Development of Phrase Translation Systems for Handheld Computers: From Concept to Field

Horacio Franco, Jing Zheng, Kristin Precoda, Federico Cesari, Victor Abrash, Dimitra Vergyri, Anand Venkataraman, Harry Bratt, Colleen Richey, and Ace Sarich[1]

SRI International, Menlo Park CA 94025, [1]Marine Acoustics, Arlington, VA 22203
hef@speech.sri.com

Abstract

We describe the development and conceptual evolution of handheld spoken phrase translation systems, beginning with an initial unidirectional system for translation of English phrases, and later extending to a limited bidirectional phrase translation system between English and Pashto, a major language of Afghanistan. We review the challenges posed by such projects, such as the constraints imposed by the computational platform, to the limitations of the phrase translation approach when dealing with naïve respondents. We discuss our proposed solutions, in terms of architecture, algorithms, and software features, as well as some field experience by users of initial prototypes.

1. Introduction

The inspiration for a spoken phrase translation system (SPTS) can be found in traditional text-based phrasebooks, where domain-specific sets of useful phrases in the user's language are listed along with translations in the target language. These translations are usually written in some phonetic form, so that the user can produce an approximation to the target language sounds. The phrase translation systems that are the subject of this paper extend the classical phrasebook in many ways: they use voice input for selecting the desired sentence, including likely variations in wording, as well as voice output to play back an associated prerecorded translation. The advantage of using voice input is that the user's eyes are kept free, and the user does not have to search for the desired phrase if he or she approximately remembers it. Voice input may also offer greater convenience than searching through a list of phrases. The capability to give spoken output can be valuable when dealing with illiterate listeners, or when it is important to keep a physical distance between the user and the addressee. The use of prerecorded voice output for the translations also ensures that a clear, native pronunciation can be conveyed and does not require the user to produce the best approximation of the often crude written phonetic representation.

The SPTS represents one of the simplest forms of voice translation available. Most of the computational effort resides in the phrase recognition component, and the translation consists of only a look-up operation for the waveform of the corresponding prerecorded translation. The moderate computational demand and the many potential applications for such a system make it an ideal application for the current generation of high-end handheld computers.

A natural extension of an SPTS is to allow spoken responses from the addressee, enabling limited bilingual communication. However, using this capability presents special challenges for the spoken phrase translation paradigm when dealing with naïve respondents, as we will discuss in the following sections.

A phrase translation system, while clearly limited in its nature, can still be very valuable in numerous situations where users need to convey directions, commands, orders, simple instructions, and even simple questions, with limited possible answers, to non-English speakers. These uses arise in activities such as peacekeeping operations, medical triage, humanitarian assistance, medical diagnostics and treatment, fire and police departments, disaster relief, business travel, customs and immigration, the travel industry, and many others.

In this paper we report on our work on the system architecture and algorithm and software development for both a unidirectional SPTS and a bilingual SPTS having several additional capabilities. There has been complementary work on the underlying speech recognition engine [1] and on the hardware platform [2], as well as on the development of Pashto language technology [3] for the bilingual SPTS that has been reported elsewhere.

2. Unidirectional phrase translation

The main challenge in developing the first unidirectional SPTS was to achieve real-time speaker-independent recognition for a significant number -- around 700 -- of simultaneously active phrases, with reasonable accuracy, on a handheld computer. Part of the work consisted of developing and optimizing our recognizer engine for embedded and mobile systems, DynaSpeak [1]. Much of this DynaSpeak development was driven by the requirements of the phrase translation system, such as the use of integer arithmetic, support for embedded grammar tags, support for dynamically loadable grammars, and low-overhead Gaussian and search pruning, which is critical for fast decoding in large search spaces. In addition, we implemented full support for SRI's different acoustic [4] and language [5] models and a simple natural language parsing capability. To increase recognition robustness, we also implemented fast online speaker and environment adaptation based on [6] and noise compensation capabilities. Finally, a multithread-safe implementation efficiently supported running multiple instances of the recognizer with different grammars, and/or different acoustic models; this is an essential feature for speech-to-speech translation systems. Other work included application-specific research and development, such as a study of search topologies, acoustic model development for robust, fast, speaker-independent recognition, and grammar optimization tools [7].

