BOOTSTRAPPING TEXT-TO-SPEECH FOR SPEECH PROCESSING IN LANGUAGES WITHOUT AN ORTHOGRAPHY

Sunayana Sitaram *Sukhada Palkar* *Yun-Nung Chen* *Alok Parlikar* *Alan W Black*

Language Technologies Institute, Carnegie Mellon University, USA
ssitaram, spalkar, yvchen, aup, awb @cs.cmu.edu

ABSTRACT

Speech synthesis technology has reached the stage where given a well-designed corpus of audio and accurate transcription an at least understandable synthesizer can be built without necessarily resorting to new innovations. However many languages do not have a well-defined writing system but such languages could still greatly benefit from speech systems. In this paper we consider the case where we have a (potentially large) single speaker database but have no transcriptions and no standardized way to write transcriptions. To address this scenario we propose a method that allows us to bootstrap synthetic voices purely from speech data. We use a novel combination of automatic speech recognition and automatic word segmentation for the bootstrapping. Our experimental results on speech corpora in two languages, English and German, show that synthetic voices that are built using this method are close to understandable. Our method is language-independent and can thus be used to build synthetic voices from a speech corpus in any new language.

Index Terms— Speech Synthesis, Synthesis without Text, Languages without an Orthography

1. INTRODUCTION

Most languages do not have a standardized writing system, yet are spoken by many people. If speech technology is to make an impact for all languages it will need to consider processing of languages without a standardized orthography. Even though there may be no standardized orthography, at least some speakers of such languages are often literate in some other languages, perhaps a former colonial language such as English or Spanish, or a nationwide language such as Hindi or Mandarin. This paper joins the growing area of investigation of speech processing for unwritten languages.

We are specifically addressing the issue of generating speech in languages without using a standardized written form for that language. We expect to be able to collect acoustics in that language and be able to know the meaning of what is said. We envisage a data collection method where, say a bi-lingual Konkani[1] speaker is given a written prompt in Hindi and they speak the sentence in Konkani with the same meaning as the Hindi prompt. Our goal is to produce a system utilizing acoustic analysis, statistical parametric speech synthesis and machine translation technologies that, given such data, will allow new prompts to be written in Hindi that will produce Konkani speech output of the same meaning. Thus allowing spoken dialog systems, information giving systems, etc. to be easily developed without having to create, teach and enforce a new writing system.

While this work is generally set in the context of speech to speech translation, this paper is specifically about building the text-to-speech component. We assume that we only have a speech corpus, with no transcriptions, from which we have to build a synthetic voice. We use an automatic speech recognition system using an acoustic model from some other language to perform phonetic decoding of the speech data to get an initial transcription of the speech. We then iteratively improve the acoustic model by re-training with the speech corpus at hand. The end result is a phonetic-like transcription of the speech corpus. We use this to build our synthetic voice.

This paper tests two languages (English and German) that actually do have writing systems, but we test how well we can produce speech output for these languages assuming we only have recorded prompts in these languages and no written form. We primarily use speech synthesis output quality to gauge our success in modeling.

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 inconsitencies (e.g. Iraqi Arabic transcription techniques in the recent Transtac Speech to Speech Translation Project, see [5]). Our proposal is to use use a phonetic-like representation of the target speech, derived acoustically as the orthography to use. In our preliminary work [6], we suggested a method to devise such an automatic writing system and we build upon this method in this work.

While the automatic orthography we propose may be difficult to write for native speakers of the target language, an SMT system can help bridge the gap by translating from the source language into this phonetic target language. [5, 7] have investigated such an approach. Changes have been proposed to SMT modeling methods [8, 9] 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 fewer data compared to speech recognition systems. Acoustic model adaptation with limited resources can be challenging[10]. [11] have proposed using a speech recognizer trained without supervision, for tasks such as topic classification. [12] 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

[1] An Indo-Aryan language spoken on the western coast of India

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.

