Challenges with Rapid Adaptation of Speech Translation Systems to New Language Pairs

Tanja Schultz and Alan W Black
Interactive Systems Laboratories, Language Technologies Institute, Carnegie Mellon University
{tanja, awb}@cs.cmu.edu

Abstract

Although we have far from solved the issues in porting speech translation systems to new languages, we have gathered sufficient experience by now to identify a number of major challenges in the process. Although well-defined processes exist for building speech recognition, speech synthesis and statistical machine translation models, they still require both significant native speaker involvement and linguistic expertise. As the core technology improves we believe we will see increasing cultural and social issues in contributions from native speakers. This paper identifies some of these issues and presents our initial attempts to build tools that we hope will eventually allow linguistically naive native informants build complete speech translation systems.

1. Introduction

As speech translation methods and systems rapidly evolve, the public demand for supporting speech translation in new language pairs increases tremendously. Due to the pivotal role of English, a newly demanded language pair often involves English as input or output language. In this case, the translation from English to a new language requires the development of speech synthesis (TTS) in this language and the translation from English to this new language (MT E-N), while the translation from the new language to English requires the development of automatic speech recognition (ASR) in the new language and translation from the new language into English (MT N-E). Language pairs not involving English can benefit from using English as a pivot language for data sparseness reasons. For instance, when translating between Mandarin and Spanish, bilingual resources are extremely sparse, while more data will be available between Chinese and English and English and Spanish. This justifies the approach of translating via English, i.e. feeding the output of a Spanish-English system as input to the English-Chinese one. Reichert showed that this approach gives significant improvements [Waibel et al., 2004]. Other pivot languages might be Mandarin for translation among Chinese languages, or Modern Standard Arabic for translation across colloquial Arabic dialects.

However, the major obstacle for the development of speech translation systems in new language pairs might not only be the data but the lack of technology experts who are fluent or at least knowledgeable in the language in question. Depending on the language familiarity it might even be challenging to find an appropriate language expert. Consequently, one of the central issues in building systems for new languages is to bridge the gap between language and technology expertise.

One solution is to build tools that solicit the required expertise from native speakers of that language no matter if they are experts or not. In this paper we discuss the challenges of developing these tools, compare the efforts in cost and time to develop speech translation systems, and describe our strategies to overcome the ever-existing data sparseness problem and the described knowledge gap.

2. Major Challenges

Availability of Suitable Data

With well established (mostly statistically based) methods, it is well known that one major problem of building speech processing systems is the lack of large amounts of data. As the community turns towards less widespread languages, the chances increase of encountering unseen challenges such as missing standardizations or even the complete lack of a writing system.

Vast amounts of data are required to develop a speech translation system in new languages. Although the amount highly depends on the task, for the development of a domain limited speech translation system we found that roughly 5-10 hours of transcribed speech data from tens of speakers can serve as a general baseline ASR component which can then be adapted by a much smaller amount of task specific speech data. For TTS it is favorable to get at least one hour of data from a single voice talent. Reasonable statistical based translation can be achieved with as little as 50k sentence aligned bilingual text corpus in a restricted domain [Schultz et al., 2005].

At CMU we developed a number of tools to retrieve text data from the web and to collect speech data from written corpora. Over the past years we collected data in around 20 languages mostly in the framework of GlobalPhone [Schultz, 2002]. These languages include Arabic (MSA, Egyptian), Bulgarian, Chinese (Mandarin, Wu), Creole (Haitian), Croatian, Czech, English, French, German, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Tamil, Thai, and Turkish. We found that it takes roughly 4 weeks for a single native non-expert to collect a ready-for-use speech and text database if these tools are applied (this includes planning, traveling, collecting, and data post processing). We found that the easy access to native speakers is the most important time factor. Therefore we usually collect in a country where the language is spoken.

For TTS we have also built many new languages. The FestVox scripts [Black and Lenzo, 2000] have been used for at least 40 different languages, while within our own group, recently we have built: Pashtu, Iraqi Arabic, Egyptian Arabic, Thai, Afrikaans, Catalan, Chinese, Spanish, Basque, Galician, and Welsh. In general we collect around one hour of high quality phonetically balanced recordings from a single speaker and use unit selection techniques to synthesis new words and phrases by selecting appropriate sub-word units.

1-4244-0469-X/06/$20.00 ©2006 IEEE

Lack of Language and Cultural Expertise

With a suitable amount of data, the development of speech translation components might be considered to be a rather straightforward task since algorithms and methods for ASR, MT, TTS are well established and in place. However, as it became obvious in the recent evaluation of the DARPA TransTac "100-day challenge", the critical factor for timely development of high performance systems in new languages is the lack of language and cultural expertise. For the purpose of speech translation it is desirable to find a person who is fluent in both languages but also understands linguistic issues.

Language expertise is indispensable for a number of tasks in speech translation. It would be desirable to have language experts who (1) highlight the major challenges of a given language and the cultural differences, (2) provide a network of people speaking this language, (3) coordinate or collect speech and text data, (4) help define useful phone sets for ASR and TTS, (5) analyze or post-edit pronunciation dictionaries, (6) clean or verify available speech transcriptions, (7) process monolingual text corpora, (8) verify alignments of bilingual corpora, (9) analyze the output of prototypical system components, and (10) identify the possible error sources that are related to the language, in order to help the technology experts improving their components.

Even when closely supervised, some of the listed tasks require not only language skills but also a general understanding of the technologies involved. Furthermore, some of the technical problems are non-trivial to solve. Social and cultural aspects may further complicate the process, e.g. native speakers may not wish to have their voices recorded (for recognition or synthesis), or because they are non-native, communication can be difficult. We have also seen an interesting social issue where naïve native speakers are eager to help, but in being helpful they are reluctant to identify parts of the system that are wrong. Thus care must be taken to best utilize their answers in order to maximally improve pronunciations and translations, for instance. Using speech and language unaware native speakers efficiently is an important requirement in building support in new languages. One route we follow is to build up tools that provide support to develop these resources. Our final goal is to provide these tools for novices in speech processing applications.

Language Peculiarities and Challenges

Over the past year we have dealt with a large variety of languages and came across a large set of diverse challenges inherent in these languages. Challenges occur on different linguistic levels and include (1) various writing systems (logographic, and segmental, syllabic, or featural phonographic scripts such as Hanzi, Kanji, Roman, Cyrillic, Devanagari, and Hangul), the Arabic script that does not provide long vowels, and languages without any written form at all (such as Egyptian and Iraqi Arabic where scripts had to be invented [Maamouri, 2004]), furthermore (2) written forms without segmentation, (3) rich morphology languages due to agglutinative or compounding rules, (4) tonal and stressed languages, and (5) a variety of different letter-to-sound relationships. These varieties have a large impact on the time and effort to be spent during the building process of speech translation systems and required detailed language knowledge to be solved appropriately.

Lack of Language Conventions

When a language is not normally written it significantly hinders text processing and lexical construction. In Arabic dialects, for instance, no standard writing system exists. Specifically, when Iraqi Arabic speakers write, they will normally use Modern Standard Arabic (MSA) which can more easily be understood throughout the Arabic speaking world. But when they speak they use their own dialect which although related to other Arabic dialects is at least difficult to understand by non-Iraqis and may be unintelligible to Arabic speakers from further away countries. In building speech-to-speech translation systems it is the spoken dialect that we need to model and not the more formal MSA that may only be spoken in official proclamations.

This issue highlights a common problem found in languages that are not normally written. Where there is a common writing system, taught within a common education system, conventions for writing become well defined. But in writing a dialect, even one where many of the speakers are literate, but in another language, there is significant disparity in the choice of spelling. Words may have several different pronunciation based on their written form, they may have several different written forms with the same pronunciations, or they may have several different written and pronunciations forms. Of course to confuse things further words that share the same pronunciation and/or written form may actually denote different semantic words.

What is required, is to explain this to a native speaker (who will typically not be aware of any these distinctions), and have them identify which can be combined and which cannot.

3. Building Components

Over building speech translation system in a wide variety of languages and domains we have discovered that the basic process still requires too much skill, and finding an informed native speaker is significantly harder, especially when we move away from the major languages of the world.

