The VoiceTRAN Speech-to-Speech Translation Communicator

JERNEJA ŽGANEC GROS, MARIO ŽGANEC
Alpineon Research and Development Division
Alpineon RTD
Ulica Iga Grudna 15, SI-1000 Ljubljana
SLOVENIA
http://www.alpineon.com

Abstract: - This paper describes the design phases of the VoiceTRAN Communicator, which integrates speech recognition, machine translation, and text-to-speech synthesis using the Galaxy architecture. The aim of the work was to build a robust multimodal speech-to-speech translation system able to translate simple domain-specific sentences in the language pair Slovenian-English.
The work represents a joint collaboration between several Slovenian research organizations that are active in human language technologies.

Key-Words: - Speech-to-speech translation, Speech recognition, Machine translation, Speech synthesis

1 Introduction

Automatic speech-to-speech (STS) translation systems aim to facilitate communication among people that speak different languages [1, 2, 3]. Their goal is to generate a speech signal in the target language that conveys the linguistic information contained in the speech signal from the source language.

There are, however, major open research issues that challenge the deployment of natural and unconstrained speech-to-speech translation systems, even for very restricted application domains, due to the fact that state-of-the-art automatic speech recognition and machine translation systems are far from perfect.

In addition, in comparison to translating written text, conversational spoken messages are often conveyed with imperfect syntax and casual spontaneous speech.

In practice, when building demonstration systems, STS systems are typically implemented by imposing strong constraints on the application domain and the type and structure of possible utterances; that is, both in the range and in the scope of the user input allowed at any point of the interaction. Consequently, this compromises the flexibility and naturalness of using the system.

The VoiceTRAN Communicator was developed in a Slovenian research project involving 6 partners: Alpineon, the University of Ljubljana (Faculty of Electrical Engineering, Faculty of Arts, and Faculty of Social Studies), the Jožef Stefan Institute, and Amebis as a subcontractor.

The work is co-funded by the Slovenian Ministry of Defense and the Slovenian Research Agency. The aim is to build a robust multimodal speech-to-speech translation communicator, similar to Phraselator [4] or Speechalator [5], able to translate simple sentences in the language pair Slovenian-English. It goes beyond the Phraselator device because it is not limited to predefined input sentences.

The application domain is limited to common application scenarios that occur in peace-keeping operations on foreign missions when the users of the system have to communicate with the local population. More complex phrases can be entered via keyboard using a graphical user interface.

First an overview of the VoiceTRAN system architecture is given. We continue to describe the individual server modules. We conclude the paper by discussing the speech-to-speech translation evaluation methods and outlining plans for future work.

2 System Architecture

The VoiceTRAN Communicator uses the DARPA Galaxy Communicator architecture [6]. The Galaxy Communicator open source architecture was chosen to provide inter-module communication support because its plug-and-play approach allows interoperability of commercial software and research software components. It was specially designed for development of voice-driven user interfaces in a multimodal platform.

The VoiceTRAN Communicator consists of a Hub and 5 servers that interact with each other through the Hub as shown in Figure 1:

Audio Server	Receives speech signals from the microphone and sends them to the recognizer. Sends synthesized speech to the speakers.
Graphic User Interface	Receives input text from the keyboard. Displays recognized source language

	sentences and translated target language sentences. Provides user controls for handling the application.
Speech Recognizer	Takes the signals from audio server and maps audio samples into text strings. Produces an N-best sentence hypothesis list.
Machine Translator	Receives N-best postprocessed sentence hypotheses from the speech recognition server and translates them from a source language into a target language. Produces a scored disambiguated sentence hypothesis list.
Speech Synthesizer	Receives rich and disambiguated word strings from the machine translation server. Converts the input word strings into speech and prepares them for the audio server.

The Hub is used as a centralized message router through which servers can communicate with one another. Frames containing keys and values are emitted by each server. They are routed by the hub and received by a secondary server based on rules defined in the Hub script.

