Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 194 – 200

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Building Statistical Parametric Multi-speaker Synthesis for Bangladeshi Bangla

Alexander Gutkin*, Linne Ha, Martin Jansche*, Oddur Kjartansson, Knot Pipatsrisawat, Richard Sproat*

Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

Abstract

We present a text-to-speech (TTS) system designed for the dialect of Bengali spoken in Bangladesh. This work is part of an ongoing effort to address the needs of new under-resourced languages. We propose a process for streamlining the bootstrapping of TTS systems for under-resourced languages. First, we use crowdsourcing to collect the data from multiple ordinary speakers, each speaker recording small amount of sentences. Second, we leverage an existing text normalization system for a related language (Hindi) to bootstrap a linguistic front-end for Bangla. Third, we employ statistical techniques to construct multi-speaker acoustic models using Long Short-term Memory Recurrent Neural Network (LSTM-RNN) and Hidden Markov Model (HMM) approaches. We then describe our experiments that show that the resulting TTS voices score well in terms of their perceived quality as measured by Mean Opinion Score (MOS) evaluations.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: TTS, Bangladesh, HMM, LSTM-RNN, acoustic modeling

1. Introduction

Developing a text-to-speech (TTS) system is a major investment of effort. For the best concatenative unit-selection systems [1], many hours of recording are typical, and one needs to invest in careful lexicon development, and complex rules for text normalization, among other things. All of this requires resources, as well as curation from native-speaker linguists.

For low-resource languages it is often hard to find relevant resources, so there has been much recent work on methods for developing systems using minimal data [2]. The downside of these approaches is that the quality of the resulting systems can be low and it is doubtful people would want to use them.

We are therefore interested in approaches that minimize effort, but still produce systems that are acceptable to users. This paper describes our development of a system for Bangla, the main language of Bangladesh and a major

* Corresponding authors

E-mail addresses: agutkin@google.com (Alexander Gutkin)., linne@google.com (Linne Ha)., mjansche@google.com (Martin Jansche)., oddur@google.com (Oddur Kjartansson)., thammaknot@google.com (Knot Pipatsrisawat)., rws@google.com (Richard Sproat).

1877-0509 © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Peer-review under responsibility of the Organizing Committee of SLTU 2016

doi:10.1016/j.procs.2016.04.049

language of India, and in particular the speech, lexicon and text normalization resources, all of which we are planning to release, under a liberal open-source license.

A core idea is the use of *multiple ordinary speakers*, rather than a single professional speaker (the normal approach). There are two main justifications. First, voice talents are expensive, so it is more cost-effective to record ordinary people; but these quickly get tired reading aloud, limiting how much they can read. We thus need multiple speakers for an adequate database. Second, there is an added benefit of privacy: we can create a natural-sounding voice that is not identifiable as a specific individual.

Unit selection[1] is a dominant approach to speech synthesis, but it is not suitable when working with multiple speakers, one obvious reason being that the system will often adjoin units from different speakers, resulting in very unnatural output. Instead we adopt a statistical parametric approach[3]. In statistical parametric synthesis the training stage uses multiple speaker data by estimating an averaged representation of various acoustic parameters representing each individual speaker. Depending on the number of speakers in the corpus, their acoustic similarity and ratio of speaker genders, the resulting acoustic model can represent an average voice that is very humanlike yet cannot be identified as any specific recorded speaker.

This paper is organized as follows: We describe the crowdsourcing approach to assemblying the speech database in Section 2. The TTS system architecture is introduced in Section 3. Next, experimental results are presented in Section 4. Finally, Section 5 concludes the paper and discusses venues for future research.

2. Crowdsourcing the speakers

We were familiar with collecting data from multiple speakers from data collection efforts for automatic speech recognition[4]. There, our goal was at least 500 speakers, of varying regional accents in different recording environments, recorded using mobile phones. For TTS, very different criterion is conventional: a professional standard dialect speaker in a recording studio. But this is expensive and cannot scale if one wants to cover the worlds many low-resource languages.

New statistical parametric synthesis methods[3] allow for building a voice from multiple speakers, but one still needs speakers that are acoustically similar. To achieve this, we held an audition to find Bangla speakers with compatible voices. 15 Bangladeshi employees at Google's Mountain View campus auditioned. From that sample, we sent a blind test survey to 50 Bangladeshi Googlers to vote for their top two preferences. Using the top choice – a male software engineer from Dhaka – as our reference, we chose 5 other male Dhaka speakers with similar vocal characteristics.

