Rapid-Deployment Text-to-Speech in the DIPLOMAT System

Kevin Lenzo, Christopher Hogan, Jeffrey Allen

Carnegie Mellon University
5000 Forbes Avenue,
Pittsburgh, PA 15203 USA
{lenzo,chogan,jeffa}@cs.cmu.edu

ABSTRACT

The DIPLOMAT project at Carnegie Mellon University instantiates a program of rapid-deployment speech-to-speech machine translation; we have developed techniques for quickly producing text-to-speech (TTS) systems for new target languages to support this work. While the resulting systems are not immediately of comparable quality to commercial systems on unrestricted tasks in well-developed languages, they are more than adequate for limited-domain scenarios and rapid prototyping -- they generalize to unseen data with some degradation, while quality in-domain can be quite good. Voices and engines for synthesizing new target languages may be developed in a period as short as two weeks after text corpus collection. We have successfully used these techniques to build a TTS module for English, Croatian, Spanish, Haitian Creole and Korean.

1. The DIPLOMAT Project

The DIPLOMAT project is an experiment in rapid-deployment, wearable, bi-directional speech translation systems. An implementation of the complete synthesis system should be available at an acceptable level of quality within a few weeks after initial recording corpus design, with continual, graceful improvement to a good level of quality over a period of months.

For a small domain, the concatenative synthesis method used here can be quite adequate; however, the current synthesis solution, called "phonebox", is a rapid prototyping tool rather than a substitute for a full-fledged speech synthesis system in the target language. The synthesizer receives text-only input through a socket or pipe, and can easily be replaced by an arbitrary synthesis engine and a small wrapper.

We must often trade off the quality synthesis in the short term in favor of development time, memory limitations, and runtime efficiency, in order to provide basic communication technologies in contexts where communication barriers are significant, and to do so on commonly available hardware. So far, DIPLOMAT has worked with English, Croatian, Haitian Creole, Spanish and Korean.

2. Building Text-to-Speech Systems Rapidly

For each language, we encounter new issues in preprocessing and recording, as well as in constrained optimization. We focus here primarily on the speech synthesis component of DIPLOMAT and the components that are needed in order to build it; other portions, such as machine translation, speech recognition, and the user interface are described in (Frederking et al., 1998)

2.1. The Text Corpus

Building a text corpus is an important first step in DIPLOMAT for several three reasons -- in order to create a language model for the recognizer, a mapping for the translation engine, and a recording corpus for both training the recognizer producing a waveform concatenative synthesizer.

Collection. The difficulty of text corpus collection varies by language. For the case of Korean, the collection of texts is a straightforward process, since information written in Korean is abundant and available from current resources on the Internet. For this case, texts were obtained from Internet broadcasting sources and the selected material did not pose any significant difficulty for Korean speakers.

The task is significantly more difficult for languages that are not widely taught, such as Haitian Creole (Allen and Hogan, 1998, Decrozant and Voss, 1998), because they are "low-density" languages, and there are few available documents in electronic form. Finding electronic texts written in Creole required about five months of part-time research on the Internet, in addition to contacting dozens of non-governmental organizations and literacy institutes worldwide that eventually provided electronic versions of their texts.

It is possible to scan and correct texts from paper documents, but our experience for Croatian and Haitian Creole was similar to that of (Decrozant and Voss, 1998) in that current OCR software packages provide poor recognition accuracy on less commonly taught languages for which customized character recognition has not been specifically developed. Our Creole corpus includes all types of text (e.g., novels, political speeches, language learning books, literacy primers, religious texts, etc.) that have been collected from all available resources whereas the Korean corpus remains in domain with abundant amounts of text.

Quality Control in Text Collection. Following text collection, it is important to have at least two native speakers read through all of the texts in order to correct typographic errors and change foreign loanwords. These types of errors and foreign intrusions adversely affect the phonetic balancing of the corpus because such graphemic sequences actually do not occur naturally in the given language.

Text Corpus Normalization. Normalizing the text to a single coding scheme also requires some effort. One must also prepare written-out versions of all symbolic forms of numerals and signs

(e.g., 1, 2, 5, %, °) because in some languages there may be multiple ways of pronouncing the same symbol. For example, in Korean the numbers under 100 can be pronounced according to the Chinese or Korean characters and over 100 are only pronounced according to Chinese characters. In Creole, dates such as 1980 can be equally pronounced as either the year one thousand nine hundred and eighty or the year nineteen hundred and eighty. From our experience, providing written-out forms for numbers and symbols reduce hand labeling tasks in the post-processing stage.