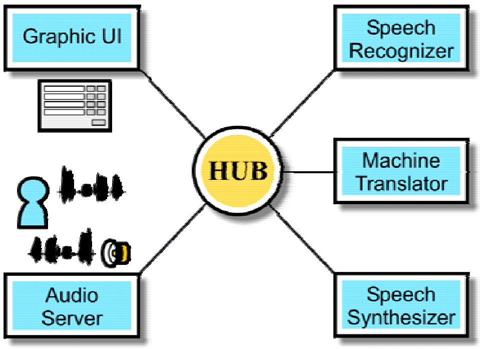

Fig. 1 The Galaxy system architecture used in the VoiceTRAN communicator.

2.1 Audio Server

The audio server connects to the microphone input and speaker output terminals on the host computer and performs recoding of user input and playing prompts or synthesized speech.

Input speech captured by the audio server is automatically recorded to files for posterior system training.

2.2 Speech Recognizer

The speech recognition server receives the input audio stream from the audio server and provides a word graph at its output and a ranked list of candidate sentences; the N-best hypotheses list, which can include part-of-speech information generated by the language model.

The speech recognition server used in VoiceTRAN is based on a Hidden Markov Model Recognizer [7]. It has been upgraded to perform large vocabulary speaker (in)dependent speech recognition on a wider application domain. A back-off class-based trigram language model is used.

Because the final goal was a stand-alone speech communicator used by a specific user, the speech recognizer has been additionally trained and adapted to the individual user in order to achieve higher recognition accuracy in at least one language pair direction.

A common speech recognizer output typically has no information on sentence boundaries, punctuation, and capitalization. Therefore, additional postprocessing in terms of punctuation and capitalization has been performed on the N-best hypotheses list before it is passed to the machine translator. Prosody information helps to determine proper punctuation and sentence accent information.

2.3 Machine Translator

The machine translator (MT) converts text strings from a source language into text strings in the target language. Its task is difficult since the results of the speech recognizer convey spontaneous speech patterns and are often erroneous or ill-formed.

A postprocessing algorithm inserts basic punctuation and capitalization information before passing the target sentence to the speech synthesizer. The output string can also convey lexical stress information in order reduce disambiguation efforts during text-to-speech synthesis.

A multi-engine based approach was used in the early phase of the project that makes it possible to exploit strengths and weaknesses of different MT technologies and to choose the most appropriate engine or combination of engines for the given task. Four different translation engines have been applied in the system. We combined TM (translation memories), SMT (statistical machine translation), EBMT (example-based machine translation) and RBMT (rule-based machine translation) methods. A simple approach to select the best translation from all the outputs was applied.

A bilingual aligned domain-specific corpus was used to build the TM and train the EBMT and the SMT phrase translation models. In SMT an interlingua approach, was investigated and promising directions pointed out in [8] have been be pursued.

The Presis translation system was used as our baseline system [9]. It is a commercial conventional rule-based translation system that is constantly being optimized and upgraded. It was adapted to the application domain by upgrading the lexicon. Based on stored rules, Presis parses each sentence in the source language into grammatical components, such as subject, verb, object and predicate and attributes the relevant semantic categories. Then it uses built-in rules for converting these basic components into the target language, performs regrouping and generates the output sentence in the target language.

2.4 Speech Synthesizer

The last part in a speech-to-speech translation task is the conversion of the translated utterance into its spoken equivalent. The input target text sentence is equipped with lexical stress information at possible ambiguous words.

The AlpSynth unit-selection text-to-speech system is used for this purpose [10]. It performs grapheme-to-phoneme conversion based on rules and a look-up dictionary and rule-based prosody modeling. Domain-specific adaptations include new pronunciation lexica and the construction of a speech corpus of frequently used in-domain phrases.

Special attention was paid to collocations as defined in the bilingual dictionary. They were treated as preferred units in the unit selection algorithm.

2.5 Graphical User Interface

In addition to the speech user interface, the VoiceTRAN Communicator provides a simple interactive user-friendly graphical user interface where input text in the source language can also be entered via keyboard or selected by pen input.