Our experience with crowd-sourced ASR data collection taught us the importance of good data collection tools. ChitChat is a web-based mobile recording studio that allows audio data to be collected and managed simply. Each speaker is presented with a series of sentences assigned to them for recording. The tool records at 48 kHz, detecting audio clipping to ensure quality, and ambient noise prior to recording each sentence, with a high noise level triggering an alert preventing further recording. Audio can be uploaded to the server or stored locally for later uploading.

For the recordings we used an ASUS Zen fanless laptop with a Neumann KM 184 microphone, a USB converter and preamp, together costing under US$2000. We recorded our volunteers over 3 days in June 2015. Each recorded about 250 phrases, averaging 45 minutes, mined from Bangla and English Wikipedia. Volunteers were first instructed on the "bright" style of voice we were interested in. After a supervised practice run of 10–15 minutes, the remainder was recorded independently while being observed remotely using ChitChats admin features. Recordings were stopped if the voice sounded tired or mouth-dry. The sessions yielded about 2000 utterances.

3. System Architecture

A typical parametric synthesizer pipeline consists of training and synthesis parts. Similar to Automatic Speech Recognition (ASR) pipeline[5], the training process consists of two steps: data preparation and acoustic model training[6]. During the data preparation step one extracts a parametric representation of the audio from the speech corpus. A typical acoustic representation includes spectral, excitation and fundamental frequency parameters, and pertinent linguistic parameters are extracted as well, which take into account linguistic and prosodic contexts for the current phoneme. Once acoustic and linguistic parameters are extracted, during the acoustic model training stage we use

machine learning techniques to estimate faithful statistical representations of the acoustic and linguistic parameters extracted by the previous step.

3.1. Phonology and lexicon

As with any TTS system, our Bangla system requires a phoneme inventory and a grapheme-to-phoneme conversion system. While the latter might be done with simple grapheme-to-phoneme rules, Bangla spelling is sufficiently mismatched with the pronunciation of colloquial Bangla to warrant a transcription effort to develop a phonemic pronunciation dictionary.

Consider the Bangla word for *telescope*, which is transcribed in IT3 transliteration[7] as `d uu ra b ii k shha nd a` and in IPA as /dur.bik.kʰɔn/. In this example there are several mismatches between the actual pronunciation and what we would expect on the basis of the spelling, including short /u/ and /i/ rather than the orthographically represented long vowels, and the cluster `k shh`, which is actually pronounced /k.kʰ/. The final letter transcribed as `nd a` has an inherent vowel, which is not pronounced in this case, but in other cases would be /o/ or /ɔ/. Indeed, the determination of the pronunciation of the inherent vowel (as /null/, /o/ or /ɔ/) is a major issue in Bangla.

Such reasons argue for the need for a hand curated pronunciation dictionary. We are aware of similar efforts[8,9], but none that are available for commercial use: in contrast, our own data is released[10].

Our phonological representation closely follows literature[11]. A team of five linguists transcribed more than 65,000 words into a phonemic representation of Bangladeshi colloquial Bangla, using a version of our phonemic transcription tools[12] and quality control methodology[13]. Our transcribers were further aided by the output of a pronunciation model, which was used to pre-fill the transcriptions of words so that transcribers could focus on correcting transcriptions, rather than entering them from scratch. The pronunciation model also provides important clues about the consistency and inherent difficulty of transcription.

In order to make our system available on mobile devices we employ LOUDS-based compression techniques[14] to encode the pronunciation lexicon into compressed representation of approximately 500 kB that is also fast enough for access.

3.2. Text normalization

The first stage of text-to-speech synthesis is text normalization. This is responsible for such basic tasks as tokenizing the text, splitting off punctuation, classifying the tokens and deciding how to verbalize non-standard words, i.e. things like numerical expressions, letter sequences, dates, times, measure and currency expressions[15]. The Google text normalization system, Kestrel[16], handles several different kinds of linguistic analysis, but here we focus on the tokenization/classification and verbalization phases, which use grammars written in Thrax[17].

For our Bangla system we benefited from already having a grammar for verbalizing numbers (used in ASR), and in addition we had a well worked out set of Kestrel grammars for the related language Hindi. Our target is Bangladesh, where very few people speak Hindi, but Bangla is also spoken in West Bengal in India. We therefore asked an Indian speaker of Bangla to translate all the Hindi content (about 1500 strings) in our Kestrel grammars into Bangla. The Hindi grammars were then converted using the Bangla translations. Inevitably some tweaking of the result was required and the fixing of issues is ongoing. However, bootstrapping a system from a closely related language is a reasonable approach if one is short of engineering resources to devote to the new language.

