Text to Speech in New Languages without a Standardized Orthography

Sunayana Sitaram, Gopala Krishna Anumanchipalli, Justin Chiu, Alok Parlikar and Alan W Black

Language Technologies Institute, Carnegie Mellon University
{ssitaram, gopalakr, jchiu1, aup, awb}@cs.cmu.edu

Abstract

Many spoken languages do not have a standardized writing system. Building text to speech voices for them, without accurate transcripts of speech data is difficult. Our language independent method to bootstrap synthetic voices using only speech data relies upon cross-lingual phonetic decoding of speech. In this paper, we describe novel additions to our bootstrapping method. We present results on eight different languages---English, Dari, Pashto, Iraqi, Thai, Konkani, Inupiaq and Ojibwe, from different language families and show that our phonetic voices can be made understandable with as little as an hour of speech data that never had transcriptions, and without many resources in the target language available. We also present purely acoustic techniques that can help induce syllable and word level information that can further improve the intelligibility of these voices.

Index Terms: speech synthesis, synthesis without text, languages without an orthography

1. Introduction

Recent developments in speech and language technologies have revolutionized the ways in which we access information. Advances in speech recognition, speech synthesis and dialog modeling have brought out interactive agents that people can talk to naturally and ask for information. There is a lot of interest in building such systems especially in multilingual environments. Building speech and language systems typically requires significant amounts of data and linguistic resources. For many spoken languages of the world, finding large corpora or linguistic resources is difficult. Yet, these languages have many native speakers around the world and it would be very interesting to deploy speech technologies in them.

Our work is about building text-to-speech systems for languages that are purely spoken languages: they do not have a standardized writing system. These languages could be mainstream languages such as Konkani (a western Indian language with over 8 million speakers), or dialects of a major language that are phonetically quite distinct from the closest major language. Building a TTS system usually requires training data consisting of a speech corpus with corresponding transcripts. However, for these languages that aren't written down in a standard manner, one can only find speech corpora. Our current efforts focus on building speech synthesis systems when our training data doesn't contain text.

It may seem futile to build a TTS system when the language at hand doesn't have a text form. Indeed, if there is no text at training time, there won't be text at test time, and then one might wonder why we need a TTS system at all. However, consider the use case of deploying a speech-to-speech translation of video lectures from English into Konkani. We have to synthesize speech in this "un-written" language from the output of a machine translation system. Even if the language at hand may not have a text form, we need some intermediate representation that can act as a text form that the machine translation system can produce. A first approximation of such a form is phonetic strings. Another use case for which we need TTS without text is, say, deploying a bus information system in Konkani. Our dialog system could have information about when the next bus is, but it has to generate speech to deliver this information. Again, one can imagine using a phonetic form to represent the speech to be generated, and produce a string of phones from the natural language generation model in the bus information dialog system.

The work we present here is our continued effort in improving text to speech for languages that do not have a standardized orthography. We have built voices for several languages, from purely speech corpora, and produced understandable synthesis. We use cross-lingual phonetic speech recognition methods to do so. Phone strings are not ideal for TTS, however, as a lot of information is contained in higher level phonological units including the syllables and words that can help produce natural prosody. However, detecting words from speech corpus alone is a difficult task. We have explored how purely acoustic techniques can be used to detect word like units in our training speech corpus and use this to further improve the intelligibility of speech synthesis.

2. Relation to prior work

Speech to speech translation typically involves a cascade of three models: an automatic speech recognition system (ASR) in the source language, a statistical machine translation system (SMT), and a text to speech engine (TTS) in the target language. Generally, these three models are developed independently of each other. Recent work such as [1, 2, 3, 4] has looked into deeper integration of this pipeline, but the general assumption here is that the target language has an orthography.

If the target language of speech to speech translation does not have a written form, it has been proposed that one be defined, though training people to use it consistently is in itself very hard and prone to inconsistencies (e.g. Iraqi Arabic transcription techniques in the recent TRANSTAC Speech to Speech Translation Project, see [5]). Our proposal is to use a phonetic-like representation of the target speech, derived acoustically as the orthography to use. [5, 6] have investigated such an approach.

Changes have been proposed to SMT modeling methods [7, 8] to specifically deal with phoneme strings in the target language. In order to induce the automatic phonetic writing form, we use an ASR system in a foreign language and

adapt the acoustic model to match the target speech corpus. Speech synthesis voices are typically built from less data compared to speech recognition systems. Acoustic model adaptation with limited resources can be challenging [9]. [10] has recently proposed a rapid acoustic model adaptation technique using cross-lingual bootstrapping that showed improvements in the ASR of under-resourced languages. Our model adaptation technique is somewhat similar to that method, but we optimize the adaptation towards better speech synthesis, and have only acoustic data in the target language.

In preliminary work in this direction [11] we proposed a method to devise a writing system. We also proposed using existing techniques to automatically induce words and syllables from a string of phonemes [12]. In this work, we propose using acoustic information to derive higher level phonological units, which is language independent and more reliable than inducing structures using noisy ASR output.

Although such representations may be difficult for a native speaker to write, an SMT system can help bridge the gap from a source language to the target phonetic representation of the language. [13] models pronunciation variability based on articulatory features and is more suited for our purpose (since ASR transcript could be noisy) and we plan to use such models in the future.

3. Data and resources

We used audio from eight languages from four diverse language families for this research. Our audio data ranged from almost two hours of speech to less than six minutes, as shown in Table 1.

Language	Size (minutes)
English	111
Dari	52
Iraqi	62
Pashto	39
Thai	25
Ojibwe	12
Inupiaq	5.5
Konkani	5.5

Table 1: Audio data sizes

Our English data was from the Blizzard Challenge [14] 2013 audio book task, recorded by a professional voice recording artist.

Dari is a dialect of Persian that is used in Afghanistan as an official language and also spoken in parts of Iran and Tajikistan. It has over 18 million native speakers. Pashto is a also an official language of Afghanistan and has over 40 million speakers. The Dari and Pashto corpora are from the DARPA TRANSTAC project. Iraqi Arabic is a dialect of Arabic spoken in Iraq and has about 15 million speakers. The Iraqi Arabic corpora were provided by BBN as part of the DARPA BOLT project.

The Thai language is spoken by over 20 million people and is the official language of Thailand. We used the Thai speech corpora from the SPICE [15] dataset.

Inupiaq is an Inuit language spoken by about 2100 people in northern and northwestern Alaska. Ojibwe is spoken in Canada and the United States and has around 56000 native speakers. Both Inupiaq and Ojibwe use the Latin script in their written forms. Our data for Inupiaq and Ojibwe came from a corpus collected as part of the Endangered Languages project at Carnegie Mellon University.

Konkani is an official language of India and is used primarily in Goa and Karnataka. It has over 8 million native speakers. Konkani does not have its own script, and native Konkani speakers use Devanagari, Latin, Kannada, Malayalam and even Arabic scripts to write it. We used a corpus of Konkani from the CMU SPICE project [15].

We used the CMU Sphinx [16] speech recognition toolkit in allphone mode as our phonetic decoder and to train new acoustic models. We used the Festvox voice building tools to build CLUSTERGEN [17] voices for the Festival [18] speech synthesizer. CLUSTERGEN is a type of statistical parametric synthesizer that is more robust to noise than other methods such as unit selection. Our method can be used with any parametric synthesis technique.

Our phonetic decoder used trigram phonetic language models built from German and Marathi data. For the German language model, we used the Europarl [19] corpus and for the Marathi language model, we used a corpus created by collecting news stories from a Marathi news website, Esakal. The words are expanded to their phonetic forms using statistically trained letter to sound rules in the respective language. We used a single acoustic model, the Wall Street Journal (WSJ) English acoustic model provided with CMU Sphinx.

We used the TestVox[20] tool to run listening tests online.

4. Overview of our approach

Figure 1 shows a block diagram of the components and flow of our approach.

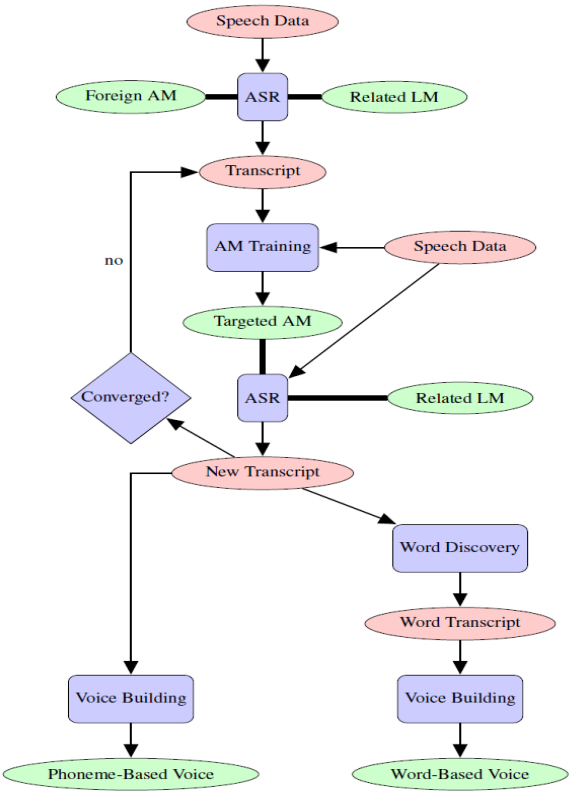

Figure 1: Overview of our approach

First, we decode the audio in the target language with a phonetic decoder using an acoustic model and language model from another language. Then, using the transcripts obtained from decoding and the speech corpus, we iteratively build new targeted acoustic models until convergence. We use the phonetic transcripts to build synthetic voices and evaluate them objectively using the Mel-Cepstral Distance (MCD) [21] and subjectively using human listening tasks. We also automatically induce syllable and word-like structures on these transcripts and build syllable and word based synthetic voices. The next few sections describe these steps in greater detail.

5. Bootstrapping synthetic voices

For phonetic decoding, we use an acoustic model and phoneme language model from a related language. For our experiments in this paper, we used the English WSJ acoustic model for decoding all languages. Using the WSJ acoustic model for decoding English speech is not fair, but we used it to keep the acoustic model consistent in all our experiments. Ideally for phonetic decoding, an acoustic model and phoneme language model from a closely related language are more appropriate. To simulate this in our experiments, we used Marathi and German phonetic language models, as listed in Table 2.

Language	Acoustic Model	Language Model
English	WSJ	German
Dari	WSJ	Marathi
Iraqi	WSJ	German
Pashto	WSJ	Marathi
Thai	WSJ	Marathi
Ojibwe	WSJ	German
Inupiaq	WSJ	German
Konkani	WSJ	Marathi

Table 2: Acoustic and Language models used for cross lingual decoding

After decoding speech in the target language using the appropriate acoustic and language models, we iteratively train new acoustic models using the decoded transcript as the text and the original audio as the speech. At each stage of the iterative process, we calculate the MCD of the voice built using the decoded transcript from that iteration. Figure 3, 4 and 5 show the MCD of voices built using these transcripts for various languages.

Figure 2 shows the MCDs of transcripts obtained on English, Dari and Iraqi Arabic. English has about two hours of speech while Dari and Iraqi Arabic have about one hour of speech. We see that there is a big drop in MCD value from the first iteration to the second, in which the targeted acoustic model is built. In the case of English, iteration 7 has the lowest MCD, after which it rises slightly. For Iraqi Arabic and Dari, the MCD continues to fall until the last iteration.

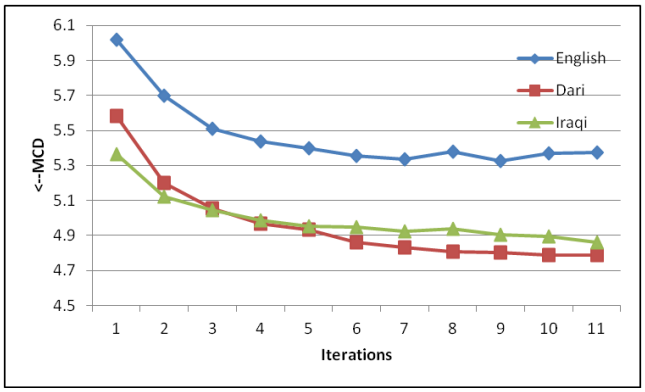

Figure 2: Iterative targeted Acoustic Models for languages with ~1 hour of speech

Figure 3 shows the MCD graph for Pashto and Thai, both of which have around 30 minutes of speech. We see that the MCD for Pashto in the first three iterations falls rapidly and then does not change much, while for Thai, there is a big drop after the first iteration, which is consistent with the results for English, Dari and Iraqi Arabic. There is a large rise in MCD at iteration 7 for Thai, but it falls again in the next iteration. We can see that even with half an hour of speech, our iterative method produces better transcripts than the base decoding with the WSJ acoustic model.

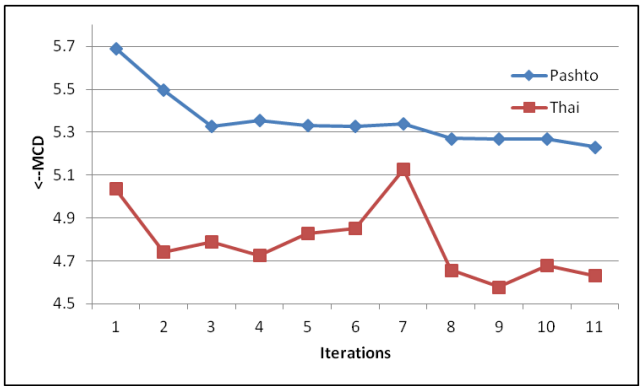

Figure 3: Iterative targeted Acoustic Models for languages with ~30 minutes of speech

Figure 4 shows results for Ojibwe, which has 12 minutes of speech and Inupiaq and Konkani, both of which have around five minutes of speech. We see that for Ojibwe, the MCD rises slightly after the first iteration and then falls after the fifth iteration, with the difference in the MCD between the base and best iteration being 1.43. This shows that even with just 12 minutes of speech, the iterative method is able to come up with a better transcript than just the base decoding. However, for both Inupiaq and Konkani, we see that the MCD rises after the first iteration. This is probably because of the phonetic complexity of these languages and the amount of speech is too small to build even targeted acoustic models.

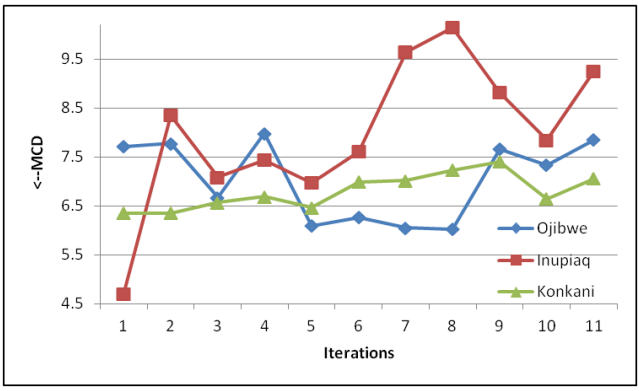

Figure 4: Iterative targeted Acoustic Models for languages with < 15 minutes of speech

Overall, we see that with a reasonable amount of speech data, the iterative targeted acoustic models produce better phoneme transcripts than just using base decoding from a cross lingual phonetic decoder, shown here on a variety of languages.

Throughout the iterations, we kept the language model used by the ASR consistent. One obvious extension of this approach is to adapt the language model at each iteration. However, preliminary experiments on interpolating the original language model at each iteration with the new transcript did not yield improvements in MCD.

6. Improved synthesis with syllables and words

So far, we have discussed the bootstrapping method which produces phoneme transcripts of the audio, which may be noisy. However, Text to Speech systems typically benefit from using syllable and word level information. So, we try to automatically induce syllables and word-like units from the phoneme transcripts.

We syllabified English and Dari transcripts with the lowest MCD in the 10 iterations. To obtain syllables, we use heuristic rules built into the Festival speech synthesizer to join phonemes in the transcripts. We treated the syllables as words and added appropriate entries in the lexicon.

For inducing word-like units, we used cross-lingual information to train a Conditional Random Field (CRF) model. We created training data for the CRF by extracting phonemes and word boundaries from the German Europarl data. We used CRF++ [22] to train a German model that could group phoneme sequences into world-like units and used the same English and Dari transcripts used for syllabification earlier to test the model. We discarded words that were rare (< 300 in frequency) and used the rest of the hypothesized words in our transcripts. We added appropriate lexical entries for these words and built voices for English and Dari.

Table 3 shows the result of syllable and word induction. We see that both for English and Dari, grouping phonemes into syllables decreases the MCD of the new voice. Surprisingly, this difference is very large in the case of Dari, even though the syllable rules were written for English. The voice built for English using CRF word induction has a slightly lower MCD than the syllable method. However, this method does not seem to make much of a difference in the case of Dari. This could be because we used a German word model, and German word rules are quite different from Dari.

Language	Best iteration	Syllables	CRF
English	5.328	5.26	5.25
Dari	4.787	4.165	4.76

Table 3: MCD comparison of voices with syllable and word induction

7. Inducing higher level phonological units

In Section 6, we have demonstrated how syllables and words can be induced from the raw phonetic transcriptions. Here we present an approach that uses the acoustic information, to derive higher order phonological units. While the approach for deriving syllables from phone strings is fairly straight forward across languages (grouping phones together, with the constraint of having one vowel per syllable), the derivation of words from phones is not one-to-one and there is little generalization that is language independent. There is additional complexities for languages with no formal notions of word, or those that are morphologically agglutinative. Here we propose derivation of a more reliable and generalizable phonological unit, the accent group.

We use the broad definition of accent group as being a group of syllables that bears only one intonational accent (a.k.a pitch accent) on them. This definition, while appealing to the idea of metrical feet, does not use pre-defined rules on which syllables should be grouped together, instead opting for a completely data-driven parsing approach (the complete description and training strategy are provided in [24]) as summarized below.

The idea is to analyze the pitch contour in tandem with the underlying syllable sequence and approximate it with a synthetic contour described as a sequence of TILT shapes [23] over parses of syllable groups. The optimal parse on the syllables is one that minimizes the reconstruction error of the target pitch contour. A stochastic context free grammar is trained on such parses of accent groups, so as to allow prediction of accent groups for unseen sequences of syllables. In order to uniquely identify syllables, we tag each syllable with the vowel name, the onset and coda categories as described in [24] (e.g: *syl_onsettype_codatype_vowel*). These categories are only a few in number and yet are language independent, allowing us to use this approach for arbitrary new languages here. This is illustrated in Figure 5 below.

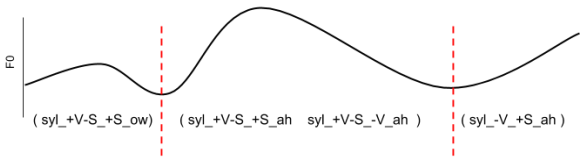

Figure 5: Acoustically derived parse over syllables into accent groups

Given such parses derived acoustically from the pitch contours on all of training data, a grammar is trained to predict parses of unseen sequences of tagged syllables. This is further improved with decision trees about the positional information of each syllable, so as to reliably estimate for each syllable boundary, if there is a accent group boundary, or not.

8. Subjective evaluation for intelligibility

From our objective results mentioned earlier, we saw that the voices built using syllables were better than the voice built on the best iteration using phonemes. Word induction seemed to help in the case of English, but not Dari.

To test this subjectively, we conducted A/B listening tests comparing the voice having the lowest MCD and the voice with syllable units for both English and Dari to see if grouping phonemes together into syllables was perceptually better. Table 5 lists the results from tests on English and Dari. In both cases, we see that participants preferred the voice with syllabified transcripts significantly more than the best iteration.

Language	# participants	Best iteration	Syllable	Can't say
English	7	4%	68%	28%
Dari	5	6%	72%	22%

Table 5: Results of listening tests for English and Dari

In order to test the higher level phonological units i.e. the words and Accent Groups, we built synthetic voices as described in the previous sections. Unseen sequences of sentences in syllabified English were synthesized by each of the above methods i) Syllable transcripts, ii) CRF induced words and iii) Automatically detected Accent Groups. In addition, we also compared the system with Accent Groups to the best iteration from our baseline approach which has no induction of higher level units. 10 sentences of each system were compared in pairs by 10 English listeners who were asked to make a preference to one of the two stimuli. The results are shown in the Table 6.

Voice A	Voice B	Prefer A	Prefer B	Can't say
Best iteration	Accent Group	12%	78%	10%
Induced words (CRF)	Accent Group	22%	70%	8%
Syllable	Accent Group	47%	43%	10%

Table 6: Results of listening tests for English

The results indicate that significant gains can be obtained by induction of the speech-derived accent group units, as opposed to word derivations through CRFs over phoneme transcriptions. While it is encouraging that the Accent Group voices perform comparably, syllable voices remain the most reliable units that can be induced in the current setting. This is perhaps due to the unavailability of sufficient data, or features that effectively capture the contextual information in building voices using higher levels of phonology.

9. Conclusion and future work

In this paper, we applied our iterative cross-lingual decoding technique to eight languages from various language families. We saw that with as little as half an hour of speech, we could get improvements in MCD over the baseline decoded transcripts.

We also used heuristics to syllabify the phoneme transcripts and a CRF to automatically induce word like units, which led to higher quality voices, both objectively and subjectively. In addition, we described a method to use acoustic information to identify accent groups to create higher level phonological units which may help improve the quality of synthesis. Our results indicate that inducing such units leads to a large improvement in both MCD and subjective preference.

From our initial experiments on building SMT systems from the source language to the target learned transcript, knowing where the word boundaries are can prove to be critical for good translation. We plan to explore other methods to automatically derive higher level units from text and acoustics.

In the future, we also plan to explore using combinations of multiple acoustic and language models instead of relying on a single model for the initial decoding pass. We also realize the importance of the initial phoneset and plan to explore more principled methods of pruning phonesets at each iteration. The next stage of this work is to extend it to use machine translation to provide a usable writing system for languages without a standardized orthography.

Acknowledgment

This research was supported in part by a Google Research Award "Text-to-Speech in New Languages without the Text"

References

[1] Bowen Zhou, Laurent Besacier, and Yuqing Gao, "On Efficient Coupling of ASR and SMT for Speech Translation," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, HI, USA, April 2007, vol. 4, pp. 101–104.

[2] Nicola Bertoldi, Richard Zens, and Marcello Federico, "Speech Translation by Confusion Network Decoding," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, HI, USA, April 2007, vol. 4, pp. 1297–1300.

[3] Pablo Daniel Agüero, Jordi Adell, and Antonio Bonafonte, "Prosody Generation for Speech-to-Speech Translation," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, May 2006.

[4] Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth S. Narayanan, "Factored Translation Models for enriching Spoken Language Translation with Prosody," in Proceedings of Interspeech, Brisbane, Australia, September 2008, pp. 2723–2726.

[5] Laurent Besacier, Bowen Zhou, and Yuqing Gao, "Towards Speech Translation of non Written Languages," in Proceedings of the IEEE Workshop on Spoken Language Technology, Palm Beach, Aruba, December 2006, pp. 222–225.

[6] Sebastian Stüker and Alex Waibel, "Towards Human Translations Guided Language Discovery for ASR Systems," in Proceedings of Spoken Language Technologies for UnderResourced Languages, 2008.

[7] Zeeshan Ahmed, Jie Jiang, Julie Carson-Berndsen, Peter Cahill, and Andy Way, "Hierarchical Phrase-Based MT for Phonetic Representation-Based Speech Translation," in Proceedings of the tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA), San Diego, CA, October 2012.

[9] Felix Stahlberg, Tim Schlippe, Stephan Vogel, and Tanja Schultz, "Word Segmentation Through Cross-Lingual Word to-Phoneme

Alignment," in Proceedings of IEEE Workshop on Spoken Language Technology, Miami, FL, December 2012.

[10] George Zavaliagkos and Thomas Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," in Proceedings of The DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[11] Sukhada Palkar, Alan W Black, and Alok Parlikar, "Text-to Speech for Languages without an Orthography," in Proceedings of the 24th International conference on Computational Linguistics, Mumbai, India, December 2012.

[12] Sunayana Sitaram, Sukhada Palkar, Alok Parlikar, and Alan W Black, "Bootstrapping Text-to-Speech for Speech Processing in Languages without an Orthography" in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, May 2013.

[13] Micha Elsner, Sharon Goldwater, and Jacob Eisenstein, "Bootstrapping a Unified Model of Lexical and Phonetic Acquisition," in Proceedings of Association for Computational Linguistics, Jeju island, Korea, July 2012.

[14] Alan W Black and Keiichi Tokuda, "Blizzard Challenge -- 2005: Evaluating corpus-based speech synthesis on common datasets" Interspeech, Lisbon, Portugal, 2005.

[15] T. Schultz, A. W. Black, S. Badaskar, M. Hornyak, and J. Kominek, "Spice: Web-based tools for rapid language adaptation in speech," in Proceedings of INTERSPEECH, Antwerp, Belgium, August 2007.

[16] Paul Placeway, Stanley F. Chen, Maxine Eskenazi, Uday Jain, Vipul Parikh, Bhiksha Raj, Ravishankhar Mosur, Roni Rosenfeld, Kristie Seymore, Matthew A. Siegler, Richard M. Stern, and Eric Thayer, "The 1996 Hub-4 Sphinx-3 System," in Proceedings of the DARPA Speech Recognition Workshop, 1996

[17] Alan W Black and Paul Taylor, "The Festival Speech Synthesis System: system documentation," Tech. Rep., Human Communication Research Centre, University of Edinburgh, January 1997.

[18] Alan W Black, "CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling," in Proceedings of Interspeech, Pittsburgh, Pennsylvania, September 2006, pp. 194–197.

[19] Philipp Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in Proceedings of Machine Translation Summit, Phuket, Thailand, September 2005, pp. 79–86.

[20] Alok Parlikar, "TestVox: Web-based Framework for Subjective Evaluation of Speech Synthesis," Opensource Software, 2012.

[21] Mikiko Mashimo, Tomoki Toda, Kiyohiro Shikano, and Wilhelm Nicholas Campbell, "Evaluation of Cross-Language Voice Conversion Based on GMM and Straight," in Proceedings of Eurospeech, Aalborg, Denmark, September 2001, pp. 361–364.

[22] Taku Kudoh, "Crf++", Software, http://crfpp.sourceforge.net/, 2007.

[23] P Taylor, "Analysis and synthesis of intonation using the tiltmodel," Journal of the Acoustical Society of America, vol. 1073, pp. 1697–1714, 2000.

[24] Gopala Krishna Anumanchipalli, Luis C Oliveira, Alan W Black, " Accent Group Modeling for Improved Prosody in Statistical Parametric Speech Synthesis", in proceedings of IEEE ICASSP 2013, Vancouver, Canada, 2013.