Proc. of the 8th Intl. Conf. on Text, Speech and Dialogue, TDS 2005, Karlovy Vary, Czech Republic, 2005.
Springer LNAI 3658, pp. 379-384

The VoiceTRAN Speech-to-Speech Communicator

Jerneja Žganec Gros[1], France Mihelič[2], Tomaž Erjavec[3], and Špela Vintar[2]

[1] Alpineon d.o.o., Ulica Iga Grudna 15,
SI-1000 Ljubljana, Slovenia
jerneja@alpineon.com,
http://www.alpineon.com
[2] University of Ljubljana,
SI-1000 Ljubljana, Slovenia
france.mihelic@fe.uni-lj.si,
spela.vintar@guest.arnes.si
[3] Jožef Stefan Insitute, Jamova 39,
SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract. The paper presents the design concept of the VoiceTRAN Communicator that integrates speech recognition, machine translation and text-to-speech synthesis using the DARPA Galaxy architecture. The aim of the project is to build a robust speech-to-speech translation communicator able to translate simple domain-specific sentences in the Slovenian-English language pair. The project represents a joint collaboration between several Slovenian research organizations that are active in human language technologies. We provide an overview of the task, describe the system architecture and individual servers. Further we describe the language resources that will be used and developed within the project. We conclude the paper with plans for evaluation of the VoiceTRAN Communicator.

1 Introduction

Automatic speech-to-speech (STS) translation systems aim to facilitate communication among people who speak in 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. Additionally, 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

Proc. of the 8th Intl. Conf. on Text, Speech and Dialogue, TDS 2005, Karlovy Vary, Czech Republic, 2005.
Springer LNAI 3658, pp. 379-384

type and structure of possible utterances, i.e. 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 is being built within a national Slovenian research project involving 5 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 project is cofunded by the Slovenian Ministry of Defense. The aim of the project is to build a robust speech-to-speech translation communicator, similar to Phraselator [4] or Speechalator [5], able to translate simple sentences in a Slovenian-English language pair.

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.

2 System Architecture

The VoiceTRAN Communicator uses the DARPA Galaxy Communicator architecture [6]. The Galaxy Communicator open source architecture was chosen to provide intermodule communication support as its plug-and-play approach allows interoperability of commercial software and research software components.

The VoiceTRAN Communicator consists of the Hub and five servers: audio server, graphic user interface, speech recognizer, machine translator and speech synthesizer (Fig. 1).

There are two ways of porting modules into the Galaxy architecture: the first is to alter its code so that it can be incorporated into the Galaxy architecture; the second is to create a wrapper or a capsule for the existing module, the capsule then behaves as a Galaxy server. We have opted for the second option since we want to be able to test commercial modules as well. Minimal changes to the existing modules were required, mainly those regarding input/output processing.

A particular session is initiated by a user either through interaction with a graphical user interface (typed input) or the microphone. The VoiceTRAN Communicator servers capture spoken or typed input from the user, and return the servers' responses with synthetic speech, graphics, and text. The server modules are described in more detail in the next subsections.

2.1 Audio server

The audio server connects to the microphone input and speaker output terminals on the host computer and performs recoding user input and playing prompts or synthesized speech. Input speech captured by the audio server is automatically recorded to files for later system training.

Proc. of the 8th Intl. Conf. on Text, Speech and Dialogue, TDS 2005, Karlovy Vary, Czech Republic, 2005.
Springer LNAI 3658, pp. 379-384

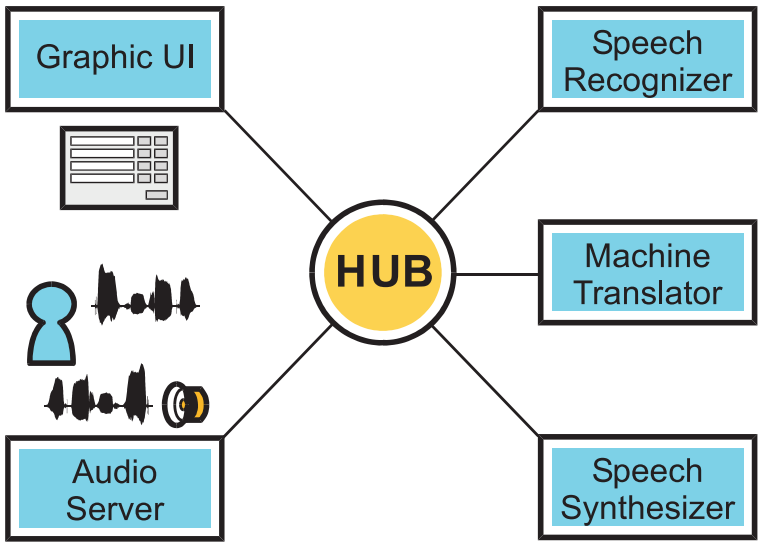

Fig. 1. The VoiceTRAN system architecture.

2.2 Speech Recognizer

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

The speech recognition server used in VoiceTRAN is based on the Hidden Markov Model Recognizer developed by the University of Ljubljana [7]. It will be upgraded to perform large vocabulary speaker (in)dependent speech recognition on a wider application domain. A bigram or a back-off class-based trigram language model will be used. Given a limited amount of training data the parameters in the models will be carefully chosen in order to achieve maximum performance.

Further in the project we want to test other speech recognition approaches with an emphasis on robustness, processing time, footprint and memory requirements.

Since the final goal of the project is a stand-alone speech communicator used by a specific user, the speech recognizer can be additionally trained and adapted to the user of the device in order to achieve higher recognition accuracy at least in one language.

A common speech recognizer output typically has no information on sentence boundaries, punctuation and capitalization. Therefore, additional postprocessing in terms of punctuation and capitalization will be performed on the N-best hypotheses list before it is passed to the machine translator. The inclusion of a

Proc. of the 8th Intl. Conf. on Text, Speech and Dialogue, TDS 2005, Karlovy Vary, Czech Republic, 2005.
Springer LNAI 3658, pp. 379-384

prosodic module will be investigated in order to link the source language to the target language, but also to enhance speech recognition proper.

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 will be 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 appropiate engine or combination of engines for the given task. Four different translation engines will be applied in the system. We will combine 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 will be applied.

A bilingual aligned domain-specific corpus will be used to build the TM and train the EBMT and the SMT phrase translation models. In SMT an interlingua approach, similar to the one described in [3] will be investigated and promissing directions pointed out in [8] will be pursued.

The Presis translation system will be used as our baseline system [9]. It is a commercial conventional rule-based translation system that is constantly being optimized and upgraded. It will be 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. It will be further upgraded within the project towards better naturalness of the resulting synthetic speech. Domain-specific adaptations will include new pronunciation lexica and the construction of a speech corpus of frequently used in-domain phrases. Other commercial off-the-shelf products will be tested as well.

Proc. of the 8th Intl. Conf. on Text, Speech and Dialogue, TDS 2005, Karlovy Vary, Czech Republic, 2005.
Springer LNAI 3658, pp. 379-384

We will also explore how to pass a richer structure from the machine translator to the speech synthesizer. An input structure containing information on POS and lexical stress information resolves many ambiguities and can result in more accurate prosody prediction.

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 a keyboard.

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, 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

For building the speech components of the VoiceTRAN system, existing speech corpora will be used [11]. The language model will be trained on a domain-specific text corpus that is being collected and annotated within the project. The AlpSynth pronunciation lexicon [10] will be used for both speech recognition and text-to-speech synthesis. Speech synthesis will be based on the AlpSynth speech corpus. It will be expanded by the most frequent in-domain utterances. For developing the initial machine translation component, the dictionary of military terminology and various existing aligned parallel corpora will be used [12].

3.1 Data Collection Efforts

The VoiceTRAN team will participate in the annotation of an in-domain large Slovenian monolingual text corpus that is being collected at the Faculty of Social Studies, University of Ljubljana. This corpus will be 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 bilingual in-domain corpus is also being collected. It will consist of general and scenario-specific in-domain sentences. The compilation of such corpora involves selecting and obtaining the digital original of the bi-texts, recoding to XML TEI P4, sentence alignment, word-level syntactic tagging and lemmatisation [13]. Such pre-processed corpora are then used to induce bi-lingual single word and phrase lexica for the MT component, or as direct inputs for SMT and EBMT systems. They will also serve for additional training of the speech recognizer language model.

Proc. of the 8th Intl. Conf. on Text, Speech and Dialogue, TDS 2005, Karlovy Vary, Czech Republic, 2005.
Springer LNAI 3658, pp. 379-384

4 Planned Evaluation

Evaluation efforts within the VoiceTRAN project will serve for two purposes: to evaluate whether we have improved the system by introducing improvement of individual components of the system; and to test the system acceptancy by potential users in field tests.

We intend to perform end-to-end translation quality tests both on manually transcribed and automatic speech recognition input. Human graders will asses the end-to-end translation performance evaluating how much of the user input information has been conveyed to the target language and also how well formed the target sentences are. Back-translation evaluation experiments involving paraphrases will be considered, as well.

We will also perform individual component tests in order to select the most appropriate methods for each application server. Speech recognition will be evaluated by computing standard word error rates (WER). For the machine translation component subjective evaluation tests in terms of fluency and adequacy are planned, as well as objective evaluation tests [14], having in mind that objective evaluation methods evaluate the translation quality in terms of the capacity of the system to mimick the reference text.

5 Conclusion

The VoiceTRAN project provides an attempt to build a robust speech-to-speech translation communicator able to translate simple domain-specific sentences in a Slovenian-English language pair. The concept of the VoiceTRAN Communicator implementation is discussed in the paper. The chosen system architecture allows for testing a variety of server modules.

6 Acknowledgements

The authors of the paper thank the Slovenian Ministry of Defense and the Slovenian Ministry of Higher Education, Science and Technology for co-funding the project.

References

1. Lavie, A., Waibel, A., Levin, L., Finke, M., Gates, D., Gavalda, M., Zeppenfeld, T., Zhan, P.: Janus-III: Speech-to-Speech Translation in Multiple Languages. In Proceedings of the ICASSP, Munich, Germany (1997) (99-102)
2. Wahlster, W.: Verbmobil: Foundation of Speech-to-Speech translation. Springer Verlag (2000)
3. Lavie, A., Metze, F., Cattoni, R., Costantin, E.,, Burger, S., Gates, D., Langley, C., Laskowski, K., Levin, L., Peterson, K., Schultz, T., Waibel A., Wallace, D., McDonough, J., Soltau, H., Lazzari, G., Mana, N., Pianesi, F., Pianta, E., Besacier, L., Blanchon, H., Vaufreydaz, D., Taddei, L.: A Multi-Perspective Evaluation of the

Proc. of the 8th Intl. Conf. on Text, Speech and Dialogue, TDS 2005, Karlovy Vary, Czech Republic, 2005.
Springer LNAI 3658, pp. 379-384

NESPOLE! Speech-to-Speech Translation System. In Proceedings of the ACL 2002 workshop on Speech-to-speech Translation: Algorithms and Systems, Philadelphia, PA (2002)
4. Sarich, A.: Phraselator, one-way speech translation system. (2001) Available at http://www.sarich.com/translator/
5. Waibel, A., Badran, A., Black, A. W., Frederking, R., Gates, D., Lavie, A., Levin, L., Lenzo, K., Mayfield Tomokyo, L., Reichert, J., Schultz, T., Wallace, D., Woscsyna, M., Zhang, J. Speechalator: Two-Way Speech-to-Speech Translation on a Consumer PDA. In Proceedings of the Eurospeech, Geneva, Switzerland. (2003) pp. 369-372
6. