TTS and Data Selection

Improving Speech Synthesis Systems for Low Resource Languages

Yocheved Levitan

Columbia University

yocheved.levitan@gmail.com

***Abstract*-- The widespread use of Text-to-Speech (TTS) technology in our lives is evidence of its success. As the technology for generating artificial speech improves, the market for TTS systems increases. Considerable progress has been made in synthesizing natural and intelligible synthetic voices with these systems. Since synthesizing speech requires large quantities of processed audio and text data, this success is limited to languages with the proper resources. Unlike English, French, and Chinese, languages like Telugu, Tok Pisin, and Lithuanian do not have adequate or satisfactory resources to facilitate the development of speech tools, including TTS systems. This reality severely hinders communication and accessibility for speakers of these low-resource languages (LRLs). Our paper describes research in the field of TTS to develop tools and algorithms to be implemented in a TTS system for LRLs. More specifically, we address the question of how low-resource TTS systems can be improved using data selection.**

I. Background

A. TTS

The two primary techniques for generating artificial speech are concatenative and HMM-based synthesis [1]. Concatenative speech systems combine small units of sound together to produce a specific word, while HMM-based systems employ Hidden Markov Models to create speech waveforms. In a setting where natural speech is required and the lexical domain is limited, concatenative systems are powerful. Because concatenative systems are built from speech recorded in an ideal environment by the target speaker, the resulting synthesis will sound like a natural recording. Recently, there has been a shift towards HMM-based synthesis [2]. This synthesis is parametric in nature and is therefore extremely flexible. Whereas in a concatenative system the synthesized voices are limited to the nature of the audio corpus, with HMM-based synthesis the audio can be manipulated to synthesize voices with different qualities. For example, expressive speech can be generated using an existing conversational corpus by manipulating parameters, eliminating the cost and effort required to record a new animated corpus (as would be necessary for concatenation). The drawback of the HMM approach is that in general, the speech produced sounds more robotic than speech synthesized by a concatenative system. Because of the many advantages of parametric systems over concatenative systems, it is important that we study how to synthesize natural speech with HMM-based synthesis.

B. Data Selection

A prerequisite for speech synthesis is the availability of text and audio data. In order to generate new speech, there must be existing speech with its corresponding transcriptions to work with. Collecting and annotating low resource audio data that is acceptable for training is a challenge since recording and transcribing audio for TTS purposes is costly and time-consuming.

Developers of low-resource tools must consider other options for collecting data. An effective alternative to developing new audio corpora is to use ‘found’ data - materials available for download from the web [3]. This is data that is free and easily accessible. The problem with found data is its unpredictability. Scraping the web for TTS training material can increase the quantity of input data, but the quality of the output data may be negatively affected. If we include unknown material in our synthesis, we might obtain substandard results. One solution to this quandary is to select only choice data to include in our training set. Data selection is the process of filtering a large dataset to remove unwanted materials. By filtering our synthesis materials by certain acoustic features, we believe that we can effectively weed out such data. The subset data that we select from our pool of found data will be limited to audio that is likely to produce quality speech.

Our task was to determine which audio features to use in filtering in order to produce intelligible and natural synthetic speech. Once this is accomplished, we can select subsets of a larger dataset with appropriate feature values to use for training. We hypothesize that applying data selection techniques will enable TTS developers to take advantage of the quantity of low-resource data without compromising the quality of the resulting speech.

II. Process

A. Overview

To test our hypothesis we set up experiments to evaluate the voice quality of voices synthesized with selected data. TTS voices are typically judged according to two measurements: intelligibility and naturalness. We developed a listening test, or Human Intelligence Task (HIT), for each of these variables using Amazon's crowdsourcing platform, Mechanical Turk. Using the information that we collected from our HITs we hope to identify which filtering features perform best in producing intelligible and natural speech. With this knowledge, we will be able to process new data by selecting the superior data segments based on their feature values. This ability is of crucial benefit to the goal of obtaining quality audio data for low resource TTS systems. We split the task into two domains: naturalness and intelligibility. We evaluated the results of each experiment individually. This paper concentrates on our research and findings for naturalness.

B. Methods

Although our end goal is to improve TTS systems for LRLs, we began experimenting with English, a familiar language. If our results are promising, we will then apply our techniques to LRLs and have native speakers of those languages evaluate our efforts. Last summer we processed the Boston Radio News Corpus [4], to optimize the data for selection. This involved segmenting the utterances into smaller units to target individual sentences rather than paragraphs. The corpus features 7 speakers, 4 male and 3 female radio broadcasters. In our initial studies, we focused on the female speakers. With this processed data as our base, we extracted features from the audio files using Praat [5], an open source toolkit for speech processing. We chose the following features: mean and standard deviation values for pitch and energy (volume), speaking rate, utterance length, and quality of articulation. These features were chosen because we suspect that they are likely to affect the perceived naturalness of synthesized speech. If an utterance is hyper-articulated, for example, it might be classified as unnatural. Once we obtained a value for each audio file, we sorted the files in ascending, descending, and middle (starting from the median value and expanding outwards in both directions) order. We organized the files in this manner to enable us to select audio files with extreme values (high/low) for these features to use for our training. A middle order was included in our sort types to contrast the extreme voices with an average model. The next step was taking these sorted lists of audio files and segmenting the data into 15 minute, 30 minute, 1 hour and 2 hour subsets. This was accomplished by adding up the durations of each audio file until the desired time limit was reached. Finally the ¼, ½, 1, and 2 hour subsets were synthesized with the HMM- based speech synthesis system (HTS) [6].

C. Evaluataion

We adhere to the standard conventions detailed in [7, 8] for evaluating synthetic speech. For assessing intelligibility we use the Semantically Unpredictable Sentences (SUS) method. The goal of this test is to transcribe a sentence as accurately as possible. If the sentence itself makes no sense in clear speech, then the user cannot infer the content of the sentence based on the context. If the user can still transcribe the words of the sentence, then we judge the voice to be intelligible. For example, the sentence "The table walked through the blue truth", if transcribed correctly, can verify that the evaluator was listening attentively, and that the speaker's speech is intelligible.

Evaluating naturalness is more challenging because it is inherently more subjective than intelligibility. It is difficult to quantify naturalness. Following the procedures outlined in [9], we designed a naturalness test. We selected arbitrary neutral sentences from the fable "Jack and the Beanstalk". The selected sentences were of varied length. The instructions were to listen to a series of voices speaking the same sentence and choose the best category to describe the naturalness of each voice from a 5 point Likert scale, where 1 = very unnatural, 2 = somewhat unnatural, 3 = neither natural nor unnatural, 4 = somewhat natural, and 5 = very natural. The listeners were unaware of how the voice had been produced, as they only saw coded audio file names.

To obtain quality results from our experiment we added the following elements to the naturalness test. First, we verified that each voice was rated by 5 unique individuals. Additionally, we included 2 reference voices to our playlist of synthetic voices - a robotic voice (generated using Mac OSX's say command, Zarvox speaker) that the investigators agreed was clearly unnatural, and a natural human voice, resynthesized. The soft check for approving subjects' ratings was whether the robotic voice was rated very unnatural or somewhat unnatural, and that the human voice was rated somewhat natural or very natural. We randomized the order of presentation of the voices, placing our reference voices in the bottom half of the list (so as not to skew the opinions of the listeners), and recorded the resulting order for each HIT to investigate order effects. Lastly, we added functionality to the HIT to guarantee that the user listened to the entire audio clip before selecting their response. In total, we synthesized 80 voices. To limit the time that it took to complete the HIT we only included 1 hour subset voices in our experiment, bringing the number of voices down to 20. With the reference voices, and an additional voice trained on the entire dataset, each playlist consisted of 23 voices.

III. Results

Within a few days of posting the naturalness HIT all of the responses were in. It was evident from the responses that the evaluators were paying attention, as all of our reference voices were rated correctly. We were able to approve 100% of the work. Table 1 shows the average rating that each voice received. The audio files with the lowest and highest average ratings are the robotic and human voice, respectively. The remaining voices received relatively poor naturalness scores, all below the average score of 3.

