MULTILINGUAL TEXT-TO-SPEECH SYNTHESIS

Alan W Black and Kevin A. Lenzo

Language Technologies Institute, Carnegie Mellon University and Cepstral, LLC
awb@cs.cmu.edu, lenzo@cepstral.com

ABSTRACT

This paper presents a framework for building multilingual text-to-speech systems. It addresses the issue at three levels. First it discusses the necessary steps required to build a synthetic voice from scratch in a new language. The second concerns the building of a new voice without recording any new acoustic data, and the restrictions that imposes. The third more speculative part discusses the steps that would be necessary to allow high quality synthesis of new languages by recording only minimal amounts in that language.

1. BACKGROUND

The construction of high quality synthetic voices is still very hard. However, with better tools, the advancement of faster computers and more disk, the job of building new synthetic voices now requires substantially less resources both in expertise and computation. But at the same time as tools and techniques made it easier to build concatenative speech synthesizers, the expectation for higher quality voices has also increased.

The FestVox [1] system provides tools and documentation for building voices in new languages for the University of Edinburgh's Festival Speech Synthesis System. The project was designed to specifically address the issues of building synthetic voices for minority languages as well as major ones.

The work of documenting the process of building voices in new languages rose out of a number of student projects carried out at Edinburgh University and elsewhere including the German diphone voices created at a summer workshop at OGI, in 1998 [2].

Although the initial tools emphasized diphone voices, the tools have matured to support generalized unit selection voices too. The projects that have used these tools have varied drastically in size and effort involving large commercial entities as well as individual students. The quality of voices built equally varies, and many find that to build a usable synthetic voice in a new language is still a substantial task even if it is easier than it was.

We are aware of at least 40 different languages that this work has been used for including, major European languages such as English., German, French, Italian and Spanish, European minority languages such as Scots and Irish Gaelic, Basque, etc., Asian languages including Chinese, Thai, Korean, Japanese, many of the Indian sub-continent languages as well as Nepali and Pashtu, and other languages from different linguistic groups such as Arabic, Turkish, Finnish, Maori and even Klingon. It seems building a new voice in a new language is understood well enough to be set as a student project.

2. BUILDING A VOICE

To build a voice one must address the following issues:

Define a phoneme set
Create a lexicon and/or letter to sounds rules
Provide text analysis
Build prosodic models
Build a waveform synthesizer

All of these basic processes can be fairly mechanistic. Although adequate solutions can be found for most languages it is very hard in general to find excellent solutions.

Many languages have had significant phonological study and a phoneme set is well defined. However, in practice it is typical to find a number of different phoneme sets defined with some ambiguity and even within a phoneme set there may be different choices in particular uses. For example, even in US English there are choices, should /dx/ (a tap) be phonetic? Or, should /axr/ be distinct from unstressed /er/? A first approximation is usually relatively easy, but there are always harder questions about the best set, eventually we would like some acoustically derived method that is correlated with the particular idiolect of the speaker being modeled.

Lexicon construction is hard, and as consistency in the entries is very important we have provided techniques that aid in the construction of new lexicons. For some languages a hand written set of letter to sound rules is possible especially where the relationship between orthography and phonetics is close. We also provide automatic learning techniques for building letter to sound rules from existing words

with pronunciations [3]. The relative success of these methods are both a measure of the consistency of the lexicons and the relative difficulty of pronunciation in a language.

A more general technique that may be adequate when no lexicon is available and the orthography is believed to be close to the phonology is to use the letters directly as phonemes. [4] showed how a letter-based phoneme set worked adequately for Spanish and could even capture dialectal variation in Castillian and Colombian Spanish, such as letter "c" as /th/ or /s/. Even for English this technique works to some degree.

For some languages, we believe a workable letter-based phone set may be successful. However in our experience with building a Pashtu synthesizer, where no standardized orthography exists, confusion between the writing system and the many varied dialects of the language lead to more problems than the orthography/phonetic relationship itself.

Statistical data-driven approaches to prosodic models, for phrasing, intonation and duration, can be build fairly easily for at least "neutral" sentences. Within a unit selection framework it is common not to explicitly model prosody but rely on the implicit modeling provided by the unit selection process.

3. UNIT SELECTION SYNTHESIS IN ANY LANGUAGE

Unit selection synthesis [5], [6] can offer high quality synthesis without the expert work that would be required to build a formant synthesizer. Although unit selection can produce high quality synthesis, the database must be properly designed to have the right coverage for the language or domain so that the quality is reasonable. [7] discusses the limitations and optimizations that can help in achieving high quality databases for unit selection.