2.2. Spoken Data

Spoken data is needed for two purposes in the DIPLOMAT system -- in order to build a recognizer and to produce an automatic segmentation of the collected speech, as well as for the speech input to the system as a whole during speech-to-speech operation at run-time.

Designing the spoken corpus. Once the text corpus has been collected and corrected for all evident errors, we then sort all sentences according to length and eliminate any examples of sentences that are two words or less in length. We also remove any examples that exceed 100 syllables in Korean or go beyond two lines on a laptop screen in Haitian Creole. From experience recording 300+ Koreans and 150 Haitians, we have found that participants tend to make more mistakes with lengthy sentences, and thus the recording process significantly slows down. Giving participants a greater number of short sentences is certainly advantageous for the recording process.

Once the short and long sentences are removed. A greedy text selection algorithm is then used to select a phonetically rich, relatively balanced set of utterances in a method similar to that described in (van Santen 1998) -- the features and contexts are enumerated for each utterance, and sentences are added in order of decreasing feature coverage until each feature is covered at least once. That is to say, utterances that have the most yet-uncovered features are selected first, and added to the corpus, until all desired feature combinations that are represented in the text corpus are represented at least once in the spoken corpus. This may be continued in order to have multiple tokens of each feature combination, redundantly filling the design matrix in order to have alternate units.

The featurally balanced corpus is then mixed randomly so that the subsets for the recognition training would not be biased in any way toward length of individual sentences, and then cut up into groups of sentences called slices.

Recording. Once the spoken corpus has been generated, sample tests are run on native-speakers to determine how many sentences an average speaker can complete in a period of 30 minutes. For each sentence, the participant reads the sentence out loud for practice, records the sentence, re-records immediately if a mistake occurs, and listens to the sentence to verify that the recorded file accurately reflects the written form. Slices of equal number of sentences are then created from the results of these sample tests.

Each individual speaker in the recognition training set records one slice of sentences in order to train the speaker-independent speech recognizer for use in live speech-to-speech translation. Individual speakers are then selected as synthesis voices, and a systematic recording, preferably of the entire corpus (all the slices), is done using that speaker; the newly-created recognizer is used to align the segment boundaries.

Koreans completed 35 sentences in 30 minutes on average, whereas Haitian Creole speakers completed approximately 25 sentences in the same amount of time. This appears to be due to the level of familiarity with the written forms of the language and is therefore a very significant factor to be considered when conducting speech processing data collection projects for less commonly taught languages. For Spanish, we were able to record some 2,000 prompts from a single speaker over three days. The English and Croatian each consisted of about 10,000 short utterances collected over about two weeks.

A simple web-based interface allows non-professionals to participate in real-time recording and verification. We use the Netscape web browser with a recording plug-in as the recording interface; it supports unicode formatting, which is necessary for our multiple language project, and provides a useful display and interaction mechanism. The simple interface has only a few functions: record, re-record, playback, next sentence, previous sentence. Common symbols for VCR/VHS and tape recorders (e.g., red dot, arrows, etc.) are used for the interface buttons, and the text of each button is translated into the different languages so that the participant may self-record.

English, Korean, Croatian, and Spanish-speaking participants, mainly students and staff at Carnegie Mellon, are often quite facile with computers, and so they require very little intervention. Creole speakers are usually less familiar with computers, and so often require considerable aid by a recording technician or assistant during the collection.

Quality Control in Recording. Lexical variation in non-standardized less commonly taught languages is important to note. The pairs (aprè and apre, jodi and jòdi) are very common examples of phonetic variation for the same lexico-semantic item in Haitian Creole. There is a high risk that participants will not pronounce a given written lexical form that is presented on the screen but may rather rely on their internalized pronunciation which reflects a slightly different written representation in Creole. Examples of this are numerous for this language. High phonetic variability for a traditionally oral language, with respect to standardized lexical forms, therefore requires much control during recording; a native or near-native speaker of the language must be present to catch words that are pronounced in a manner different from the written form. In such a controlled situation, errors are caught immediately and thus the rejection rate for misaligned sound files with the text is very low.

2.3. Building a Recognizer/Aligner Rapidly

We have explored two methods for reducing the development time of a TTS system. The "Assimilation" method can be used when acoustic models exist for languages that are phonetically similar to the target language. Otherwise, the "Simplified Model" method can be used.

Assimilation. In the case that the language to be developed is phonetically similar to another language for which acoustic models already exist, the Assimilation method may be used to perform indexing without requiring that new acoustic models be built. Acoustic models for the existing language are remapped

so as to cover the phonemes in the new language. These models are then used to segment the new language.

We have successfully applied the assimilation method to segment Croatian speech. The phoneme set of English was deemed to be sufficiently close to that of Croatian, and English models were adapted and used to segment the Croatian. This technique has also been successfully used to build a speech recognizer for Croatian (Frederking et al. 1997).

Simplified Model. The assimilation technique can only be used in a very small number of circumstances where the language to be studied is sufficiently close to one for which acoustic models already exist. Because this is rare, we have developed other techniques that do not rely on this state of affairs.

In the simplified model approach, we use the synthesis data to bootstrap its own speaker-dependent acoustic models, which can then be used to segment the data. This relies on the fact that a forced alignment is produced during the Baum-Welch phase of the acoustic training process. Usually, the amount of data collected for synthesis is not enough to build an accurate continuous speech recognizer, and such a recognizer would certainly not be speaker independent. However, by working with context-independent phonemes, and restricting to a single speaker, we have been able to automatically segment the synthesis data.

3. Preprocessing for Synthesis

A number of steps are needed to prepare the synthesis corpus for use in DIPLOMAT; these are segmentation, normalization, indexing, and minimizing the corpus size.

3.1. Segmentation/Alignment.

After recording, and building a recognizer/aligner, the data must be segmented for use by the synthesis system. Word and triphone segmentations are produced by SPHINX II or III (the speech recognition system) running in forced-alignment mode, either as a by-product of recognizer training, or from the speaker-dependent models. The segment labels include segment identity, the file name of the utterance that contains the unit, start and end times, and the associated acoustic scores for the triphones (phone with left and right context) and words.

3.2. Normalization

The audio files for each utterance are normalized to a common maximum volume by finding the loudest speech sounds and scaling that to be a fixed fraction of the dynamic range. This reduces the effects of changes in volume during data collection, and helps reduce discontinuities during the waveform concatenation at run-time.

3.3. Indexing and Selection of Units

Given the segmentations of the synthesis corpus, a series of tables is created to index every available unit. In particular, each phone (in context) and word is indexed by identity, position within the container (word for phones, utterance for words), and acoustic score as given by the alignment stage. These indices are used for rapid lookup at run-time.

3.4. Minimizing the Synthesis Corpus

Design factors for the DIPLOMAT project have imposed severe constraints on the implementation of the synthesis system. The intended deployment platform for the system is wearable, low-profile and lightweight. Currently, such systems offer little in the way of processor, disk or memory.

Given that this concatenative technique works by storing a large database of speech segments online, the storage requirements of even a modest system can be substantial, even hundreds of megabytes. In order to minimize the footprint while maintaining sufficient quality and speed to be usable, we use an algorithm to scale the corpus size to the condition and eliminate redundant units. The basis of this algorithm is the same greedy search technique described for spoken corpus design above, except focusing strictly on a limited set of features for the task.

A dense set of utterances which will cover a set of desired features may be much smaller than the complete corpus. The greedy algorithm is usually run to completion four times in order to cover high-frequency in-domain units in context. We limit the size of the resulting corpus to about the best 90 megabytes, and have found that this is usually sufficient to cover completely the first three sets of features, and about half of all the words in the corpus.

Real-time operation is very important in a dialogue system, and users are willing to accept lower quality in exchange for faster interaction; this can be achieved by reducing the number of features during the greedy selection process.

3.5 Building the Text-to-Phoneme converter

A text-to-phoneme converter is built for each target language. This may be one of two types: a production-rule system, if the language is orthophonetically simple, or a decision tree built from a pronunciation dictionary, if one is available. The decision tree method is described in (Black et al., 1998) and (Pagel et al., 1998); the production rules are designed by hand.

4. Run-time

The concatenative synthesis scheme in phonebox is similar to other selection-based waveform concatenation systems that do not use prosody modification, such as the Festival system from the University of Edinburgh (Black and Taylor, 1997).