Recognized sentences in the source language along with their translated counterparts in the target language are displayed.

A push-to-talk button is provided to signal an input voice activity, and a replay button serves to start a replay of the synthesized translated utterance. The translation direction can be changed by pressing the translation direction button.

3 Language Resources

Some of the multilingual language resources needed to set up STTS systems and include Slovenian are presented in [11].

For building the speech components of the VoiceTRAN system, existing speech corpora have been used [12]. The language model has been trained on a domain-specific text corpus that was collected and annotated within the project.

The AlpSynth pronunciation lexicon [10] has been used for both speech recognition and text-to-speech synthesis. Speech synthesis is based on the AlpSynth speech corpus. It has been expanded by the most frequent in-domain utterances.

For developing the initial machine translation component, a dictionary of military terminology [13] and various existing aligned parallel corpora were used [14, 15, 16].

3.1 Data Collection

We have syntactically annotated an in-domain large Slovenian monolingual text corpus that was collected at the Faculty of Social Studies, University of Ljubljana. This corpus has been used for training the language model in the speech recognizer, as well as for inducing relevant multiword units (collocations, phrases, and terms) for the domain.

Within VoiceTRAN, an aligned bi-lingual in-domain corpus is also being collected. It consists of general and scenario-specific in-domain sentences. The compilation of such corpora involves selecting and obtaining the digital original of the bi-texts, re-coding to XML TEI P4, sentence alignment, word-level syntactic tagging, and lemmatization [17].

The corpus has been used to induce bi-lingual single word and phrase lexica for the MT component, and as direct input for SMT and EBMT systems. It was also used for additional training of the speech recognizer language model.

4 Evaluation

The evaluation tests of a speech-to-speech translation system were designed on order to serve two purposes:

1. to evaluate whether we have improved the system by introducing improvement of individual components of the system [18, 19, 20];

2. to test the system acceptance by the end users in field tests [21, 22].

A detailed description of the VoiceTRAN communicator STS system evaluation tests is provided in [23].

5 Conclusion

We presented the VoiceTRAN multimodal speech-to-speech translation communicator, which is able to translate simple domain-specific sentences in the language pair Slovenian-English.

The concept of the VoiceTRAN Communicator implementation is discussed in this paper. The chosen system architecture makes it possible to test a variety of server modules. The end-to-end prototype is ready for end user field trials in military exercises.

6 Acknowledgements

The authors wish to thank the Slovenian Ministry of Defense and the Slovenian Research Agency for co-funding this work under contract no. M2-0019.

References:

[1] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavaldà, T. Zeppenfeld, and P. Zhan, "Janus-III: Speech-to-Speech Translation in Multiple Languages," Proceedings of the ICASSP, Munich, Germany, 1997, pp. 99–102.

[2] W. Wahlster, Verbmobil: *Foundation of Speech-to-Speech translation*, Springer Verlag, 2000.

[3] A. Lavie, F. Metze, R. Cattoni, E. Costantin, S. Burger, D. Gates, C. Langley, K. Laskowski, L. Levin, K. Peterson, T. Schultz, A. Waibel, D. Wallace, J. McDonough, H. Soltau, G. Lazzari, N. Mana, F. Pianesi, E. Pianta, L. Besacier, H. Blanchon, D. Vaufreydaz, and L. Taddei, "A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System," Proceedings of the ACL 2002 Workshop on Speech-to-Speech Translation: Algorithms and Systems, Philadelphia, PA, 2002.

[4] A. Sarich, "Phraselator, one-way speech translation system," available at http://www.sarich.com/ translator/, 2001.

[5] A. Waibel, A. Badran, A.W. Black, R. Frederking, D. Gates, A. Lavie, L. Levin, K. Lenzo, L. Mayfield, L. Tomokyo, J. Reichert, T. Schultz, D. Wallace, M. Woscsyna, and J. Zhang, "Speechalator: Two-Way Speech-to-Speech Translation on a Consumer PDA," Proceedings of the Eurospeech'03. Geneva, Switzerland, 2003, pp. 369–372.

[6] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, P. and V. Zue, "Galaxy-II: A Reference Architecture for Conversational System Development," Proceedings of the ICSLP'98, Sydney, Australia, pp. 931–934, available at http://communicator. ourceforge.net/, 1998.

[7] S. Dobrišek, "Analysis and Recognition of Phrases in Speech Signals," PhD Dissertation, University of Ljubljana, Slovenia, 2001.

[8] H. Ney, "The Statistical Approach to Spoken Language Translation," Proceedings of the International Workshop on Spoken Language Translation, Kyoto, 2004, pp. 15–16.

[9] M. Romih and P. Holozan, "A Slovenian-English Translation System," Proceedings of the 3rd Language Technologies Conference, Ljubljana, Slovenia, 2002, p. 167.

[10] J. Žganec Gros, A. Mihelič, M. Žganec, N. Pavešić, F. Mihelič, and V. Cvetko Orešnik, "AlpSynth Corpus-Driven Slovenian Text-to-Speech Synthesis: Designing the Speech Corpus," Proceedings of the Joint Conferences CTS+CIS. Computers in Technical Systems, Intelligent systems, Rijeka, 2004, pp. 107–110.

[11] J. Žganec Gros, F. Mihelič, Š. Vintar, and T. Erjavec, "The Voicetran Speech-to-Speech Communicator," *Lecture Notes in Compueter Science: Text, Speech and Dialog*ue, Editors: Václav Matoušek, Pavel Mautner, Tomáš Pavelka, 2005, pp. 379–385.

[12] F. Mihelič, J. Žganec Gros, S. Dobrišek, J. Žibert, and N. Pavešić, "Spoken Language Resources at LUKS of the University of Ljubljana," *Int. Journal on Speech Technologies*, Vol. 6., No. 3, 2003, pp. 221–232.

[13] T. Korošec, "Opravljeno je bilo pomembno slovarsko delo o vojaškem jeziku," *Slovenska vojska*, Vol. 10, No. 10, 2002, pp. 12–13.

[14] T. Erjavec, "The IJS-ELAN Slovene-English Parallel Corpus," *International Journal on Corpus Linguistics*, Vol. 7, 2002, pp. 1–20.

[15] Erjavec, T. "MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora," Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC'04, Lisbon, Portugal, 2004, pp. 1535–1538.

[16] T. Erjavec, C. Ignat, P. Pouliquen, and R. Steinberger, "Massive Multi-lingual Corpus Compilation: Acquis Communautaire and Totale," Proceedings of the 2nd Language and Technology Conference, Poznań, Poland, 2005.

[17] T. Erjavec and S. Džeroski, "Machine Learning of Language Structure: Lemmatising Unknown Slovene Words," *Applied Artificial Intelligence*, Vol. 18, No. 1, 2004, pp. 17–41.

[18] S. Rossato, H. Blanchon, and L. Besacier, "Speech-to-Speech Translation System Evaluation: Results for French for the Nespole! Project First Showcase," Proceedings of the ICSLP. Denver, CO, 2004.

[19] MT Evaluation Kit. "NIST MT Evaluation Kit Version 11a," available at http://www.nist.gov /speech/ tests/mt, 2002.

[20] Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsuji, "Overview of the IWSLT04 Evaluation Campaign," Proceedings of the

International Workshop on Spoken Language Translation, Kyoto, Japan, 2004, pp. 1-9.

[21] R. Eklund, B. Lyberg, "Inclusion of a Prosodic Module in Spoken Language Translation Systems," In Proceedings of the ASA 130th Meeting, St. Louis, MO, 1995.

[22] F. Lefevre, J.L. Gauvain, L. Lamel, "Improving Genericity for Task-Independent Speech Recognition," In Proceedings of the Eurospeech'01, Aalborg, Denmark, 2001, pp. 1241-1244.

[23] J. Žganec Gros, M. Žganec, "A Slovenian-english speech-to-speech translation system," *WSEAS Transactions on Systems*, 2006, accepted for publication.