In our present set-up a reasonable database can be found by first selecting a large body of text in the target language (millions of words or more is good). Then using a synthesizer front end, that can segment the text into sentences and then convert the text to phoneme strings. We can then select sentences that will best cover the desired phonetic space of the language, optimizing for diphone/syllable coverage depending on the language. The object of the exercise is to find a relatively small set of utterances that are both natural and phonetically balanced. We typically put other restrictions on the selection such as ensure all words are in the lexicon, and limit sentences to under 20 words in length. This makes the utterances easier to say, reducing the effort required from the voice talent and minimizing errors in their performance. Having around 1000 sentences (perhaps around 40,000 phonemes) seems to be reasonable.

We have also experimented with a more elaborate selection technique, [8] where we first model a particular speaker's acoustic variation and select data based on their actual usage rather than just general phonemes. This may perform better but it is more computationally expensive, and requires an existing model of the speaker, which may not be available when building a new language.

We used the simpler technique in building the CMU ARCTIC voices [9], and have successfully used very similar techniques for a wide range of languages including as Croatian, Thai and Spanish. Also we note that given a suitably balanced set of utterances we can more accurately automatically label the data using acoustic modeling HMM tools such as [10].

The quality and ease with which a synthesizer can be built is still very dependent on the quality of the voice talent and of the recording set up. Even with professional voice actors we have found that speakers who have recorded for speech synthesizers before perform better. Thus there is a consistency and style of delivery which leads to a better synthesizer. Perhaps one should always throw away the first recordings and make the speaker do it a second time.

4. EVALUATION

Evaluation of text-to-speech is very hard as the ultimate quality is based on the perception of the listener. The more the listener listens to the voice the more accustomed they are to its irregularities. This is, perhaps, why ones own synthesizer always sounds better than others.

It is very important to understand that synthesis in languages you are less familiar with, always sounds better than those that you are fluent in. In building synthetic voice for new languages, it is important to include a formal method for evaluation to ensure that the voice quality is as required. Just because it "sounds Chinese" to the Western listener does not mean it does so to Chinese native speakers.

We have defined 5 levels of diagnostic evaluation:

1. Text analysis
2. Lexical and letter-to-sound rule coverage
3. Prosodic/style
4. Phonetic/metrical coverage
5. Word/sentence coverage

The first two can be quantitatively measured, and good front ends and lexical components can be expected to be making less 1% error per token type.

Phonetic coverage can be explicitly checked through DRT and MRT tests and MOS listening tests [11]. Though, it should be noted that high accuracy in isolated confusable words in unit selection synthesizers does not guarantee the same accuracy in fluent text.

In unit selection synthesizers we find that in-domain sentences (where there is a target application), and SUS (semantically unpredictable sentences) [12] stress the unit se-

lection system well and improvements for such sentences make a difference to the overall quality.

Prosodic measures are harder, although there are objective measures it is well known that they only partially correlate to human perception.

The purpose of providing evaluation strategies, is to make it easier for less experienced people to find where the problems are.

5. MULTILINGUAL VOICES

The above build process works, and to a large extent documented [1], and we are aware of many users. Although it is possible to get a voice in a new language in as little as a few days, realistically to produce a good voice you need to spend much longer on it than that.

But this is only one of the problems. We would like to build voices that are capable of multiple languages.

Individual voices that cover multiple languages can be built by recording speakers who are (reasonably) fluent in multiple languages. In the simple case where the speaker is not fully bi-lingual the resulting synthesizers are accented. This is also true whenever we build voices in a language other than the speaker's native language. It is worth pointing out that accented speech is not necessarily a bad thing in speech synthesis. We have run simple tests with US English synthesizers built from a Scottish English speaker and a Chinese English speaker. US listeners are more accepting of errors in the accented voices even when there are unit selection errors.

We must also consider mixed-lingual synthesis where multiple languages are contained within the one utterance as words or phrases. Phonetic coverage can be achieve with multilingual speech data, but specialized text analysis is also required. [13] gives a good overview of the problems and solutions.

6. NEW LANGUAGES WITHOUT RECORDING

At present to support any new languages well it is necessary to record some phonetic examples in the target language. Recording data may not be an option when rapid deployment of a system is required.

Cross language synthesizers are possible. We have done this in a number of cases. One of the early non-English voices in Festival was Basque and we used an existing Spanish diphone synthesizer for waveform synthesis. This is not as ridiculous as it might first appear, although the resulting synthesizer was Spanish accented, it is not unusual for Basque speakers to also be native Spanish speakers. This allowed us to have a speaking Basque synthesizer much earlier in development.

We include support to map native phones in the target language into phones within an existing language so that a working system can be more quickly built. Although when these mapping are used between unrelated languages the result can sound almost silly, such as using English for Chinese.