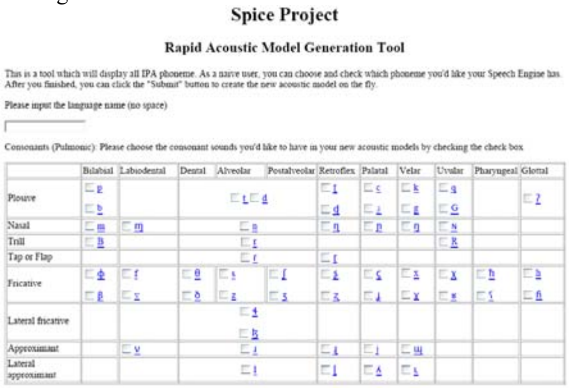
Phone Set Definition

From the phonological point of view we need not necessarily discover the actual phoneme set, but discover an appropriate discrete set of labels which can be distinguish automatically by speech recognition and synthesis techniques. [Kominek and Black, 2005] proposes a method to discover acoustic distinctions based on initial labels using the orthography alone. The simple experiment tried to separate English fricatives /jh/, /zh/, /ch/ and /sh/. The issues investigated include the best distance metric to measure the differences, and the best machine learning technique to do the classification. We envisage using this technique to better separate predictable acoustic dissimilarity in defining the phone set.

Lexical Issues

Traditionally building pronunciation lexicons for new languages was a very time-consuming process, but with the demand for support for new languages a more efficient method is required. We have been using statistical methods to predict pronunciations of unknown words for some years [Black et al., 1998], but these require some initial data to train from. Rather than simply collecting word pronunciation data randomly, we have devised an efficient method to collect the pronunciations for the most useful words, both in their contribution to the overall accuracy of the pronunciation of general text, and in aiding the training of statistical models for unknown words. The basic technique, more fully described in [Maskey et al., 2004], starts with a set of a few hundred high frequency words, including words that give alphabet coverage for the language. The pronunciations of these words must be hand transcribed but once entered we build basic letter-to-sound models. Then in batches of around hundred words, selected as high frequency words across a number of documents, these are presented to the native speaker with the predicted pronunciations, and the native speaker must accept or correct the pronunciations. These results are fed back into the letter-to-sound model building process to generate a new version of the pronunciation prediction models. This technique has been shown to be effective for a number of languages, including, English, German, Thai, and Nepali.

Similar work [Davel and Barnard, 2004], [Barnard and Davel, 2004] uses the additional feedback of synthesizing the prediction pronunciations to the native speaker with a simple phoneme based synthesizer. Their experiments show that this additional feedback mechanism allows both faster and more accurate corrections from the native speaker, whether they are expert phonologists or not.



Spice Project

Rapid Acoustic Model Generation Tool

This is a tool which will display all IPA phoneme. As a naive user, you can choose and check which phoneme you'd like your Speech Engine has. After you finished, you can click the "Submit" button to create the new acoustic model on the fly.

Please input the language name (no space)

Consonants (Pulmonic): Please choose the consonant sounds you'd like to have in your new acoustic models by checking the check box

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	☐ p ☐ b			☐ t ☐ d		☐ ʈ ☐ ɖ	☐ c ☐ ɟ	☐ k ☐ g	☐ q ☐ G		☐ ʔ
Nasal	☐ m	☐ ɱ		☐ n		☐ ɳ	☐ ɲ	☐ ŋ	☐ ɴ		
Trill	☐ ʙ			☐ r					☐ ʀ		
Tap or Flap				☐ ɾ		☐ ɽ					
Fricative	☐ ɸ ☐ β	☐ f ☐ v	☐ θ ☐ ð	☐ s ☐ z	☐ ʃ ☐ ʒ	☐ ʂ ☐ ʐ	☐ ç ☐ ʝ	☐ x ☐ ɣ	☐ χ ☐ ʁ	☐ ħ ☐ ʕ	☐ h ☐ ɦ
Lateral fricative				☐ ɬ ☐ ɮ							
Approximant		☐ ʋ		☐ ɹ		☐ ɻ	☐ j	☐ ɰ			
Lateral approximant				☐ l		☐ ɭ	☐ ʎ	☐ ʟ			

Figure 1: SPICE Tools - screenshot

4. Efficiency Tools

The primary focus of SPICE (**S**peech **P**rocessing - **I**nteractive **C**reation and **E**valuation Toolkit for new Languages), a three years program sponsored by NSF, is to significantly reduce the amount of time and effort involved in building speech processing systems by providing innovative methods and tools for novices to develop speech processing models, collect appropriate data to build these models, and evaluate the results allowing iterative improvements [Schultz, 2004]. Building on the existing GlobalPhone and FestVox projects, knowledge and data will be shared between recognition and synthesis such as phoneme sets, pronunciation dictionaries, acoustic models, and text resources. User studies will indicate how well speech systems can be built, how well tools support the efforts and what must be improved to create even better systems. Archiving the data gathered on-the-fly from many cooperative users will significantly increase the repository of languages and resources. Figure 1 shows a screenshot of the acoustic model bootstrap part of this project.