Languages are not simply sounds: there are words and sentences. Typical speech synthesizers are trained on such higher level structures rather than simply phonemes. In this work, we have used existing techniques[13, 14] to induce words from phonemes. [15] 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 this model in the future. We also used an automatic part-of-speech induction technique [16] over the generated words to use as features in our synthesizer.

3. RESOURCES AND DATA

We used the Festival[17] speech synthesizer for this research. We used Festvox to build CLUSTERGEN[18] voices for synthesis in the target language. CLUSTERGEN is a statistical parametric synthesis system, a form of synthesis that is particularly suited to dealing with noisy data. A parametric synthesizer is more robust to noise than other synthesis methods, such as unit selection. Our method doesn't particularly depend on CLUSTERGEN, any other parametric synthesis technique can also be used. We used the Sphinx-3[19] system as our phonetic decoder and also to train new acoustic models.

A realistic data set for us to use would be some parallel speech corpus (say, English utterances and their Hindi equivalent speech). We also require the data to be of modest size, so that an SMT system can be trained on it. We do not have such a parallel speech corpus. We have started collecting such a corpus. For work reported in this paper however, we simulated the presence of such a corpus. We started with the BTEC corpus[20], which is a Text–Text parallel corpus between English and German (and other languages). We then used a speech synthesizer to generate speech for the corresponding utterances. We then pretend the text transcription never existed, and use only the generated speech in our experiments. We realize that having natural speech is better, but our goal was to study the feasibility and success of our methods before carrying out an expensive corpus elicitation task.

We applied our methods to two languages: English and German. For both languages, we selected speech corpora of two sizes: (a) 1000 utterances, and (b) 5000 utterances. We also chose two acoustic models in each language to start bootstrapping from. One of the acoustic models is the Wall-Street-Journal (WSJ) English acoustic model, provided with CMU Sphinx, trained from [21]. The other is an acoustic model built on a combined corpus of Indic languages taken from the IIIT-H data set[22]. The WSJ acoustic model uses the CMU-DICT phone set, whereas the Indic acoustic model uses a subset of the Sampa phoneset.

Our phonetic decoder uses a tri-gram language model built from text in a related language. In this case, we used the Europarl[23] corpus and built phonetic language models from English and German data. When decoding English speech, we used the German Phonetic Language model, and when decoding German speech, we used the English Phonetic Language model. Before building the language models, we mapped the phonemes into the phoneset appropriate for the acoustic models.

We used the TestVox[24] tool to run listening tests online.

4. OVERVIEW OF OUR APPROACH

Our proposed method basically takes speech data and produces a transcription of it using automatic speech recognition. The process is explained briefly in this section. Figure 1 shows a block diagram of the various components and the process flow.

We first decode (cross-lingually) our speech corpus with a phonetic decoder. We then use iterative decoding to build a new targeted acoustic model using our target speech corpus. After convergence, we use the phonetic transcription obtained to build our synthetic voice. We also automatically induce word like structures over the phonetic sequences and build another synthetic voice. We evaluate these phonetic and wordified voices objectively using the Mel-Cepstral Distance[25] (MCD) and subjectively using human listening tasks. These steps are described in detail in the next sections.

