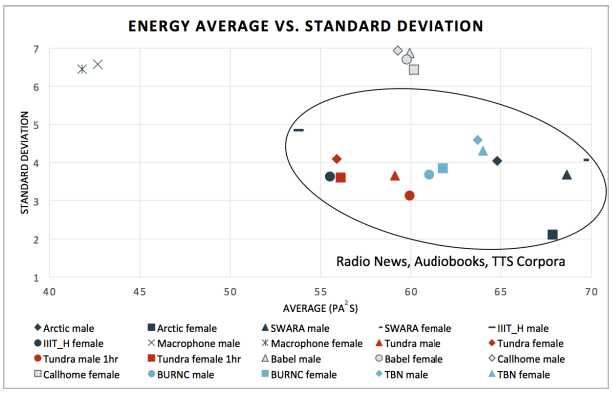

Figure 2: *Audiobooks, broadcast news, and TTS corpora have lower standard deviations for energy.*

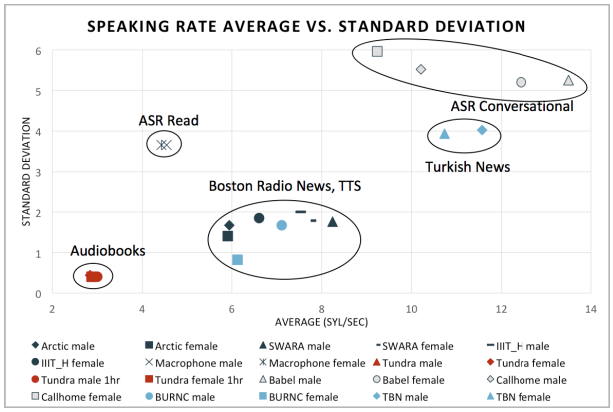

Figure 3: *Audiobooks, broadcast news, and TTS corpora have lower average speaking rates and SDs than found data.*

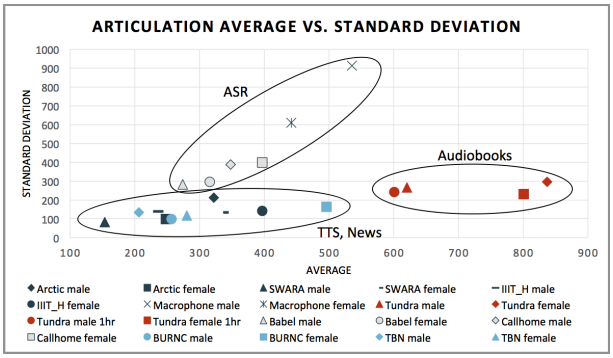

Figure 4: *Audiobooks have the highest levels of articulation.*

6. Speaking Rate and Articulation

We found that the different corpora (divided by gender) showed clear clusters with respect to mean and standard deviation of speaking rate (Figure 3). ASR corpora had both a high mean and standard deviation of speaking rate, and audiobooks had the lowest values for both. The TTS corpora clustered together with the broadcast news. MACROPHONE, which is read speech, showed a slower speaking rate as well. There did not appear to be significant gender differences.

Audiobooks show the highest level of articulation (Figure 4), as expected, explained by the fact that speakers in audiobooks may over-articulate to portray particular characters. The conversational data shows much more variation. TTS data and news show the lowest standard deviation for articulation. We also observe that the 1-hour Tundra subsets selected with an active learning based approach [12] show both a lower mean and standard deviation for articulation as compared to the full Tundra corpus, indicating that lower mean and variance of articulation corresponds with their learned model of human judgments of neutral style. In fact, the audiobooks show a substantially higher level of articulation than the TTS data, for languages for which both genres are available (English and Romanian) (Figure 5). In previous work, we have found that selecting less-articulated utterances for training TTS voices on found data produces better voices, both in terms of naturalness and intelligibility, likely because this data is more similar to TTS data [5, 6, 7].

