RAPID DEVLOPEMENT OF SPEECH-TO-SPEECH TRANSLATION SYSTEMS

Alan W Black[1]*, Ralf D. Brown*[1]*, Robert Frederking*[1]*, Kevin Lenzo*[2]*,*
John Moody[4]*, Alexander Rudnicky*[1]*, Rita Singh*[3]*, Eric Steinbrecher*[4]
[1] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA
[2] ISRI, Carnegie Mellon University, Pittsburgh, PA,
[3] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA,
[4] Lockheed Martin Systems Integration, Owego, NY

ABSTRACT

This paper describes building of the basic components, particularly speech recognition and synthesis, of a speech-to-speech translation system. This work is described within the framework of the "Tongues: small footprint speech-to-speech translation device" developed at CMU and Lockheed Martin for use by US Army Chaplains.

1. INTRODUCTION

The DIPLOMAT system [1] was developed as a speech-to-speech translation system that could be readily adapted to new languages. It was designed to run on a small platform, such as a laptop or wearable. These requirements introduce different constraints on the system when compared with the larger more general speech-to-speech translation systems such as those in CSTAR.

The later TONGUES project [2] was to built a prototype speech-to-speech translation system designed to run on a sub-notebook computer for use by US Army Chaplains for communicating with locals on issues of refugees etc. This prototype was tested in the field in Zagreb in April 2001, see [3] for a full description of that evaluation.

In DIPLOMAT and TONGUES it was not just the end system that was being developed, it was the processes involved in building the components, so application to new languages and domains require less effort. DIPLOMAT developed basic versions in: Croatian, Korean, Spanish and Haitian Creole. TONGUES targeted Croatian alone though almost all aspects of the models were rebuilt for that system.

This paper describes the technique we used to develop the basic models for each of the components in the system.

2. PRELIMINARIES

As the intention is rapid development of new languages we cannot afford the time to do full linguistic analysis of each new language. therefore we have mostly adopted a approaches relying on data-driven techniques.

First we wished to identifying the domain. Although everyone wants their translator to be as general as possible, in all practical situations it will in fact be restricted to one or more domains. Therefore the first task in the work was to record dialogs in the expected domain.

In the first months of the TONGUES project we collected together a number of US Army Chaplains. Each was provided with a headset with a head mounted microphone and asked to role play short dialogs. These were recorded in stereo (one person on each channel). The chaplains were accustomed to such role playing and devised their own scenarios involving refugees, medical emergencies, food distribution, etc. Both sides were playing in English.

46 dialogs (and some monologues) were recorded over two sessions. These ranged from just over one minute to just over 15 minutes, with an average of 5 minutes 26 seconds.

These recordings were segmented automatically into non-speech and speech regions, giving a total of 4.26 hours of speech. This speech, mostly spontaneous conversational speech, was hand transcribed at the word level. In addition to actual words spoken, false starts, filled pauses etc were explicitly labeled.

3. SPEECH RECOGNITION MODELS

Four speech models are necessary for the speech-to-speech translation system: acoustic models and language models for each languages.

For this project we use the CMU Sphinx II recognition engine [4]. A semi-continuous HMM-based recognition engine which requires relatively low computational requirements to run.

For English we used the 4.25 hours of speech from chaplain dialogs to train new acoustic models using the SphinxTrain acoustic model package. The advantages that the training data was in the intended domain, and the right channel and recording conditions outweighed the lack of data. Also some (not all) of the participants in the original dialogs would be actual users of the system.

Another advantage is that the chaplain, as a regular user

of this device, is likely to become an expert user. They will learn how to speak to the machine so that it works.

The chaplain database was too small to build a reasonable language model. Although we did include that data, we also took data from chaplain hand books to produce a larger text set. Word-trigram models were built with absolute discounting.

The Croatian models were not as straight forward to build. In the original DIPLOMAT system we used English HMM models to seed Croatian models and adapted them with some data, a much more elaborate method along these lines is described in [5]. However, in TONGUES, we decided to try to record sufficient data to build new models from scratch.