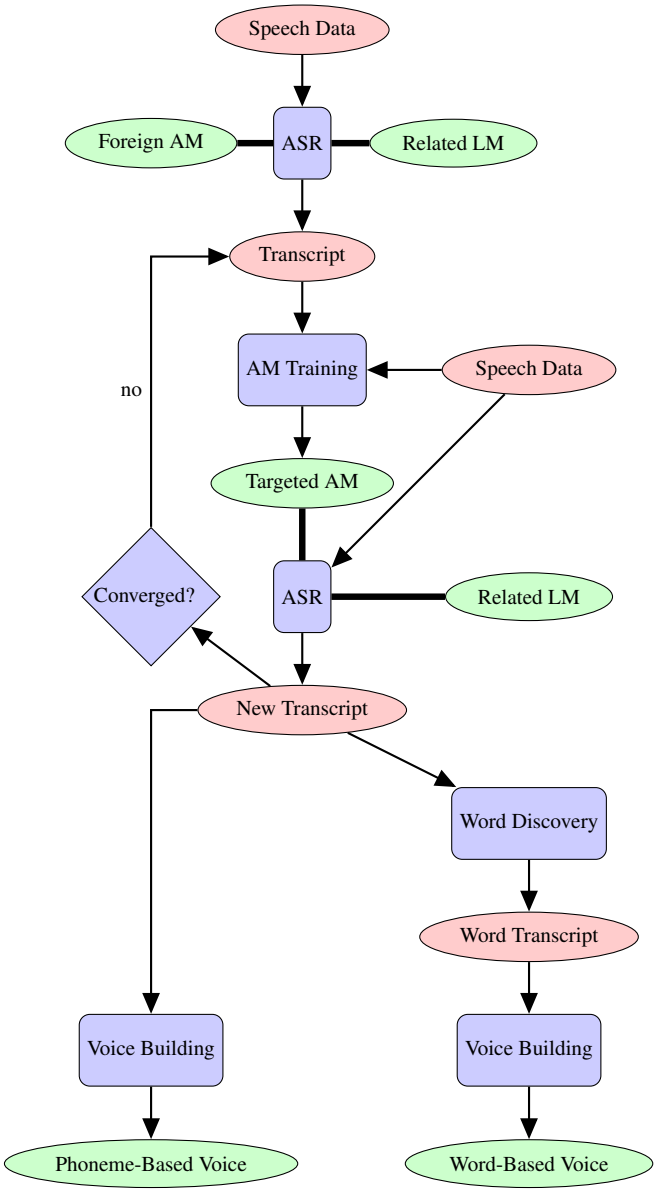

Fig. 1. Overview of our Approach

5. BOOTSTRAPPING PHONETIC TRANSCRIPTIONS

To bootstrap phonetic transcriptions, we take our speech corpus and cross-lingually decode it using a phonetic ASR system. We use an

acoustic model built from data in a foreign language. The language model is built using corpus from a phonetically related language. For new languages, we can choose a related language from the same family.

In the case of our German speech, we used two acoustic models: (i) The WSJ acoustic model in English, and (ii) an Indic acoustic model. For English, we initially used (i) The Indic acoustic model, and (ii) An acoustic model built on the Globalphone data[26]. However, because of a gender-mismatch between the particular subset of the Globalphone data that we have, and our target speaker, we could not get reliable results with the Globalphone acoustic model and fell back up on the WSJ acoustic model. Using the WSJ acoustic model to decode English speech is not realistic, but we present the results here for comparison with how well the English decoding performs relative to the Indic model.

We built four language models. For German, we used the Europarl English data and built phonetic language models using (i) The WSJ phone set, and (ii) The Indic phone set. For English, we used the Europarl German data and built two language models for the two phonesets.

We first phonetically decode our speech corpus using one of these appropriate models. After we obtain the initial transcription, we use this transcription and the corresponding speech as parallel data to train a new acoustic model. This acoustic model gets built on only the speech corpus we have. Using this new acoustic model, and the same language model we used before, we decode our speech corpus again. This results in another transcription of the speech. We rebuild another acoustic model, repeat the decoding process and iterate. At each step, we build a synthetic voice and measure the quality of synthesis in terms of the MCD score on a helded out test set. The iteration that gets the lowest MCD is deemed to be the best iteration. The transcription obtained from this iteration is deemed to determine the automatic orthography for the language in the speech corpus.

Figure 2 shows the results of our iterative process for German. We can make the following observations: (i) The quality of synthesis improves substantially over the iterations. (ii) Results are better when we use larger data (5000 utterances versus 1000 utterances), and (iii) Using the WSJ acoustic model gives better voices, compared to using the Indic model.