7. Voice Quality

The main differences we observed for gender were mainly differences in voice quality features, although some corpus effects could be observed as well. Speakers from the ARCTIC, SWARA, and BURNC corpora tended to be on the lower end of both average and standard deviation of shimmer and NHR *for their gender*, with female speakers typically having lower values overall for average and standard deviation of NHR and jitter (a less clear gender effect was observed for shimmer). The CALLHOME female data was a notable outlier for all three features, having high means and standard deviations, and not clustering well with the other female data as a result. Furthermore, we observed that the 1-hour subsets of the Tundra data selected for neutral style had a lower standard deviation for jitter, and lower means and standard deviations for shimmer and NHR, than the full Tundra data. Plots of the data for each gender in each corpus can be seen in Figures 6, 7, and 8.

8. Discussion

We have measured whether TTS data does in fact follow the recommendations typically given to speakers, who are typically instructed to speak with very little variation in their voice quality, speaking style, pitch, volume, or tempo [1]. However, when we compare these features across multiple genres we we find that TTS speakers do not always differ from speakers in other genres in these characteristics. More importantly, we have found that several found-data genres do model TTS corpora in important ways.

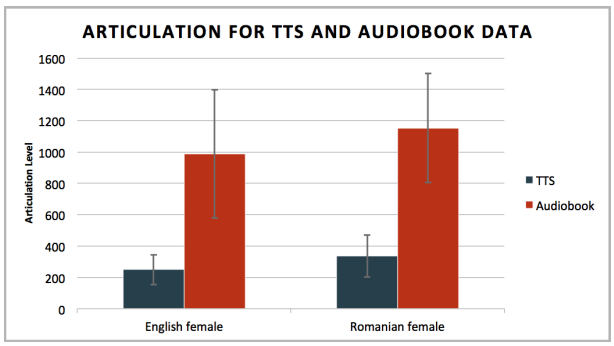

Figure 5: *Audiobooks are substantially more articulated than TTS corpora.*

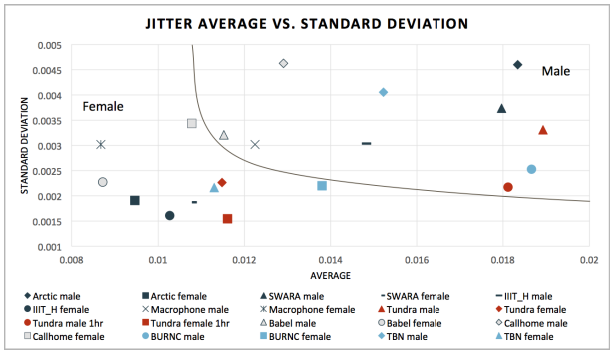

Figure 6: *Female speakers tend to have lower mean and standard deviation for jitter.*

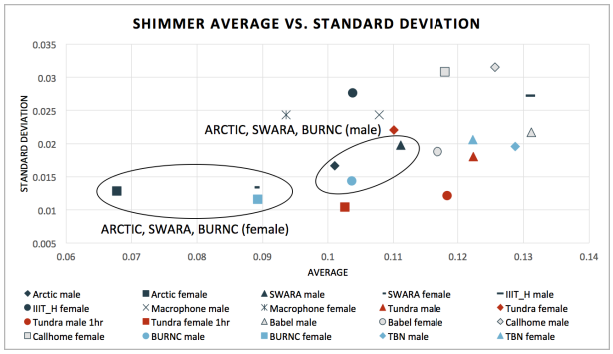

Figure 7: *The ARCTIC, SWARA, and BURNC corpora have relatively low means and standard deviations for shimmer.*

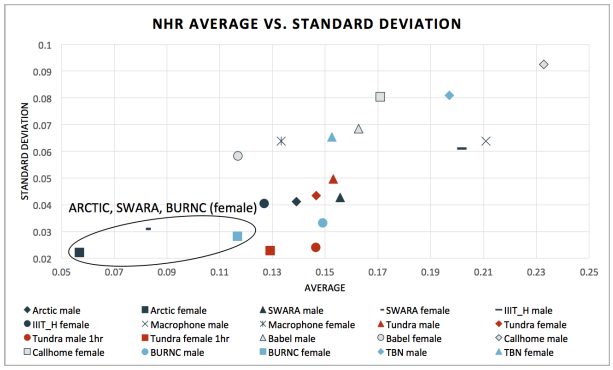

Figure 8: *Female speakers tend to have lower means and standard deviations for NHR.*