2.1. Architecture of the unidirectional SPTS

In Fig. 1 the dotted box delimits the components that correspond only to the unidirectional SPTS. At initialization, the application loads English acoustic models into the recognition engine, and the user, either by using the GUI or by voice commands, chooses one of the available recognition grammars, usually corresponding to specific application domains. Each recognition grammar is composed of a number of finite state subgrammars, each subgrammar representing the canonical form of a sentence and likely variations in wording. Each sentence is explicitly represented by concatenating the corresponding word models in a sequential path. Alternative branches in the subgrammar allow the representation of variations. In the simplest case, each sentence subgrammar is associated with a unique tag that the recognizer passes through to the output and that is used to find a unique translation and display string. The grammars can be specified in a simple text form using the JSGF format. The tag is used to look up a corresponding translation in a set of compressed audio waveforms spoken in the desired output language. The user may select the output language from a set of available languages in a given application.

The user interface uses a push-to-talk (PTT) button to input speech into the recognizer. While the system is active, audio is continuously acquired into a circular buffer and the PTT beginning and end signals are used only to cue likely endpoints. This enables the search for the actual endpoint to start before the PTT beginning signal, so the system is robust to user synchronization errors that would otherwise result in the speech being truncated.

2.2. System features

A valuable consequence of the use of speaker-independent acoustic models is that no acoustic training for individual speakers is required to begin using the SPTS. A result of the use of current state-of-the-art large-vocabulary speech recognition technology is that the acoustic models are phonetically based, allowing the easy addition of new phrases through entering ordinary spelling. It is thus straightforward to create phrase sets for new domains with relatively simple tools. All that is needed besides text input are the recordings of the associated translated waveforms. It is also easy to add sentence variations either by defining new sentences associated with the same tag/translation, or by editing the subgrammar associated with the canonical form of a sentence.

Adding new languages is also easy, requiring only (oral) translations and recordings of the translations. There is no practical limit to the number of target languages, as each normally requires only a few megabytes and many can be stored on a flash memory card.

The system user has the choice between having the recognized phrase translated immediately, or using a verification mode, in which the recognized phrase is first displayed, and if correct, played in translation by the user pushing a second button. This verification mode can also be speech based, so the operation can be eyes free. This simple feature adds reliability to critical communications or when recognition accuracy may be degraded because of environmental conditions. It is also possible to use stylus input to select the desired sentence to translate on a scrolling display window.

3. Bilingual phrase translation

The next step in the development of phrase translation systems was to allow limited bilingual communication. In principle this can be implemented with two unidirectional systems, one for each language. Unfortunately, simply putting together two unidirectional SPTSs would not work for untrained respondents, as they may not be aware of the allowable phrase answers. Furthermore, a naïve respondent may produce unexpected but critical speech input. Our goal was to meet challenges without greatly increasing the computational demands, so that the same hardware could support the new, bidirectional system. One approach to handling the speech of the naïve respondent is to reduce recognition perplexity by using question-dependent recognition grammars for the respondent; that is, at each point in the dialog, the last English question determines which Pashto grammar will be active during the respondent's turn. The English questions are also designed to constrain respondent answers by using careful wording, and even suggesting possible alternative answers as part of the question. To deal with unexpected input, which by definition cannot be captured by the question-dependent grammars, we implemented a wordspotting mechanism that can be used either to cue the system user about certain critical information or to help the system user reengage in the kind of directed dialog that the system is capable of handling.