The WSJ acoustic model yields better transcriptions for German. While one could think that the English phoneset is better for German than the Indic one, this may not be the only reason behind the results we obtained. When we look at the number of phonemes that are found in the transcript, we see that the Indic transcripts have fewer phoneme types (31) than the WSJ transcript (39). Having more phoneme types can help capture the variability in speech better, resulting in better synthesis.

Figure 3 shows the results of our iterative process for English. We can make the following observations: (i) The quality of synthesis after the iterative process can be substantially better than with the initial labeling, (ii) Using larger speech corpus yields better transcriptions, and (iii) The WSJ model performs better than the Indic model for larger data, yet the Indic model beats the WSJ acoustic model for smaller data.

WSJ is an English acoustic model, and the WSJ phoneset is thus theoretically best suited for English. The results we obtained on the 1000 utterance data set are slightly puzzling. Transcriptions with the Indic acoustic models are better than those with WSJ models, even if the number of phone types used in the Indic transcriptions (26) was lower than the WSJ case (31).

We also initially attempted to include Mandarin Chinese in our tests and found similar trends. The biggest improvement in MCD

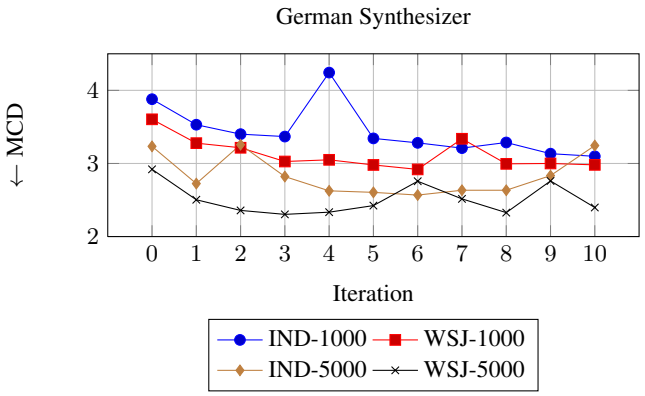

Fig. 2. Iterative Targeted Acoustic Models for German using Different Phonesets and Data Sizes

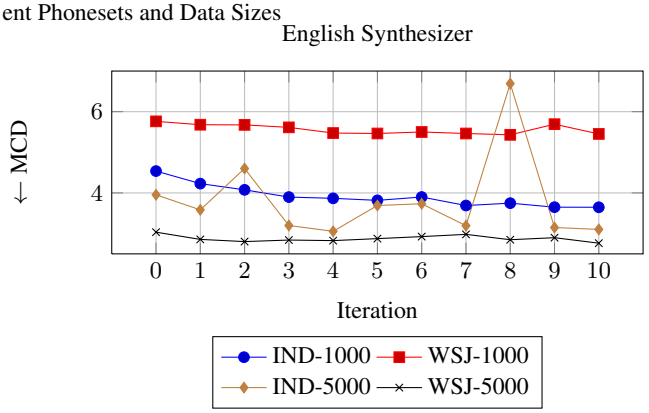

Fig. 3. Iterative Targeted Acoustic Models for English using Different Phonesets and Data Sizes

comes with the first iteration that uses the target language data to build the acoustic model. In all cases the number of phones at each stage remained the same to decreased by one.

6. IMPROVED SYNTHESIS WITH WORDS AND SYLLABLES

The proposed bootstrapping method produces transcriptions that are only phonetic strings. Text to speech systems typically produce better results when syllable or word level information is available. In addition, the automatic transcription has some noise in it. To study the impact of these two issues on synthesis quality, we ran an oracle experiment in both languages. We used actual transcriptions at word level and built a voice. We then converted these true transcriptions into phonetic strings, to get true phone sequences. Table 1 shows this comparison. We can clearly see the gap between word based synthesis and phoneme based synthesis. In addition, the gap between using true phonemes and the phonemes our method generates is also substantial. Thus, there is great scope for improving these models further.

We ran experiments to see whether word level information can help improve the quality of synthesis, starting from the best automatic transcription that we have. We used two different methods to group together phonemes before synthesis. For the German voice, we used the best transcriptions (WSJ-5000-iter3) and for the English voice, we used the best Indic transcriptions (IND-5000-iter4) to run word induction on.

Language	System	MCD
English	Best IND-1000 Iter	3.647
	Best IND-5000 Iter	3.055
	Best WSJ-1000 Iter	5.447
	Best WSJ-5000 Iter	**2.802**
	Oracle Phones	2.465
	Oracle Words	2.157
German	Best IND-1000 Iter	3.098
	Best IND-5000 Iter	2.568
	Best WSJ-1000 Iter	2.919
	Best WSJ-5000 Iter	**2.304**
	Oracle Phones	2.120
	Oracle Words	1.753

Table 1. Comparison with Oracle Synthesis Quality. Our best system is shown in bold.

In our first method, we used heuristic rules to clump together groups of phonemes into syllables. Festival has a built in syllabifier that works with the WSJ phoneset. We tweaked it to also work with the Indic phoneset. We marked individual syllables as "words" in the transcript, and added appropriate lexical entries. We also ran automatic Part-of-Speech induction[16] on these super-phonetic units and added them into Festival.

For the second method, we used cross-lingual information for inducing word like units. For German synthesis, we started with the English Europarl data and extracted phonetic information along with word boundaries from it. We then trained a Conditional Random Field model that can chunk phoneme sequences into word like units. We then used this "English chunker" on the German transcription. We discarded the hypothesized German word units based on a frequency cutoff of 500, and chose the remainder as super-phonetic units. We added appropriate lexical entries and also induced part-of-speech tags on these word units.

Table 2 shows the result of these word induction experiments. We see a small improvements for German and good improvements for English in the MCD in the word based voices, compared to the base phonetic voices. While these numbers show the correct trend (that more word-like units help improve the quality), the German improvements themselves may not be perceptually significant, since [27] has shown that only improvements over 0.05 can be perceived. Syllabification on the English voice gives good improvements. Festival's syllabifier follows standard phonetic rules for syllabification, and while it was tested more on English, it wasn't built specifically for English. It seems grouping phoneme sequences together is a good idea and we will explore better methods of doing so.

Language	System	MCD
English	Best Proposed Method (IND)	3.055
	Syllabification (IND)	2.873
	CRF Based Words (IND)	3.006
German	Best Proposed Method (WSJ)	2.304
	Syllabification (WSJ)	2.279
	CRF Based Words (WSJ)	2.276

Table 2. Improvement in Synthesis Quality using Word Induction Techniques

7. SUBJECTIVE EVALUATION

Our objective results showed that among the phonetic voices, the oracle phonetic transcription has best synthesis. We also saw that the first pass of transcription obtained (using non-targeted cross-lingual phonetic decoder) has worse synthesis than the one obtained using iterative bootstrapping.

We ran subjective tests on English with these three models. We took 25 utterances from held out test data, synthesized them with the three models. We ran two A-B listening tests: (i) Comparing the oracle phonetic synthesis with the best bootstrapped voice, and (ii) Comparing the best bootstrapped voice with the zeroth iteration. Our A-B tests were conducted online, on Amazon Mechanical Turk. We presented the same utterance synthesized using the two models to compare, and asked which one they think is more understandable. We had upto ten participants listen to each clip, so a total of 250 data points per listening task. Table 3 shows the percentage of votes each system received in the A-B task. We can make two observations: (i) The iter-4 voice is perceptually better than the iter-0 voice, and (ii) the iter-4 voice is not very different compared to the oracle voice. We also ran an A-B test with local volunteer participants and observed a similar pattern. We did an informal subjective evaluation for German synthesis, and found the same trend.