The SPICE tools have so far been used in our lab to rapidly bootstrap a speech recognition system in Bulgarian and Vietnamese [Le et al., 2006], as well as for rapid development of an Afrikaans-English speech translation system [Engelbrecht and Schultz, 2005]. The reported numbers are related to the latter project and indicate how much time and effort is involved in the process of developing different components of a full speech translation system. In this case the user was not a novice but a well-trained expert in speech recognition. We differentiate between data preparation, training and tuning of components, the evaluation of the final system, and building a prototype demonstrator. The demonstrator runs on a laptop, accepts spoken speech in both languages as input and synthesizes the translated text in the corresponding language. For training and evaluation we used the Janus Speech Recognition Toolkit (JRTk) [Finke et al., 1997] for ASR, the CMU SMT decoder [Vogel et al., 2003] for MT, and the FestVox tools [Black and Lenzo, 2000] for TTS. For this development, transcribed speech data in Afrikaans was given, together with a dictionary. The English ASR and TTS were borrowed from similar tasks. As can be seen in Figure 2 the most time consuming process is the preparation of data for the purpose of training the language models (11 days for ASR and another 3 days for MT), acoustic models (5 days), translation models (2 days), and TTS (2 days). The dictionary, although given, required one unit to match existing formats and 3 days to generate letter-to-sound rules for TTS. The actual component training took 8 days for the acoustic models, and less than 1 day for language models, translation decoders, and TTS. The ASR took another 7 days for parameters tunings and settings. Finally, 5 days were used to evaluate the end-to-end system and build the demonstrator, respectively.

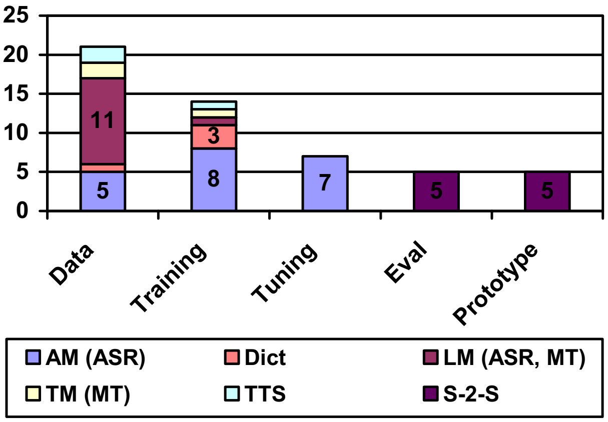


Figure 2: Development Time for Speech Translation

For TTS, we ideally need only collect a small amount of acoustic data. Standard FestVox unit selection voices, although require substantially less data than some commercial systems, still require around 1000-2000 utterances, which can be 30-60 minutes of speech. It is worth noting that speaking consistently for that amount of time and correctly reading prompts is actually a hard thing to do, especially when the scripting language is unfamiliar to the speaker. Thus such a collection will have a substantial amount of errors, and error identification methods are necessary. A more feasible technique is following the GlobalPhone technique [Schultz and Waibel, 2001]. [Latorre et al., 2005] uses HMM-based generation synthesis to build basic acoustic models from one language and adapts them to the target language. The technique still requires 10-20 minutes of transcribed speech in the target language. Theoretically this could even be from different speakers, but collection of a single speaker is likely to be better. Where an acoustic model is available for an acoustically close language, simpler voice transformation techniques can be adequate only requiring 30-50 phonetically balanced utterances.

For Iraqi we used our existing Arabic phoneme set plus /ch/ and /jh/, used in Persian (Farsi) loan words, /p/ and the pharyngealized liquid /lq/. We used an LDC provided lexicon that was automatically built from Iraqi transcripts. This provides the yellow romanized, vocalized versions of each Arabic script word. Using our letter-to-sound rules techniques we built pronunciations for addition words. The initial version of the LTS-rules provided only 35% word accuracy. We then used greedy selection techniques to find in-domain sentences that were phonetically balanced. This stage from starting to a prompt set of around 1500 sentences took three work days. The recording (done at the Defense Language Institute) took place over two days, which included hand fixing of some of the prompts to better reflect Iraqi Arabic. Then an automatic build of the voice took another three work days. This unchecked voice was then played to native speakers and we identified a number of small lexical errors which were fixed (mainly to do with distinguishing word initial and final use of some phonemes). This voice was used in our initial speech translation evaluation, where although far from perfect the synthesizer was mostly understandable. A later version of the lexicon and corrected mapping rules to our phoneme set produced a lexicon where our LTS-rules predicted a held out set at 60% accuracy. This lexicon was used to re-label the data, which in native listening tests provides a much better quality voice.