First, we have found that a low standard deviation of f0 does not in fact consistently characterize TTS corpora, and thus may not be an important feature for selecting found data for building TTS systems. While female TTS data does have a lower SD of f0 relative to other types of data, especially compared to conversational ASR speech, this is not the case for male TTS data. However, we have found that TTS data does tend to have a lower *mean* pitch relative to other genres. We have also found that TTS data has a lower SD of energy, similar to broadcast news and audiobooks, but different from ASR corpora. TTS data is also quite similar to broadcast news in terms of speaking rate and level of articulation (relatively low mean and SD), whereas audiobooks tend to have an even lower speaking rate and a very high level of articulation. Finally, TTS speakers exhibit low mean and SD for shimmer and NHR, much like professional broadcast news speakers and audiobook readers. These findings suggest objective justification for building TTS systems from particular found-data genres but also indicate the criteria that should be used to select data from other, less similar corpora, for building TTS systems.

9. Conclusions and Future Work

In this work, we have identified features that characterize TTS corpora as well as which found-data corpora are most similar to TTS data – radio news and audiobooks. In so doing, we have also identified which features are important to use when choosing subsets of utterances from the less similar genres we examined, such as conversational corpora. Thus, we can not only demonstrate *why* certain genres are particularly well adapted for TTS voice construction but we can predict, from empirical findings, what subsets of other corpora should be either included or excluded from TTS voice construction. In the future, we would like to extend this work to additional languages for which we do have TTS data in order to enable a fuller comparison across languages. We would like to further examine language effects versus genre effects as well, since our features may be language-dependent. We would also very much like to compare commercial TTS corpora to the research TTS corpora used in these experiments.

We would also like to use our findings to construct machine learning based approaches for identifying the best parts of a found-data corpus to use for building TTS voices, similar to [12], which used active learning to discover sentences similar to a small set of audiobook utterances labeled as 'neutral' by a human listener. There are additional challenges we would like to explore for automatically selecting the most neutral or TTS-like utterances from other genres of found data other than audiobooks, such as variation in recording quality, presence of background noise, and a potentially large number of different speakers. If we can develop approaches to select the most TTS-like utterances from heterogeneous sources of found data, this will enable us to more quickly and easily build intelligible and natural-sounding TTS voices for many low-resource languages.

10. Acknowledgements

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

11. References

[1] J. Matoušek, D. Tihelka, and J. Romportl, "Building of a speech corpus optimised for unit selection tts synthesis," *LREC*, 2008.

[2] 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.

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

[6] E. Cooper, A. Chang, Y. Levitan, and J. Hirschberg, "Data selection and adaptation for naturalness in HMM-based speech synthesis," *INTERSPEECH*, 2016.

[7] 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.

[8] J. Kominek, A. W. Black, and V. Ver, "CMU Arctic databases for speech synthesis," Tech. Rep., 2003.

[9] A. Stan, F. Dinescu, C. Ţiple, Ş. Meza, B. Orza, M. Chirilă, and M. Giurgiu, "The SWARA speech corpus: A large parallel romanian read speech dataset," *th Conference on Speech Technology and Human-Computer Dialogue*, 2017.

[10] K. Prahallad, E. N. Kumar, V. Keri, S. Rajendran, and A. W. Black, "The IIIT-H Indic speech databases," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[11] 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.

[12] 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," *8th ISCA Speech Synthesis Workshop*, 2013.

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

[14] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME american english speech corpus ldc97s42," *DVD*, 1997.

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

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

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

[18] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods, vol. 41, no. 2, pp. 385-390*, 2009.

[19] P. Boersma, "Praat, a system for doing phonetics by computer," *Clot International, vol. 5, no. 9-10, pp. 341–345*, 2001.