This method has primarily been supported to allow the ability to label recordings in the target language. For example, in building a Korean diphone synthesizer we map Korean phones to English ones, a process that will loose information, as for example our English diphone synthesizer does not distinguish between aspirated and non-aspirated stops which are phonetic in Korean. We used a DTW (dynamic time warping) algorithm to align the synthesized prompt with English phones with the spoken Korean prompts. The following table compares how the DTW results match with hand-labeled boundaries, this table also compares labeling within language and across dialect (UK to US English).

	type	RMSE	stddev
KED-KED	self	14.77ms	17.08
MWM-KED	US-US	27.23ms	28.95
GSW-KED	UK-US	25.25ms	23.92
KED-WHY	US-Kor	28.34ms	27.52

We have used this cross-lingual labeling technique for many languages. It is quite adequate when applied to short words and sentences. This method works because even though there may be variations in the target language that are not in the source language, in almost all cases, a vowel in one language is more like a vowel than a consonant in another language.

Availability of existing diphone and unit selection synthesizers as in the MBROLA databases [14], can make bootstrapping voices in new languages much quicker. Although there are many existing databases available there has not yet been an organized effort to try to cover major language groups in the world that would make the use of existing databases for related languages more practical.

7. NEW LANGUAGES WITHOUT (MUCH) RECORDING

The next level is to use voice conversion techniques to try to modify some existing database toward the target language. This would require at least some examples in the target language but not as much as would be required to build a whole diphone or unit selection voice.

There has been work in the area, e.g. [15], but it currently requires a least one bilingual database, from which to pre-build a mapping for a new speaker. Rather than supporting new languages, this work is targeting cross-lingual modification of voices. This technique is very useful in speech-to-speech translation where speaker style, (e.g. command

vs compassionate) should be translated from the source to the target speaker.

We are still substantially far way from being able to build synthesizers in new languages without recording substantial phonetic and prosodic examples in that language.

8. DISCUSSION

Although we now have a defined method for building new voices in new languages, it still requires a substantial degree of skill, expertise and care to build high quality voices in new languages. As researchers and speech technologists we may feel we have solved this problem but there are still many languages in the world that do not have support for synthetic voices, and given the lack of literacy outside the top languages these may particularly benefit more from speech technology.

To make this task easier we still need to develop better methods to answer such questions as "how can be find the most appropriate phoneme set from data", "what are the speaker-specific pronunciation rules?". We also need to better understand cross-lingual voice conversion if we are to build voices in new languages more easily.

Improvements in building voices are continuing and are likely to involve automatic adaptation of some "close" language as well as improving tools and evaluation techniques to make the building of voices easier.

9. ACKNOWLEDGMENTS

This work was funded in part by NSF grant 0121631 "Avenues" and NSF grant 0219687 "ITR/CIS Evaluation and Personalization of Synthetic Voices". The opinions expressed in this paper do not necessarily reflect those of NSF.

10. REFERENCES

[1] A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system," http://festvox.org/bsv/, 2000.

[2] M. Macon, A. Kain, A. Cronk, H. Meyer, K. Mueller, B. Saeuberlich, and A. Black, "Rapid prototyping of a german tts system," unpublished report Oregon Graduate Institute, http://www.cslu.ogi.edu/tts/research/multiling/de-report.html, 1998.

[3] V. Pagel, K. Lenzo, and A. Black, "Letter to sound rules for accented lexicon compression," in *ICSLP98*, Sydney, Australia., 1998, vol. 5.

[4] A. Black and A. Font Llitjós, "Unit selection without a phoneme set," in *IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, CA., 2002.

[5] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP-96*, Atlanta, Georgia, 1996, vol. 1, pp. 373–376.

[6] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, 1999, pp. 18–24.

[7] A. Black, "Perfect synthesis for all of the people all of the time," in *IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, CA., 2002.

[8] A. Black and K. Lenzo, "Optimal data selection for unit selection synthesis," in *4th ESCA Workshop on Speech Synthesis*, Scotland., 2001.

[9] J. Kominek and Black A., "The CMU ARCTIC speech databases for speech synthesis research," Tech. Rep. CMU-LTI-03-177 http://festvox.org/cmu_arctic/, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003.

[10] Carnegie Mellon University, "SphinxTrain: building acoustic models for CMU Sphinx," http://www.speech.cs.cmu.edu/SphinxTrain/, 2001.

[11] J. Logan, B. Greene, and D. Pisoni, "Segmental intelligibility of synthetic speech produced by rule," *Journal of the Acoustical Society of America*, vol. 86(2), pp. 566–581, 1989.

[12] C. Benoit, M. Grice, and V. Hazan, "The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences.," *Speech Communication*, vol. 18, pp. 381–392, 1996.

[13] B. Pfister and H. Romsdorfer, "Mixed-lingual text analysis for Polyglot TTS synthesis," in *Eurospeech*, Geneva, Switzerland., 2003.

[14] T. Dutoit, V. Pagel, N. Pierret, O. van der Vreken, and F. Bataille, "The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes," in *ICSLP96*, Philadelphia, PA., 1996, vol. 3, pp. 1393–1397, http://tcts.fpms.ac.be/synthesis/mbrola.html.

[15] M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and N. Campbell, "Cross-language voice conversion evaluation using bilingual databases," *IPSJ*, vol. 43, no. 7, pp. 2177–2185, 2002.