4.1 Normalization

During synthesis, the text is tokenized. Since the “text” to be spoken is actually the output of the machine translation component, there is little or no punctuation -- lookup is largely lexical, and each utterance is given as a single production, so the normalization component of the input text is minimal in comparison to a full Text-to-Speech system.

The word sequences are converted to phone sequences using the Text-to-Phoneme conversion module built for the language as in (Black et al., 1998) and (Pagel et al., 1998).

4.2 Search

A maximal span of word samples is computed for regions that have word samples, and maximal subspans of phoneme sequences are generated for the remainder, accounting for the phonemes at the word boundaries as context. The set is ranked according to duration-normalized acoustic score and feature satisfaction of the left and right contexts. The criterion effectively minimizes the number of join points in the output and uses the best acoustic match score.

This method is similar to that described in (Black and Taylor, 1997) for the Festival system and (Campbell and Black, 1996) for the CHATR system at ATR, Japan.

4.3 Waveform Concatenation

Once a set of candidate units are selected, adjacent candidates are cross-correlated in order to find the harmonically most similar join points within a threshold determined by the segmental identity. The units with the highest cross-correlation peaks are then cut out of their containing utterance with a fixed amount of extra samples on the left and right; in the case that there is only one unit, it is the best one; for corpora with small disk footprints, this may be the case. The extra samples are used to crossfade within a 10 ms window, with no attempt to optimize based on any other criteria. No prosody modification is used whatsoever.

5. Conclusion and Discussion

The DIPLOMAT project has rapidly produced synthesizers for several languages. Although the quality does not equal that of commercial systems, the result is useable and communicative, with a short development cycle. Future work will focus on automatic prosodic modeling, the quality of synthesis, and making the entire voice-generation process a turnkey operation.

We are considering adopting the Festival synthesis system and working within its framework to create a set of tools for rapid voice and language creation. Festival is a well-developed system with more functionality than phonebox; working within the common reference could make it easier for us to share voices and data with other sites, as well as mitigate the transition to a full Text-to-Speech system. All of the steps outlined in this paper still apply, as we attempt to rapidly synthesize new target languages.

5. References

1. Allen, Jeffrey and Christopher Hogan. “Evaluating Haitian Creole Orthographies from a Non-Literacy-based Perspective,” paper presented at the annual meeting of the Society for Pidgin and Creole Linguistics and Linguistics Society of America, New York City, 9-10 January, 1998.

2. Black, Alan, Kevin Lenzo, and Vincent Pagel, “Issues in Building General Letter to Sound Rules,” ESCA Speech Synthesis Workshop at the International Conference on Spoken Language Processing 1998, Sydney, Australia.

3. Black, Alan and Paul Taylor, “Festival Speech Synthesis System: system documentation (1.1.1),” Human Communication Research Centre Technical Report HCRC/TR-83, 1997.

4. Campbell, Nick, and Alan Black, “Prosody and the Selection of Source Units for Concatenative Synthesis,” in "Progress in speech synthesis", eds. J. van Santen, R Sproat, J Olive and J. Hirschberg, pp. 279-282, Springer Verlag, 1996.

5. Decrozant, Lisa, and Clare Voss. “Cross-linguistic Resources for MT evaluation and Language Training,” paper presented at the International Conference on Natural Language Processing and Industrial Applications (NLP+IA 98), Moncton, New Brunswick, Canada, 18-21 August, 1998.

6. Frederking, Robert, Alexander Rudnicky, and Christopher Hogan. “Interactive Speech Translation in the DIPLOMAT Project,” In Proceedings of the Spoken Language Translation Workshop at Association for Computational Linguistics (ACL 97), Madrid, Spain. pp. 61-66, 1997.

7. Frederking, Robert, Alexander Rudnicky, Christopher Hogan, and Kevin Lenzo, “Interactive Speech Translation in the Diplomat Project,” Machine Translation Journal, Special Issue on Spoken Language Translation. Guest Editor: Steven Krauwer (Utrecht University). Expected in 1998 (submitted).

8. Pagel, Vincent, Kevin Lenzo, and Alan Black, “Letter to Sound Rules for Lexicon Compression,” International Conference on Spoken Language Processing 1998, Sydney, Australia.

9. van Santen, Jan, and Richard Sproat. “Optimal Text Selection,” in Multilingual Text-to-Speech Synthesis: The Bell Labs Approach; Richard Sproat (ed.), Ch. 2, section 4, pp. 17-20, Kluwer Academic Press, 1998.