3.1. Architecture of the bilingual SPTS

In Fig. 1 we show the different components of the bilingual SPTS. The system consists of two similar SPTSs, one for English and one for a target language, in this case, Pashto. The user interface allows the operator to direct input speech to either recognizer. The English subsystem is very similar to the unidirectional SPTS discussed above, with the addition that the identity of the recognized English sentence is used by the grammar selection block of the Pashto SPTS. The Pashto SPTS uses this information to load the response grammar for the Pashto recognizer that matches the last question from the English speaker. In other words, the Pashto recognition grammar at any time is dependent on the previous English question. This feature is essential to reducing the recognition perplexity in the Pashto system to only a few semantically distinct alternatives.

In our approach, respondent answers fit into 1-of-N synonymy classes, where each synonymy class is a set of answers having the same meaning. Each synonymy class is modeled by a subgrammar that represents most of the expected variability in the responses of naïve speakers to a given question. Each synonymy class is linked to a "canonical translation" that conveys the meaning back into English. Note that the synonymy classes can be represented by quite large grammars to encode syntactic, lexical, and dialectal variations of answers with the same meaning. This complexity of the synonymy class subgrammars is compensated for by the fact that only a few of the synonymy class subgrammars are active at a given time, because of the question-dependent architecture for the Pashto recognizer.

3.2. Structured value sentences

A particular type of input, structured value sentences, which convey information about quantities, time, order, and so on, is handled by parsing the input to obtain a language-independent

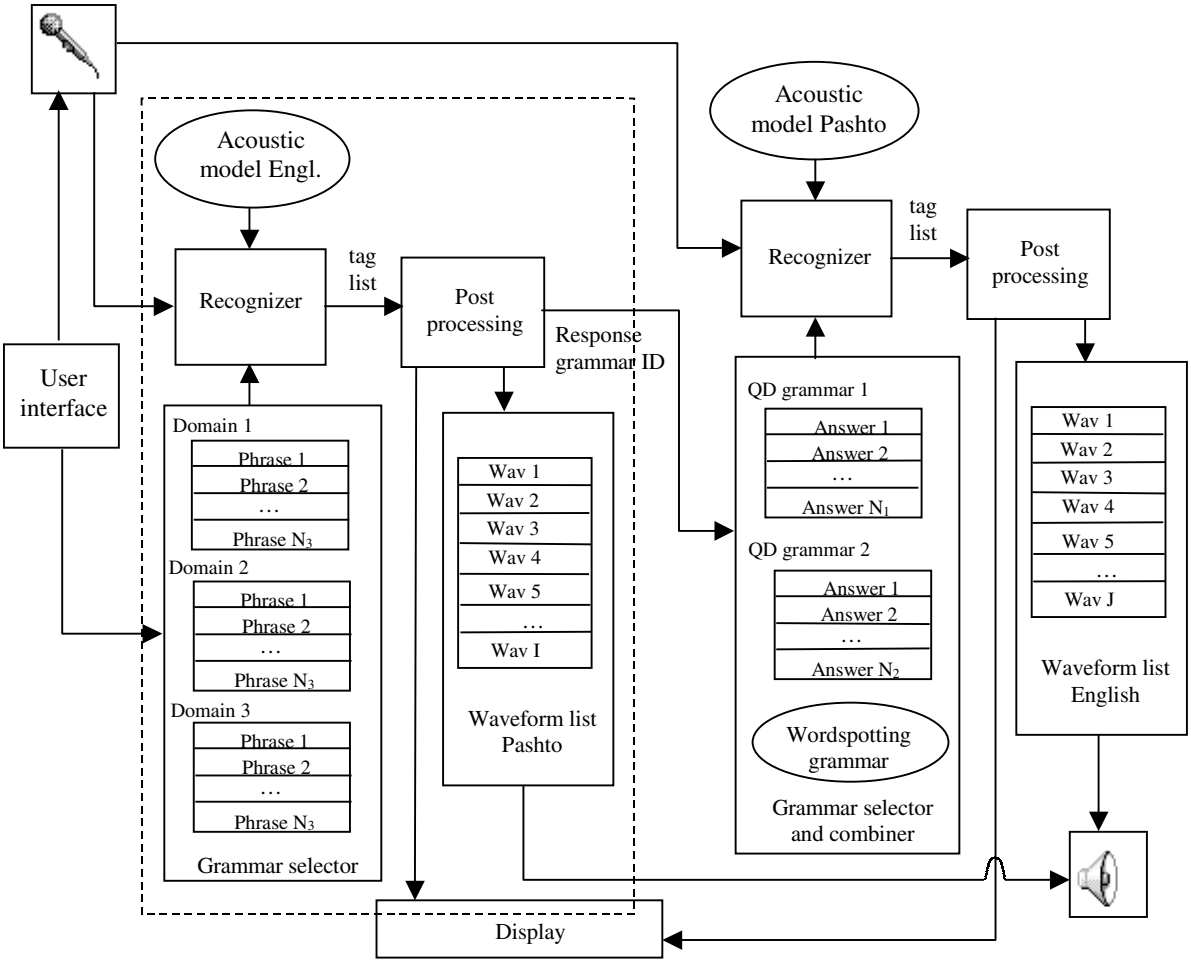

Fig. 1: Bilingual SPTS. The user interface controls the audio gating to the English or the Pashto recognizer; it also can select, by voice or stylus, a domain grammar from those available. The English recognizer output is used to select the translation waveform/s to playback, the display string, and the recognition grammar for the Pashto recognizer. The question-dependent grammar is combined with the wordspotting grammar in the Pashto recognizer. The Pashto recognizer output, a single tag, or a tag list for value structured sentences, is processed to produce the Pashto waveforms play list and the display string corresponding to the canonical translation. The dotted box delimits the unidirectional SPTS components.

representation of the value embedded in the sentence, and then generating the translation by rule from a series of prerecorded waveforms that are dynamically concatenated. We implemented support for a mechanism by which the structured values can be embedded in the recognition grammars as well as in the definition of the canonical translations. We used DynaSpeak's capability to output tags associated with subgrammars to encode in these tags relevant information. We scan the output of the recognizer for these tags and process them to extract the correct values in the right places. The associated audio translation is formed by concatenating prerecorded waveforms according to a rule-generated waveform list associated with the decoded value. We added additional functionality to handle tense and number agreement between the generated components. Both recognizers can use this capability.

3.3. Wordspotting capability

A significant capability was the capacity to handle important but unexpected input, or "hot words", from the respondent. To address this need, we added wordspotting. Our approach to wordspotting attempts to capture most of the features of the high-performance wordspotters based on continuous speech recognition technology [8], such as using statistical language models and explicit acoustic modeling of context words leading to the "hot words". At the same time, we limit the size of the total wordspotting model to operate in real time on the handheld hardware. In our implementation we used a class-based bigram language model composed of "hot words", likely context words surrounding the hot words, and a general filler model to represent any other speech not explicitly modeled. The wordspotting grammar is combined with the active question dependent grammar at a given time.

The final stage of the wordspotting algorithm is based on word confidence estimation derived from a posterior probability weighted word graph derived from N-best decoding. Our experiments showed that this approach had five times fewer false alarms than classical wordspotting methods using only hot word and filler acoustic models.

3.4. Additional features

A confirmation feature was implemented for the Pashto speaker, which plays back the Pashto waveform meaning "Did you mean …" concatenated with a canonical Pashto waveform (or waveforms in the case of structured value sentences) associated with the recognized Pashto synonymy class. This feature is particularly important for languages like Pashto where there is no widespread use of written language.

Confidence scores, a byproduct of the wordspotting mechanism, can also be presented to the user to help determine when the recognition accuracy may be low, and the confirmation mechanism engaged.