The first task was to collect Croatian speech. We did not have easy access to Croatian speakers, nor were the available speaker capable of role playing chaplain/refugee dialogs. As the device is designed for Chaplains to talk to a varied population of Croatian speakers, we would like a wider range of speakers. Following techniques in data selection for optimal acoustic coverage for speech synthesis [6], we constructed sets of utterances to record which would provide the desired phonetic variation.

We took the translations of the chaplain dialogs and used the basic Croatian synthesizer to generate phoneme strings for each utterance. We then greedily selected the utterances that had the best diphone coverage, (phone plus previous phone). The selected utterance were then removed from the pool and selection was re-applied. This was repeated over the whole corpus thus partitioning the utterances into sets of about 250, each with good phonetic coverage.

We then recorded 15 different native Croatian speakers each reading one of the sets (there were 5 sets used in total). Each speaker was prompted with displayed Croatian text and recorded (possibly with corrections if they made mistakes). This had bypassed the need for hand transcription, which we would have had to teach our Croatian speaking helpers. Of course there is also the disadvantage that this is read speech not spontaneous speech.

This provided 4.0 hours of Croatian speech, from 13 female speakers and 2 male. The bias for female speech was due to scarcity of available male Croatian speakers. This data alone was used to build Croatian acoustic models.

The gender bias was not a problem when we ran actual experiments in Zagreb, as both male and female speaker we recognized alike. The problem of using read speech was also not so much of a problem as we feared. In such systems, the time taken for each turn is much longer than would be used for a turn in single language dialogs. The speech must be recognized, checked, translated, checked, and synthesized. Each of these stages is done before the next stage starts, although they could be partially streamlined, we deliberately wanted users to have the opportunity for correction. Each of these processes takes time, thus speakers quickly learn that spontaneous speech is not the best way to transmit content and they tend to produce much simpler well-structured utterances. Of course that the recognizer (and translation engine) also work better on those types of sentences, will also encourage users towards that style.

One problem we identified was that because there were no filled pauses, equivalent to “um” and “eh” in the Croatian data, short function words were often spuriously recognized. We have discussed one method to alleviate this, though have not tested it yet: using English noise models directly in the Croatian recognizer is a low cost solution though as we have not tested, we do not have any results on its suitability.

As with English, Croatian language models, were word tri-grams built with absolute discounting.

The language model-vocabularies were 2900 words for English and 3900 words for Croatian. In pilot experiments with held out test sets, the word error rates were found to be below 15% for English and below 20% for Croatian.

4. SPEECH SYNTHESIS MODELS

The CMU FestVox project [7] provides documentation, tools, and explicit walk throughs for building synthetic voices in new languages for the Festival Speech Synthesis System [8].

A synthetic voice requires the following modules:

text analysis: takes strings of characters and finds the words required to speak them, expanding numbers, abbreviations, symbols etc.

lexicon: a method for finding pronunciation of words, either through an explicit word list and/or letter to sound rules (which may be hand written or trained from data).

prosody models: to provide phrasing, duration and intonation.

waveform synthesis: converting strings of phonemes (with prosodic and metrical structure) into waveforms.

For the TONGUES project we were only constructing a Croatian voice. For the English side we used a standard US English voice.

For Croatian we first defined a phoneme set (shared with the Croatian recognition system). Croatian orthography is closely related to its pronunciation so a set of letter to sound rules were written by hand with little problem. In addition basic symbols were added explicitly to a lexicon. The combined lexicon and letter to sound rule set were used for both the synthesizer and the speech recognition engine.

The next stage was to define other text analysis. With aid from Croatian native speakers (and substantial amounts of example text), we defined some standard abbreviations.

Real examples are important when discussing text analysis with native speaker who have never considered speech synthesis before. While the expansions are typically trivial they do need to be codified. Numbers were simply treated as string of digits. Proper treatment of numbers would require identification of case.