The various components of the normalization system are efficiently represented in our system as archives of finite state transducers (FSTs). There are there FST archives: rewrite grammar handles the basic rewriting of the incoming text and necessary unicode normalization, tokenizer and classifier grammar is responsible for text tokenization and detection of critical verbalization categories. Finally the verbalization grammar converts main verbalization categories into natural language text[16]. In the final system each grammar archive is losslessly compressed. The sizes of various Thrax FSTs before and after compression (and the corresponding compression ratios) are given in Table 1. The Bangla Kestrel grammars will be released along with the voice data. Also, in order for these to be useful, have developed a lightweight version of Kestrel called Sparrowhawk. This is already in the public domain and is in the process of being integrated with Festival open-source speech synthesis system[18].

Table 1. FST grammars and their disk footprint (in kilobytes).

Archive Type	Original (kB)	Compressed (kB)	Ratio
rewrite	117	22	×5.3
tokenize/classify	5429	1687	×3.2
verbalize	14190	3330	×4.3
total	19736	5039	×3.9

3.3. Synthesizer

The synthesis stage consists of two steps: First, a sentence is decomposed into corresponding linguistic parameters and acoustic model is used to predict a sequence of optimal acoustic parameters that correspond to linguistic ones. Second, the signal processing component, a vocoder, is used to reconstruct speech from the acoustic parameters [19]. In our system we use the state-of-the-art Vocaine algorithm [20] for the vocoding stage.

We have explored two acoustic modeling approaches. It is important to note in both approaches that we train all the speakers together and that the statistical nature of the acoustic modeling has the effect of averaging out the differences between the speakers in the original dataset. While the resulting acoustic parameters do not represent any particular person they can still nevertheless be used to reconstruct naturally sounding speech.

The first approach uses Hidden Markov Models (HMMs), and is a well-established parametric synthesis technique [21] In this approach we model the conditional distribution of an acoustic feature sequence given a linguistic feature sequence using HMMs.

One of the main limitations of HMMs is the frame independence assumption: HMM models typically assume that each frame is sampled independently despite concrete phonetic evidence for strong correlations between consecutive frames in human speech. One promising alternative approach that provides an elegant way to model the correlation between neighboring frames is Recurrent Neural Networks (RNNs) [22]. RNNs can also use all the available input features to predict output features at any given frame. In RNN-based approaches a neural network acoustic model is trained to map the input linguistic parameters to output acoustic parameters. In our work we use Long Short Term Memory (LSTM) architecture that has excellent properties for modeling the temporal variation in acoustic parameters and especially long-term dependencies between them [23,6]. LSTM models can be quite compact, making them particularly suitable for deployment on mobile devices.

4. Experiments

4.1. Experimental Setup

We experimented with a multi-speaker Bangla corpus totaling 1,891 utterances (waveforms and corresponding transcriptions) from five speakers selected during crowdsourcing process described in Section 2. The script contains total of 3,681 unique Bangla words which are covered by 40 monophones from Bangla phonology given in Section 3.1. Phone-level alignments between the acoustic data and its corresponding transcriptions have been generated using HMM-based aligner bootstrapped on the same corpus.

In order to account for phonemic effects such coarticulation the monophones were expanded using the full linguistic context. In particular, for each phoneme in an utterance we take into account its left and right neighbors, stress information, position in a syllable, distinctive features and so on, resulting in 271 distinct contexts. Expanding monophones in this fashion resulted in 21,917 unique full-context models to estimate.

The speech data was downsampled from 48 kHz to 22 kHz, then 40 mel-cepstral coefficients [24], logarithmic fundamental frequency (log F0) values, and 5-band aperiodicities (0-1, 12, 2-4, 4-6, 6-8 kHz) [25] were extracted every 5 ms. The output features of LSTM-RNNs were phoneme-level durations. The output features of the acoustic LSTM-RNNs were acoustic features consisting of 40 mel-cepstral coefficients, log F0 value, and band 5 aperiodicity. To model log F0 sequences, the continuous F0 with explicit voicing modeling approach [26] was used; voiced/unvoiced binary value was added to the output features and log F0 values in unvoiced frames were interpolated.