TABLE 1 – RESULTS SUMMARY

Subsets with Feature ID	**Avg Rating**
robotic	1.03
meanHighPitch	1.97
highArticulation	2.08
meanHighEnergy	2.08
mediumDuration	2.08
lowSpeakingRate	2.13
stdvHighEnergy	2.13
stdvMiddleEnergy	2.28
shortDuration	2.33
lowSpeakingRate	2.13
stdvLowEnergy	2.37
stdvHighPitch	2.37
middleSpeakingRate	2.4
meanMiddleEnergy	2.42
meanLowEnergy	2.42
longDuration	2.5
highSpeakingRate	2.55
meanMiddlePitch	2.55
stdvMiddlePitch	2.6
stdvLowPitch	2.62
brnf_nonsat (no data selection)	2.68
lowArticulation	2.7
meanLowPitch	2.7
natural_vocoded	4.95

Our findings are consistent with studies that show that naturalness is perceptible [10]. The question of whether the naturalness of these synthetic voices approach the naturalness of a human voice, remains. Both the robotic and human voices received appropriate scores on the Likert scale. The synthetic subset voices received an average rating above the average rating for the robotic voice, and below the average rating for the human voice. However, the differences between the ratings for the synthetic voices on average are slight. Despite our precautionary tactic of placing the reference voices in the second half of the playlist, it is possible that the listeners were influenced by these baselines and that this skewed their opinions. Another possible explanation for our results is the fact that our voices were too alike to begin with. If all of the synthetic voices were easily distinguishable from each other, it is likely that our results would be similarly distinguishable. A third possibility is that we should have ranked or compared the voices instead of rating each voice individually.

IV. CONCLUSION

At present we have performed a general survey of the experiment data. Beyond these preliminary statistics, there is much to be explored. As mentioned above, we would like to study how listening order affects the naturalness scores. In addition, we plan to compare the similarity between ratings of different Turkers for each voice. Based on our results it is difficult to answer the question of which features are the best features to use in filtering - the differences between the ratings are too small. However, the framework that we created for extracting features from audio files, sorting the data, creating subsets, and synthesizing training material, will ease the process of performing and analyzing additional experiments to pursue these questions.

ACKNOWLEDGMENT

Many thanks to Dr. Julia Hirschberg of Columbia University for being the quintessential research mentor, and to PhD student Erica Cooper for her guidance throughout this project. I would also like to thank the sponsors and coordinators of the DREU (Distributed Research Experience for Undergraduates) program for their generosity, enabling this research.

REFERENCES

[1] Tiomkin, Stas, et al. "A hybrid text-to-speech system that combines concatenative and statistical synthesis units." *Audio, Speech, and Language Processing, IEEE Transactions on* 19.5 (2011): 1278-1288.

[2] Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*(11), 1039-1064.

[3] Mendels, Gideon, et al. "Improving Speech Recognition and Keyword Search for Low Resource Languages Using Web Data."

[4] Ostendorf, Mari, Patti J. Price, and Stefanie Shattuck-Hufnagel. "The Boston University radio news corpus." *Linguistic Data Consortium* (1995): 1-19.

[5] Boersma, Paul, and David Weenink. "Praat, a system for doing phonetics by computer." (2001): 341-345.

[6] Tokuda, K., Zen, H., Yamagishi, J., Black, A., Masuko, T., Sako, S., Toda, T., Nose, T., and Oura, K., 2008. The HMM-based speech synthesis system (HTS). http://hts.sp.nitech.ac.jp/.

[7] Benoît, Christian, Martine Grice, and Valérie Hazan. "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility Semantically Unpredictable Sentences." *Speech Communication* 18.4 (1996): 381-392.

[8] Karaiskosa, Vasilis, et al. "The blizzard challenge 2008." (2008).

[9] Georgila, Kallirroi, et al. "Practical Evaluation of Human and Synthesized Speech for Virtual Human Dialogue Systems." *LREC.* 2012.

[10] Mixdorff, Hansjörg, and Dieter Mehnert. "Exploring the naturalness of several German high-quality-text-to-speech systems." EUROSPEECH. 1999.