Prosody is phrasing, intonation and duration. In this system we fell back on punctuation alone for phrasing. As sentences are generated by the translation engine, there is no punctuation generated, so each utterance is treated as a single phrase. This would be unacceptable for reading paragraphs of text but this device will be typically used with short utterances. Duration models are trained directly on the data recorded for waveform generation. As the duration models were based directly on the recorded speech they produced appropriate durations.

We first attempted to build Croatian intonation models but they were not good, therefore we used English intonation models. We had Croatian speakers listen to them. All of the Croatians preferred the English model (though did not know it was an English model). This is not because English intonation models are similar to Croatian but that we could not reliably extract F0 contours from the our recorded speech, and eve if we could there was probably insufficient data.

Although a diphone based synthesizer would have been adequate for this application (the English voice is a diphone voice), we wanted to take advantage of some of the aspects of domain synthesis [9]. Thus we built a unit selection synthesizer using a database containing sentences selected from the translations thus using in-domain sentences.

The original chaplain dialogs were first translated into Croatian. These were then split into sentence sized chunks. These were given to the basic synthesizer and converted from text to phoneme lists. The utterances were then greedily selected finding the set of utterance that had the maximum diphone coverage. This gave rise to 638 utterances. A second set were selected from other Croatian text data, that set was 975 utterances, these sentences which were typically much longer. The first set plus around 70 of the second set were recorded.

The data was then autolabelled with a cross-linguistic phoneme aligner. We first generate synthetic versions of the utterances by mapping the Croatian phones to English phones and synthesizing the speech. These synthetic utterances sound very English-like but because they were synthesized we knew where the phones started and ended, and they are never used except for alignment. Using a DTW (dynamic time warping) technique based on [10] we align the synthetic utterances with the natural Croatian ones allowing us to find label boundaries. The labels were then hand checked.

5. TRANSLATION MODELS

The translation engine used was a Multi-Engine MT (MEMT) system [11], whose primary engines were an Example-Based MT (EBMT) engine [12] and a bilingual dictionary/glossary. Carnegie Mellon's EBMT system uses a "shallower" approach than many other EBMT systems; examples to be used are selected based on string matching and inflectional and other heuristics, with no deep structural analysis. The MEMT architecture uses a trigram language model of the output language to select among competing partial translations produced by several engines. It is used in this system primarily to select among competing (and possibly overlapping) EBMT translation hypotheses.

The translated chaplain dialogs provided some of the training but we also relied pre-existing parallel English-Croatian corpora.

AN addition finite-state word reordering mechanism was added to improve placement of clitics in Croatian.

6. OVERALL INTERFACE

You cannot just bolt recognizers, synthesizers and translation engines together and except to have a working translation aid. A well designed interface is required to take into account the known limitations of components.

The device is designed to be primarily "driven" by the English speaker, but expects the non-English speaker to be new. Hence a number of pre-recorded utterances in Croatian were available directly from the main screen. These included basic commands, such as "Halt", and informative messages, like "We are here to help," and basic instructions and description of the machine itself.

As the English speaker goes first the non-English speaker can see the basic operation. The English speaker says their sentence and the recognizer prints the recognized utterance word by word. The speaker can correct this using the keyboard or by respeaking it. Then the speaker can translate the English utterance into Croatian. As the English speaker does not know if the translation is reasonable, they may also back translate the Croatian sentence into English. If the message is clear after the double translation it is assume the Croatian is probably correct (though a bi-lingual informant who watched over our evaluation said that often the multiple translations caused speakers to change their utterance even when the translation was acceptable). Once they are content with the translation they can use the synthesizer to render it as speech. The Croatian speaker follows a similar route, but with more instructions (in their language).

As we wished to allow the system to improve with use, we included the facility to add new words and fixed translations to the system.

7. EVALUATION

As part of the TONGUES project we had the opportunity to test the system. In April 2001 a group of US Army chaplains and some the authors took two of the devices to Zagreb, Croatia, to carry out field tests. These tests actually took place in rooms in the University so the environment was not as noisy as it could be in real use.