We built three parametric speech synthesis systems. The first configuration is an HMM system, which fits well on a mobile device[27]. This system is essentially similar to the one described by Zen et. al.[25]. We also build two LSTM-RNN acoustic models that are essentially the same apart from the number of the input features. The LSTM-RNN configuration with fewer (270) features is slightly smaller, portable (we excluded one feature that is resource-intensive to compute) and fast enough to run on a modern mobile device. In addition, for the embedded configuration we use audio equalizer to boost the audio volumes on the device. No dynamic range compression is employed for this configuration. Further details of LSTM-RNN configurations are described by Zen and Sak[6]. For all the configurations, at synthesis time, predicted acoustic features were converted to speech using the Vocaine vocoder[20].

To subjectively evaluate the performance of the above configurations we conducted a mean opinion score (MOS) tests. We used 100 sentences not included in the training data for evaluation. Each subject was required to evaluate a maximum of 100 stimuli in the MOS test. Each item was required to have at least 8 ratings. The subjects used headphones. In the MOS tests, after listening to a stimulus, the subjects were asked to rate the naturalness of the stimulus in a 5-scale score (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). 13 native Bangladeshi Bangla speakers participated in the experiment. Each participant had an average of minute and a half to rate each stimuli.

4.2. Results and Discussion

The results of MOS evaluations are shown in Table 2. In addition to regular MOS estimate we also report robust MOS estimate which is a mean opinion score computed using trimmed means (smallest and largest value are removed before computing a mean response for each stimuli). The MOS scores reported in Table 2 indicate that the three

Table 2. Subjective 5-scale MOS scores: regular (MOS) and trimmed (Robust MOS) estimates for speech samples produced by LSTM-RNN and HMM configurations, shown along with confidence intervals.

Model Type	5-scale MOS	5-scale Robust MOS
Server LSTM-RNN	3.403±0.098	3.424±0.101
Embedded LSTM-RNN	3.519±0.102	3.526±0.106
HMM	3.430±0.091	3.394±0.102

multi-speaker configurations are acceptable to the evaluators both in terms of naturalness and intelligibility – all the scores centering around the median between “Fair” and “Good”. The embeded LSTM-RNN configuration is preferred over server LSTM-RNNs. Since the number of input features for both models only differs by one, we hypothesize that the quality difference is due to the use of an audio equalization post-processing step which is employed in the embedded LSTM-RNN system.

The robust MOS confidence intervals (the numbers shown after the ± sign) for each configuration reported in Table 2 indicate no statistically significant difference between server and embedded LSTM-RNN configurations. This is indicated by the confidence interval overlap. On the other hand, the difference between HMMs and embedded LSTM-RNNs is statistically significant.

Interestingly enough, the HMM system did reasonably well: according to regular MOS score it is second behind the embedded LSTM-RNN. According to the robust MOS scores, the HMM system comes out worst out of the three systems but it is not very far behind the server LSTM. The difference in robust MOS scores between the two systems is 0.03, which is not very significant. We hypothesize that this is due to the size of the training corpus – HMM configuration may generalize reasonably well on a small dataset, whereas LSTM-RNNs may struggle with a small amount of data because there are too many parameters to estimate.

Following the subjective listening tests, the native speakers used the system in real-life scenarios (e.g., as part of machine translation). Out of approximately 25 bugs reported most of them were pronunciation errors due to the errors in lexicon transcription (or missing pronunciations) or text normalization issues. No reported problems are related to the actual quality of acoustic models.

5. Conclusion and Future Work

We described the process of constructing a multi-speaker acoustic database for Bangladeshi dialect of Bangla by the means of crowdsourcing. This database is used to bootstrap statistical parametric speech synthesis system that scores reasonably well in terms naturalness and intelligibility according to mean opinion score (MOS) criteria. We belive that the proposed approach will allow us to scale better to further under-resourced languages. While the results of our experiments are encouraging, there is still further research required into improving the scalability of the linguistic components: phonological definitions, lexica and text normalization. We would like to focus on this line of research next.

As we mentioned in this paper, we released the phonology and lexicons used in this work [10]. We are also finalizing the integration of Sparrowhawk text normalization framework with the Festival [18] system and will soon release the Bangla recordings and transcriptions used in our experiments.

References

1. Hunt, A.J., Black, A.W.. Unit selection in a concatenative speech synthesis system using a large speech database. In: *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*; vol. 1. IEEE; 1996, p. 373–376.
2. Sitaram, S., Palkar, S., Chen, Y.N., Parlikar, A., Black, A.W.. Bootstrapping text-to-speech for speech processing in languages without an orthography. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE; 2013, p. 7992–7996.
3. Zen, H., Tokuda, K., Black, A.W.. Statistical parametric speech synthesis. *Speech Communication* 2009;**51**(11):1039–1064.
4. Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P.J., LeBeau, M.. Building transcribed speech corpora quickly and cheaply for many languages. In: *INTERSPEECH*. 2010, p. 1914–1917.
5. Gales, M., Young, S.. The application of hidden Markov models in speech recognition. *Foundations and trends in signal processing* 2008; **1**(3):195–304.
6. Zen, H., Sak, H.. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In: *Acoustics, Speech and Signal Processing*. 2015, p. 4470–4474.
7. Prahallad, K., Elluru, N.K., Keri, V., Rajendran, S., Black, A.W.. The IIIT-H Indic Speech Databases. In: *INTERSPEECH*. 2012.
8. Alam, F., Habib, S., Sultana, D.A., Khan, M.. Development of annotated Bangla speech corpora. In: *Spoken Language Technologies for Under-resourced Languages (SLTU10)*; vol. 1. 2010, p. 35–41.
9. Habib, S.M., Alam, F., Sultana, R., Chowdhur, S.A., Khan, M.. Phonetically balanced Bangla speech corpus. In: *Proc. Conference on Human Language Technology for Development*. 2011, p. 87–93.
10. Google, . Bangla Phonology and Lexicon. `http://github.com/googlei18n/language-resources/tree/master/bn/data`; 2016.
11. Ud Dowla Khan, S.. Bengali (Bangladeshi Standard). *Journal of the International Phonetic Association* 2010;**40**(2):221–225.
12. Ainsley, S., Ha, L., Jansche, M., Kim, A., Nanzawa, M.. A Web-Based Tool for Developing Multilingual Pronunciation Lexicons. In: *INTERSPEECH*. Citeseer; 2011, p. 3331–3332.
13. Jansche, M.. Computer-Aided Quality Assurance of an Icelandic Pronunciation Dictionary. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., et al., editors. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4; 2014, p. 2111–2114. ACL Anthology Identifier: L14-1299.
14. Fuketa, M., Tamai, T., Morita, K., Aoe, J.i.. Effectiveness of an implementation method for retrieving similar strings by trie structures. *International Journal of Computer Applications in Technology* 2013;**48**(2):130–135.
15. Sproat, R., Black, A.W., Chen, S., Kumar, S., Ostendorf, M., Richards, C.. Normalization of non-standard words. *Computer Speech & Language* 2001;**15**(3):287–333.
16. Ebden, P., Sproat, R.. The Kestrel TTS text normalization system. *Natural Language Engineering* 2015;**21**(03):333–353.
17. Tai, T., Skut, W., Sproat, R.. Thrax: An open source grammar compiler built on OpenFst. In: *IEEE Automatic Speech Recognition and Understanding Workshop*. 2011.
18. Taylor, P., Black, A.W., Caley, R.. The architecture of the Festival speech synthesis system. In: *The Third ESCA Workshop in Speech Synthesis*. International Speech Communication Association; 1998, p. 147–151.
19. Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H.. TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In: *ICASSP 2008. IEEE International Conference on*. IEEE; 2008, p. 3933–3936.
20. Agiomyrgiannakis, Y.. VOCAINE the vocoder and applications in speech synthesis. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2015, p. 4230–4234.
21. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: *EUROSPEECH*. 1999, p. 2347–2350.
22. Tuerk, C., Robinson, T.. Speech synthesis using artificial neural networks trained on cepstral coefficients. In: *EUROSPEECH*. 1993, p. 1713–1716.
23. Fan, Y., Qian, Y., Xie, F., Soong, F.K.. TTS synthesis with bidirectional LSTM based recurrent neural networks. In: *Proc. Interspeech*. 2014, p. 1964–1968.
24. Fukada, T., Tokuda, K., Kobayashi, T., Imai, S.. An adaptive algorithm for mel-cepstral analysis of speech. In: *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*; vol. 1. IEEE; 1992, p. 137–140.

25. Zen, H., Toda, T., Nakamura, M., Tokuda, K.. Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE transactions on information and systems* 2007;**90**(1):325–333.
26. Yu, K., Young, S.. Continuous f0 modeling for hmm based statistical parametric speech synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on* 2011;**19**(5):1071–1079.
27. Gutkin, A., Gonzalvo, X., Breuer, S., Taylor, P.. Quantized HMMs for low footprint text-to-speech synthesis. In: *INTERSPEECH*. 2010, p. 837–840.