4. Field experience

While still being improved, the unidirectional SPTS has been deployed and used in several real-world situations allowing the collection of valuable feedback. While limited, the concept of phrase translation seems to fill many relevant needs, mostly when the needed communication is asymmetric. Some users have reported that the SPTS provided a great boost to the capability to function with foreign nationals and communicate the user's intent to target audiences.

When just introduced to the system, people often start out by assuming that the device can translate spontaneous speech, but learn very quickly to adapt to the limitations of the SPTS approach. It is not surprising that users have often expressed a desire for bidirectional input.

The design of the phrase lists and organization of the phrases into meaningful groups seems to be very important. Some users have said that they would like to be able to organize the phrases themselves for maximum individual convenience, or just to group phrases into subdomains that users can interpret and manage easily. Being able to navigate a meaningful hierarchy of topics seems to make a big difference to how easy or hard it is to find a desired phrase.

A number of field users have taken advantage of the provided tools to build new domains or translate old domains into new languages. The capability supported by these tools appears to be highly appreciated and very valuable in rapidly changing situations.

Overall, perhaps the most important factor in the perceived utility of the unidirectional SPTS is whether the desired application is well matched to its capabilities and limitations. It cannot be overemphasized that the SPTS is not the solution to the overall language barrier problem; nevertheless, it appears to be a viable solution to parts of the problem.

5. Summary and discussion

We have described the system architecture and the underlying algorithms used to develop the software for a unidirectional spoken phrase translation system that would run on a handheld computer. The architecture was extended to accept constrained responses from untrained non-English speakers, as well as to support structured value sentences and wordspotting capabilities. Feedback from initial deployments of the SPTS seems to indicate that the approach fills the need for communication in many important situations, especially when the user wants to convey information in specific semantic domains and the interaction is rather asymmetrical; that is, the English speaker mostly directs the dialog by providing information, giving directions and orders, or by asking questions with simple answers. The capability for rapid prototyping of application packages for new domains and for new target languages is perhaps one of the most salient characteristics of the SPTS approach, providing a very valuable tool to enable basic, essential communication in situations where international teams need to provide disaster relief, humanitarian assistance, or other kinds of services, and a fast response is essential.

Acknowledgments

The work on the unidirectional SPTS was funded by a Defense Advanced Research Projects Agency (DARPA) Small Business Innovative Research grant to Marine Acoustics, Inc. The work on the bidirectional SPTS was funded by a DARPA contract under the DARPA Babylon program and by the Advanced Concepts Technology Demonstration program, Language and Speech Exploitation Resources program.

References

[1] Franco, H., Zheng, J., Butzberger, J., Cesari, F., Frandsen, M., Arnold, J., Gadde, V. R. R., Stolcke, A., and Abrash, V., DynaSpeak: SRI's Scalable Speech Recognizer for Embedded and Mobile Systems. *Proc. Human Language Technology Conference*, San Diego, CA, pp. 25-30, 2002

[2] http://www.phraselator.com

[3] Precoda, K., Non-mainstream Languages and Speech Recognition: Some Challenges, to appear in *CALICO Journal*, vol. 21, 2003.

[4] Digalakis, V., Monaco, P., and Murveit, H., Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers. *IEEE Trans. Speech and Audio Processing,* 4, pp. 281-289, 1996.

[5] Stolcke, A., SRILM – An Extensible Language Modeling Toolkit, *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, Denver, CO, pp. 901-904, 2002.

[6] Digalakis, V., and Neumeyer, L., Fast Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures. *Proc. ARPA Spoken Language Systems Technology Workshop,* Austin, TX, 1995.

[7] Zheng, J., and Franco, H., Fast Hierarchical Grammar Optimization Algorithm Towards Time and Space Efficiency. *Proc. ICSLP'2002,* Denver CO, 2002.

[8] Weintraub, M., LVCSR Log-likelihood Ratio Scoring for Keyword Spotting, *Proc. IEEE Intl. Conf. on Speech and Signal Processing*, vol. 1, Detroit, pp. 297–300, 1995.