Full details of the test and evaluation are given [3], but a brief description is given to back up our conclusions. The tests were run over three days, in all 28 dialogs were collected, though 10 of these had less that three Croatian turns in them, hence were ignored.

The dialogs took from 14 to 68 minutes, with an average of 24 minutes. Each dialog typically started with 6 (or 7) prerecorded prompts lasting around a minute.

On average the English side took 2.67 turns more than the Croatian, though it was noted that often when the answer to a question posed by the US Army chaplain was yes or no, the translation device was not actually used.

The following shows the number of words and turns over all the transactions. Pre-recorded phrases are not included here, the translated English words are counted for the Croatian turn rather than the Croatian words.

	words	turns	w per t
English	1019	218	4.67
Croatian	355	101	3.51

As we can see expressions are typically short. In user questionnaires, when asked "what works" almost half said "short sentences". There are probably a number of reasons for this. First limitations of the components themselves but also limitations of the whole system itself. Its hard to pass lots of information in a conversation where each turn takes an average of 81 seconds.

8. CONCLUSION

The TONGUES project was not just to deliver a running prototype, it was also to investigate the amount of effort required in building a domain target speech-to-speech translation system. For the most part the basic tools had already been developed before the start of the one year project. However some continuing development did take place. We estimate there was around 2 person-years total effort by the senior staff plus part-time Croatian informants, chaplains and some student helpers. Although most of the data was collected as part of this project we did use some previously collected data in the development, (especially bi-lingual parallel text corpora).

We also feel that although the technology of building new models in new languages is becoming better, the decisions about how best to use that technology are still not automated. Another aspect of the work that should not be underestimated is the amount time required to manage and train labelers and translators, who although native speakers of the target language typically have little computational skills, or appropriate linguistic awareness of their language.

As a final note, we felt that the device performed in the tests adequately it. It did aid successful communication between two parties who did not speak the same language. We are aware that the field test was far from real usage but it was much more realistic than laboratory testing.

9. REFERENCES

[1] R. Frederking, A. Rudnicky, C. Hogan, and K. Lenzo, "Interactive speech translation in the diplomat project," *Machine Translation Journal*, vol. special issue on spoken language translation, 2000.

[2] A. Black, R. Brown, R. Frederking, R. Singh, J. Moody, and E. Steinbrecher, "Tongues: Rapid development of a speech-to-speech translation system," in *HLT2002*, San Diego, 2002, pp. 2051–2054.

[3] R. Frederking, A. Black, R. Brown, J. Moody, and E. Steinbrecher, "Field testing the tongues speech-to-speech machine translation system," in *LREC*, 2002.

[4] X. Huang, F. Alleva, H.-W. Hon, K.-F. Hwang, M.-Y. Lee, and R. Rosenfeld, "The SPHINX-II speech recognition system: an overview," *Computer Speech and Language*, vol. 7(2), pp. 137–148, 1992.

[5] T. Schultz and A. Waibel, "The globalphone project: Multilingual lvcsr with janus-3," in *Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop*, Plzen, Czech Republic, 1997, pp. 20–27.

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

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

[8] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," http://festvox.org/festival, 1998.

[9] A. Black and K. Lenzo, "Limited domain synthesis," in *ICSLP2000*, Beijing, China., 2000.

[10] F. Malfrere and T. Dutoit, "High quality speech synthesis for phonetic speech segmentation," in *Eurospeech97*, Rhodes, Greece, 1997, pp. 2631–2634.

[11] R. Frederking and R. Brown, "The Pangloss-Lite Machine Translation System," in *AMTA*, Montreal, Quebec, Canada, October 1996, pp. 268–272.

[12] R. Brown, "Example-based machine translation in the Pangloss system," in *Proceedings of COLING-96*, Copenhagen, Denmark, 1996, pp. 169–174.