Conclusion

In the paper we have begun to identify the harder tasks of efficiently using native speakers to build speech translation models. As we support more and more languages, native speaker availability and ability to use tools efficiently will be crucial in the success of rapid portability. The SPICE project tools are our first attempt at mechanizing this process to allow a linguistically naive native speaker to build speech recognition and speech synthesis models for a new language.

Acknowledgements

This work is in part supported by the US National Science Foundation under grant number 0415021 "SPICE: Speech Processing Interactive Creation and Evaluation Toolkit for new Languages." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[Barnard and Davel, 2004] Barnard, E., and Barnard, E., *The efficient generation of pronunciation dictionaries: human factors during bootstrapping.* ICSLP, Jeju, Korea, 2004.

[Black et al., 1998] Black, A., Lenzo, K. and Pagel, V., *Issues in Building General Letter to Sound Rules*, Proc. ESCA Workshop on Speech Synthesis, pp 77-80, Jenolan Caves, Australia, 1998.

[Black and Lenzo, 2000] Black, A., and Lenzo, K., *The FestVox Project: Building Synthetic Voices*, http://festvox.org/bsv/ 2000.

[Davel and Barnard, 2004] Davel, M., and Barnard, E., *The efficient generation of pronunciation dictionaries: machine learning factors during bootstrapping*. ICSLP, Jeju, Korea, 2004.

[Engelbrecht and Schultz, 2005] Engelbrecht, H. and Schultz, T., *Rapid Development of an Afrikaans-English Speech-to-Speech Translator.* Proceedings of the IWSLT, Pittsburgh, PA, October 2005.

[Finke et al., 1997] Finke, M., Geutner, P., Hild, H., Kemp, T. Ries, K. and Westphal, M., T*he Karlsruhe Verbmobil Speech Recognition Engine.* ICASSP, Munich, Germany, 1997.

[Kominek and Black, 2005] Kominek, J. and Black, A, *Measuring Unsupervised and Acoustic Clustering through Phoneme Pair Merge-and-Split Tests* Interspeech, Lisbon, Portugal, 2005.

[Latorre et al., 2005] Latorre, J., Iwano, K., and Furui. S., *Polyglot synthesis using a micture of monolingual corpora.* ICASSP, Philadelphia, PA, 2005.

[Le et al., 2006] Le, V.B., Besacier, L., Schultz, T., *Acoustic-Phonetic Unit Similarities for Context Dependent Acoustic Model Portability.* Submitted to ICASSP 2006.

[Maskey et al., 2004] Maskey. S, Black, A. and Tomokiyo, L. *Bootstrapping Phonetic Lexicons for New Languages.* ICSLP, Jeju, Korea, 2004.

[Maamouri, 2004] Maamouri, M, Graff, D, Hubert, J, Cieri, C. and Buckwalter, T. (2004) *Dialectal Arabic Orthography-based Transcription and CTS Levantine Arabic Collection*, COLING, Geneva, Switzerland, 2004.

[Schultz and Waibel, 2001] Schultz, T. and Waibel, A., *Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition*, Speech Communication, Volume 35, Issue 1-2, pp. 31-51, August 2001.

[Schultz 2002] Schultz, T., *GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University.* ICSLP, Denver, CO, 2002.

[Schultz, 2004] Schultz, T., *Towards Rapid Language Portability of Speech Processing Systems.* Conference on Speech and Language Systems for Human Communication, Delhi, India, November 2004.

[Schultz et al., 2005] Schultz, T., Black, A., and Woszczyna, M., *Flexible Speech Translation Systems.* Special Issue in Speech Translation, IEEE Transactions of Speech and Audio Processing, Accepted for publication, August 2005.

[Vogel et al., 2003] Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., and Waibel, A., *The CMU Statistical Translation System.* Proceedings of the MT Summit IX. New Orleans, LA. September 2003.

[Waibel et al., 2004] Waibel, A., Schultz, T., Vogel, S., Fügen, C., Honal, M., Kolss, M., Reichert, J., and Stüker, S., *Towards Language Portability in Statistical Speech Translation.* ICASSP, Montreal, Canada, May 2004.