Model A	*Model B*	*A better*	*B better*	*Can't Say*
IND-5000 iter 4	IND-5000 iter 0	46.0%	40.0%	14.0%
Oracle Phonetic	IND-5000 iter 4	40.5%	43.6%	15.9%

Table 3. A/B Listening Task: Understandable Synthesis

8. CONCLUSIONS AND FUTURE WORK

In this paper, we experimented with different amounts of data and saw the effect it had on synthesis quality. We feel it is feasible to collect a few thousand utterances in a target language, but we know that collection more than that requires a more dedicated voice talent.

We can see from the difference between Oracle Word and Oracle Phone synthesizers that there is still a lot to gain by finding better ways to discover words in the target language. Also in our initial experiments in building SMT systems from our source language to our target (learned) phone set, knowing where the word boundaries are critical to good translation. We feel the direction described in [15] is particularly relevant, as it can be seen as also mapping the output of a noisy acoustic phonetic model to normalized word segmentation.

We also are aware that the number of phones produced by the initial cross-lingual decoding as being important, and findings to have a larger initial set (and finding principled methods to reduce them) would offer an advantage, and we feel that using multi-stream decoders may help us here (c.f. [28]).

Our results indicate that we are producing a phonetic-like transcription resulting in somewhat understandable speech (if still not as good as Oracle cases). The next stages of this work are to investigate our techniques with real speech, produce better word segmentation and then also extend it to use machine translation for providing a usable writing system.

9. ACKNOWLEDGMENT

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

10. 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] 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.

[7] Sebastian Stüker and Alex Waibel, "Towards Human Translations Guided Language Discovery for ASR Systems," in *Proceedings of Spoken Language Technologies for Under-Resourced Languages*, 2008.

[8] 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] Man-Hung Siu, Herbert Gish, Arthur Chan, and William Belfield, "Improved topic classification and keyword discovery using an hmm-based speech recognizer trained without supervision," in *Proceedings of Interspeech*, Makuhari, Japan, September 2010.

[12] Ngoc Thang Vu, Franziska Kraus, and Tanja Schultz, "Rapid building of an ASR system for Under-Resourced Languages based on Multilingual Unsupervised Training," in *Proceedings of INTERSPEECH*, Florence, Italy, August 2011.

[13] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson, "A Bayesian framework for word segmentation: Exploring the effects of con text," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.

[14] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2001, pp. 282–289.

[15] 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.

[16] Alexander Clark, "Combining distributional and morphological information for part of speech induction," in *Proceedings of European Chapter of Association for Computational Linguistics*, Budapest, Hungary, August 2003, pp. 59–66.

[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] 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.

[20] Toshiyuki Takezawa, "Multilingual Spoken Language Corpus Development for Communication Research," in *Chinese Spoken Language Processing*, Qiang Huo, Bin Ma, Eng-Siong Chng, and Haizhou Li, Eds., 2006, vol. 4274 of *Lecture Notes in Computer Science*, pp. 781–791.

[21] John Garofalo, David Graff, Doug Paul, and David Pallett, *CSR-I (WSJ0) Complete*, ldc93s6a edition, 1993.

[22] Kishore Prahallad, E. Naresh Kumar, Venkatesh Keri, S. Rajendran, and Alan W Black, "The IIIT-H Indic Speech Databases," in *Proceedings of Interspeech*, Portland, OR, USA, September 2012.

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

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

[25] 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.

[26] Tanja Schultz, "Globalphone: a multilingual speech and text database developed at Karlsruhe University," in *Proceedings of the Seventh International Conference on Spoken Language Processing*, 2002.

[27] John Kominek, *TTS From Zero: Building Synthetic Voices for New Languages*, Ph.D. thesis, Carnegie Mellon University, 2009.

[28] Qin Jin, Tanja Schultz, and Alex Waibel, "Speaker Identification Using Multilingual Phone Strings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, FL, USA, May 2002.