Text-To-Speech for Languages without an Orthography

Sukhada Palkar *Alan W Black* *Alok Parlikar*
Language Technologies Institute
Carnegie Mellon University
Pittsburgh (PA), USA
{spalkar, awb, aup} @cs.cmu.edu

ABSTRACT
Speech synthesis models are typically built from a corpus of speech that has accurate transcriptions. However, many of the languages of the world do not have a standardized writing system. This paper is an initial attempt at building synthetic voices for such languages. It may seem useless to develop a text-to-speech system when there is no text available. But we will discuss some well defined use cases where we need these models. We will present our method to build synthetic voices from only speech data. We will present experimental results and oracle studies that show that we can automatically devise an artificial writing system for these languages, and build synthetic voices that are understandable and usable.

TITLE AND ABSTRACT IN MARATHI

अक्षरपद्धती नसलेल्या भाषांसाठी वाणी संस्लेषण

ध्वनिमुद्रित वाक्यांच्या कोषापासून वाणी संस्लेषणाची संगणकीय प्रतिरुपे बनविण्यासाठी त्या कोषाची अचूक लिखित प्रतिलिपी उपलब्ध असावी लागते. जगातील अनेक भाषा मात्र मानांकित अक्षरपद्धती वापरत नाहीत. प्रस्तुत काम हे अशा भाषांसाठी संस्लेषित आवाज बनविण्याचा एक पहिला प्रयास आहे. मुळात अक्षरपद्धतीच नसताना त्या भाषेच्या लिखित पाठ्याचे वाणी संस्लेषण करण्याचे तंत्र हे व्यर्थ वाटू शकते. पण प्रस्तुत लेखात आम्ही या संस्लेषण प्रणालीचे काही प्रमुख उपयोग सुचवीत आहोत. केवळ ध्वनिमुद्रित वाक्यांचा कोष वापरून संस्लेपित आवाज बनविण्याची आमची पद्धत या लेखात आपण पाहू. आम्ही केलेले प्रयोग व विश्लेषण असे दर्शवितात की आपण एखादी अक्षरपद्धती आपोआप शोधू शकतो, जिचा वापर करून केलेले वाणी संस्लेषण सुगम व वापरण्याजोगे असते.

KEYWORDS: Speech Synthesis, Synthesis Without Text, Low Resource Languages, Languages without an Orthography.

KEYWORDS IN L_2: वाणी संस्लेषण, पाठ्याशिवाय संस्लेषण, संसाधन-दुर्लभ भाषा, अक्षरपद्धती नसलेल्या भाषा.

Proceedings of COLING 2012: Posters, pages 913–922,
COLING 2012, Mumbai, December 2012.

1 Introduction

Of the many languages in the world, most actually are only spoken, and do not have a writing system. Even for many of the languages that do have writing systems, the orthography is poorly standardized. Such languages typically have speakers that are not literate in those languages, even if they may be literate in other languages such as English.

Speech processing should offer the opportunity to communicate in all languages, and is perhaps even more valuable for languages where a written form is not well defined. This paper investigates how to build a text-to-speech system in languages where no well-defined writing system exists.

If text is fundamental to speech synthesis, what does it even mean to synthesize in a language that does not have text? We propose the following: Given a speech corpus in such a language, we automatically derive a writing system appropriate for that language. This could be a phonetic writing system that uses either a universal phone set, or a phone set from a closely related language. We use Automatic Speech Recognition technology to develop this writing system. This automatically derived, artificial writing system can then be used as "text" that is input to our text-to-speech system.

At first it might seem futile to develop a speech synthesis system without a related writing system. But consider these two use cases that highlight the need of such a system. The clearest use case, that underlies the reason for this work, is the development of a speech to speech translation system from a language that has a written form, into a language that does not. If we attempt to collect "parallel data" for training translation systems, we will end up with text in the source language, and only speech in the target language. But standard methods of machine translation require text to be present on both source and target sides. Our proposed artificial phonetic orthography can be used as the text in the target language to enable training of machine translation models. Note that such a system will essentially translate words in the source language into phonetic units of our artificial writing system. But such translation systems have been shown (Stueker and Waibel, 2008) to be possible. Another use case of our proposed method is in dialog systems. If the language of a dialog system does not have a written form, how will people write the prompts? And how will the synthesis happen? Our proposal will allow system developers to use the automatically derived writing system to write prompts that can then be synthesized.

Our goal is to develop synthetic voices in languages without orthography. However, in order to test our methods and illustrate our techniques, we have in fact used languages that do have well defined written forms, and speech and text corpora available. We have particularly used Marathi as the language in this research, built synthetic voices by pretending it did not have a writing system, and we will show results about how well these models do. We will also present similar results for Hindi and Telugu.

This paper is organized as follows. Section 2 describes all the data we have used in this research. Section 3 describes the basic strategy of developing a synthetic voice from speech data that has no transcriptions. In Section 4, we discuss a novel method of improving the quality of the synthetic voice. In Section 5, we comment on the nature of the artificial writing system we devise for languages at hand, and present conclusions towards the end.

2 Data and Resources

We used Indian languages (Marathi, Hindi, Telugu) in this work. Our method uses two resources: (i) Speech data in the target language, and (ii) Text data in a related high resource language.

For Marathi, we used about 30 minutes of speech data made available by Parlikar and Black (2012). For Hindi and Telugu, we used the speech corpora collected by Prahallad et al. (2012). These have about an hour each of single speaker speech. We used a corpus for Hindi and Marathi text made available by IIT Bombay CFILT, and crawled wikipedia articles for Telugu text. All the speech data we used is single channel clean speech, recorded in a studio setting at 16KHz.

Note that these three languages all have well defined written forms. Hindi and Marathi use the Devanagari script, and Telugu uses its own script. The speech corpora described above all have an associated transcript. For purposes of this research, we did not use the transcripts except to run an oracle evaluation of our models.

We used publicly available tools for speech recognition and speech synthesis. For recognition, we used the CMU Sphinx3 (Placeway et al., 1996) system. For building synthetic voices, we used the Festvox (Black and Lenzo, 2002) suite of tools. The voices we build use the clustergen (Black, 2006) method of statistical parametric synthesis. We used the Festival (Black and Taylor, 1997) system for speech synthesis.

3 Basic Approach to Synthetic Voices without Orthography

We have speech data in our target language, and there is no well defined orthography for transcriptions. A simple method to deal with this situation is to run an automatic speech recognizer over available speech data and use its output as transcriptions.

The caveat with using a speech recognizer is that because our target language does not have a text form, a speech recognizer will not exist in that language. We hence have to use a speech recognizer in another language: a language that has an orthography, and large corpora to train speech recognizers. This presents another caveat: we are recognizing in a different language than the models are trained for. Using the default language model is thus not ideal, and we need to adapt it so that it is suitable for our target language. We also use **phonetic** decoding instead of word level decoding.

We propose the following: (i) Choose an appropriate acoustic model for speech recognition, then (ii) Choose a language that has an orthography and is phonetically close to our target language, and then build a phonetic language model on text in this language. (iii) Run phonetic decoder on our target speech data with these two models and obtain transcripts. (iv) Build a voice using the speech data and the phonetic transcripts just obtained. Figure 1 illustrates this method.

We used this method and ran experiments on our Marathi data. We assumed Marathi to be the language that has no orthography. We considered English and Hindi to be the languages that have high resources available, and those that have an orthography.

3.1 Decoding with an English Acoustic Model

We used an English acoustic model trained on the Wall Street Journal data that we obtained from the CMU-Sphinx website. This model uses the CMU-DICT US English phone set, which

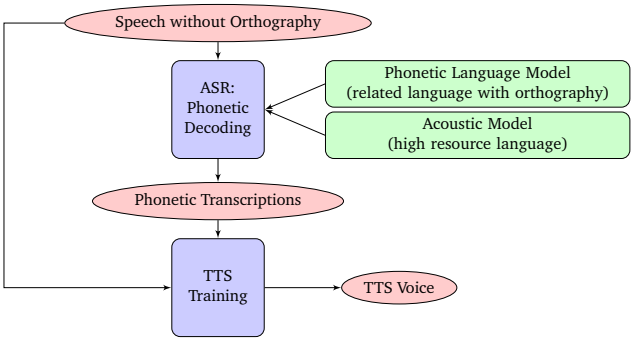

Figure 1: An overview of our basic approach

consists of 39 phones.

Keeping the acoustic model fixed, we decoded our speech with three different language models: (i) An English Phonetic Language Model trained on part of the BTEC Data , (ii) A Hindi Phonetic Language Model trained on the Hindi text corpus, and (iii) A Marathi Phonetic Language Model trained on the Marathi text, for oracle comparison.

Since the phone set of the acoustic model is CMU-DICT, we wrote a tool that converts Indic Unicode Script into CMU-Dict phone strings. The language models from Hindi and Marathi text mentioned above were built on this phonetized text.

With these acoustic and language models, we decoded the Marathi speech data and obtained transcriptions. We then built a phone-based clustergen voice using this data. We held out 10% of the data for evaluating the synthetic voice. We synthesized the test set, aligned it using dynamic-time-warping to the original speech, and computed the spectral distance (MCD) (Mashimo et al., 2001) between the two as the evaluation measure. Since this is a distance, lower is better. Kominek (2009) has showed that a difference of 0.1 in the MCD is perceptually significant.

Table 1 shows the quality of synthesis obtained using the different language models. We see that using a language model trained on Hindi is better than one trained on English. This could be because Marathi is phonetically much closer to Hindi than to English. Notice also that the model obtained with Hindi language model is almost as good as the oracle result of using a Marathi language model. This shows promise in the use of sister languages for language modeling.

Language Model	**MCD of Synthesis**
Phonetic English	7.391
Phonetic Hindi	7.124
Phonetic Marathi (Oracle Result)	7.117

Table 1: MCD of Synthesis using English acoustic model and different language models

3.2 Decoding with an Hindi Acoustic Model

The CMU-DICT phone set is very different from the set of phones that Marathi uses. We investigated whether using an acoustic model from a closely-related language could yield improvements. We used the Hindi speech data we have to train an acoustic model. However, this data was only an hour of female speech. Our Marathi data is recorded by a male speaker. The gender mismatch, and the small size of training data yielded a very weak acoustic model. After decoding with this acoustic model and a language model trained on the larger Hindi text corpus and repeating our voice build, we were left with a synthesizer that had an MCD of 7.868. We believe that with more training data for a Hindi acoustic model, we might have a better voice than with an English acoustic model.

3.3 Extended CMU-DICT phone set

CMU-DICT is an English phone set, and English is not very similar phonetically to Marathi. We hence explored enhancing the CMU-DICT phoneset. Specifically, we investigated whether splitting English vowels into finer groups of short and long vowels could improve our models. We decoded the speech data with the previous English acoustic model. We then aligned the speech to these phonetic transcripts using an EHMM alignment tool (Prahallad et al., 2006). We determined the duration of different vowels and clustered them into two groups based on the duration. We then labeled these vowel clusters as being two different vowels when training the synthetic voice. We saw marginal improvements to the MCD of synthesis using this method, but did not yet explore this in more detail.

4 Targeted Acoustic Model for Improved Synthetic Voices

In the previous section, we saw that our best baseline synthetic voice comes from transcriptions derived using an English acoustic model. We explored if we could target the acoustic model to the speech database at hand and get an improved result over all.

4.1 Method Description

We use a bootstrapping method. First, we use the English acoustic model and obtain baseline transcriptions for our target speech. Using these transcriptions and the speech data, we train a targeted acoustic model. This model is a small acoustic model, but it is specific to the data we are using. Using this new acoustic model, keeping rest of the decoding process similar, we decode our speech data again. We get a new set of transcriptions. We train another targeted acoustic model with these new transcriptions and repeat the iterations until the MCD on a held out test set stops improving. Figure 2 shows a flow diagram for this training.

4.2 Experiments and Results

We started with our Marathi speech data (assumed again, that Marathi had no writing system). We used the baseline speech recognition system as described in Section 3. We then applied the described iterative method to build and use a targeted acoustic model. We obtained very good improvements as evaluated objectively using the MCD distance. We then repeated similar experiments for Hindi and Telugu. We assumed that Hindi had no orthography, used the Wall Street Journal acoustic model and a Marathi Phonetic Language Model for recognition. For Telugu, we used the same acoustic model and the Hindi Phonetic Language Model.

The results of these experiments are plotted in Figure 3. We see that for all three languages,

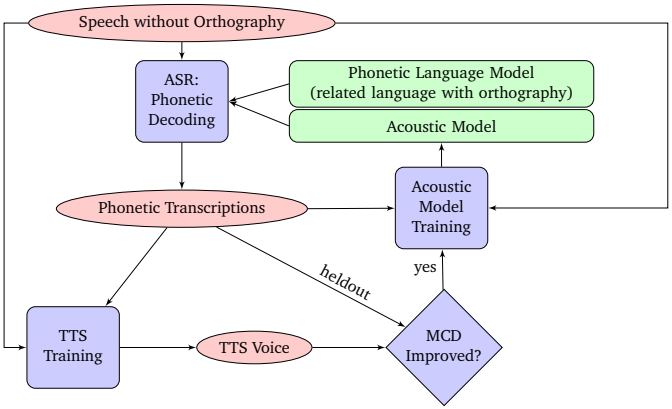

Figure 2: Building a targeted acoustic model for improved synthesis

the iterative targeting of acoustic model ultimately produces a synthesizer that is better than the baseline. The range of MCD values in each language depends on the recording conditions and the speaker. We see that the improvement over the baseline is more for Telugu, compared to the other two languages and we are investigating why.

Given that we see consistent improvements in MCD using the proposed method, we evaluated whether these improvements are perceptually meaningful. We ran listening tests on the Marathi data. We compared the baseline Marathi voice to the voice after 6 iterations of acoustic model targeting. We synthesized 20 utterances using both voices and ran an A/B test. Each participant was presented with the utterance in both voices and they had to pick the utterence that they thought was more understandable. We had 6 native speakers of Marathi take the test. Figure 4 shows that the improvements we obtain out of the proposed method are indeed perceptually significant.

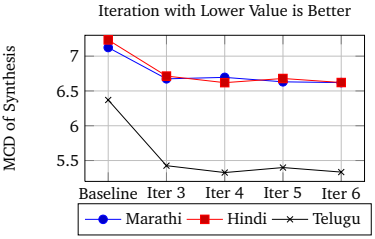

Figure 3: Objective Improvements using Iterative Targeted Acoustic Models

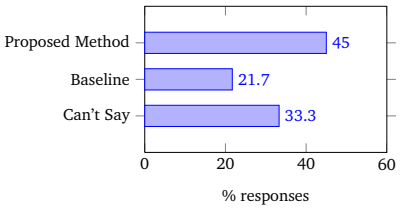

Figure 4: Subjective Preference between Synthetic Voices

5 Phonetic Writing System: Discussion

Our goal is to build synthetic voices for languages without an orthography. Our proposal is to automatically invent a phonetic writing system in that language, and then use it to build synthesis systems. We have used speech recognition techniques to devise these writing systems. This raises the questions of whether the artificially invented writing system is valid, or even useful.

In our experiments we used an English acoustic model for speech recognition. The output of the recognition decoder became the new writing system for the language at hand, Marathi. There are two main issues with using this writing system. (i) A smaller phone set (CMU-DICT) where Marathi actually has more phonemes. and (ii) Errors that speech recognition introduces in the phone strings. These two issues are explained below.

5.1 Effect of Phone-set Divergences on Synthesis Quality

The actual Marathi phone set is bigger than the CMU-DICT phone set that the English models have used. This leads to multiple Marathi phones getting mapped down to the same English phone. We looked at how using our ASR-based script for Marathi compares to using real Marathi text for building synthetic voices.

We built a standard Grapheme-based clustergen voice for Marathi. Each Unicode grapheme of the Devanagari alphabet was considered to be an independent phone. Because Devanagari is a phonetic alphabet, this voice provides us with an oracle data point: if we had an artificial language with a very good phone set, how good could our synthesis be?

Table 2 shows the comparison of the models we built in our work against the oracle voice. We observe that while our proposed method of acoustic model targeting gives good improvements over the baseline, there is a big gap in synthesis quality between using a CMU-DICT based writing system and the oracle writing system. This suggests that we should explore using more sophisticated acoustic models, such as those that use globalphone (Schultz and Waibel, 2001), or investigate phone splitting and phone joining techniques in future work.

Writing System	**MCD of Synthesis**
ASR-Based (CMU-DICT)	7.124
ASR-Based (CMU-DICT) (Targeted Acoustic Model)	6.620
Devanagari (Actual Marathi) (Oracle Result)	5.780

Table 2: Comparing ASR-based writing system to actual Marathi orthography

5.2 Effect of Errors Introduced by Speech Recognizers

Speech Recognition is often not perfect. Well-trained phonetic decoding can make mistakes when decoding speech in a language it was trained in. In our work, we are using a CMU-DICT based English acoustic model to decode a different language: Marathi. This discrepancy can introduce gross errors in the transcriptions generated. The writing system we invent is thus tainted.

We performed an oracle experiment to study the effect of noisy ASR on the quality of synthesis we ultimately achieve. We used the CMU-DICT as the phone set for our writing method. We used our own tool to map the original Indic script for Marathi into the CMU-DICT phone set. These transcriptions can be thought of as the output of a "perfect ASR" system. We then built an oracle voice using these transcripts and compared it to the voices built using automatically derived transcription language.

Table 3 shows the comparison of our models to the oracle voice. We observe that our baseline model is quite a bit weaker than the oracle voice. The targeted acoustic model helps build a voice that is better, but there is a good scope for future improvements in this direction. Good ways to detect noise introduced by ASR and methods to ignore the noise can potentially help bridge this gap and make synthesis even better.

Writing System	**MCD of Synthesis**
ASR-Based (CMU-DICT)	7.124
ASR-Based (CMU-DICT) (Targeted Acoustic Model)	6.620
Phonetized Devanagari (CMU-DICT) (Oracle Result)	6.006

Table 3: Effect of ASR noise on synthesis quality

5.3 Validity of the Phonetic Writing System

The automatically generated ASR-based writing system does generate understandable synthesis. It could thus be used as an intermediate language in speech to speech translation if the target language has no orthography. However, if we were building a dialog system in the target language, some person will have to write down text in the language as designed by ASR. No matter what phone set we use, ASR language can potentially be tainted by ASR errors. We need to measure the effort that a human would require in generating prompts in the artificial phonetic language. However, this is outside the scope of this paper.

6 Conclusions and Future Work

We have addressed a novel problem of building speech synthesizers for languages without an orthography. In our solution, we proposed automatically developing a writing system for the language, using a speech recognition system. Our iterative method to build targeted acoustic models yield very good improvements in synthesis quality. We showed objective and subjective results, as well as oracle results on Marathi, which show that our direction to building synthesis models without written text is promising. We also showed similar results on Hindi and Telugu, thus showing that our methods are language independent.

We have shown that the ASR-based writing system helps us build understandable synthesis. Two improvements we want to explore are: (i) using a large acoustic model trained on a larger phone set, or a universal phone recognizer such as (Siniscalchi et al., 2008), and (ii) Detecting noise in ASR transcript and mitigating the effects of that noise in synthesis output. We also plan to build speech translation systems for languages without orthography. The idea is to use the writing system we developed in this work and train statistical machine translation. We also plan to develop a written system for a real world language that has no orthography, and evaluate the user effort required in using the system to type real text.

References

Black, A. W. (2006). CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling. In *Proceedings of Interspeech*, pages 194–197, Pittsburgh, Pennsylvania.

Black, A. W. and Lenzo, K. (2002). Building voices in the festival speech synthesis system.

Black, A. W. and Taylor, P. (1997). The festival speech synthesis system: system documentation. Technical report, Human Communication Research Centre, University of Edinburgh.

Kominek, J. (2009). *TTS From Zero: Building Synthetic Voices for New Languages*. PhD thesis, Carnegie Mellon University.

Mashimo, M., Toda, T., Shikano, K., and Campbell, W. N. (2001). Evaluation of cross-language voice conversion based on GMM and straight. In *Proceedings of Eurospeech*, pages 361–364, Aalborg, Denmark.

Parlikar, A. and Black, A. W. (2012). Data-driven phrasing for speech synthesis in low-resource languages. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan.

Placeway, P., Chen, S. F., Eskenazi, M., Jain, U., Parikh, V., Raj, B., Mosur, R., Rosenfeld, R., Seymore, K., Siegler, M. A., Stern, R. M., and Thayer, E. (1996). The 1996 hub-4 sphinx-3 system. In *Proceedings of the DARPA Speech Recognition Workshop*.

Prahallad, K., Black, A. W., and Mosur, R. (2006). Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 853–856, Toulouse, France.

Prahallad, K., Kumar, E. N., Keri, V., Rajendran, S., and Black, A. W. (2012). The iiit-h indic speech databases. In *Proceedings of Interspeech*, Portland, OR, USA.

Schultz, T. and Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(1):31–51.

Siniscalchi, S. M., Svendsen, T., and Lee, C.-H. (2008). Toward a detector-based universal phone recognizer. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, USA.

Stueker, S. and Waibel, A. (2008). Towards human translations guided language discovery for asr systems. In *Proceedings of Spoken Language Technologies for Under-Resourced Languages*.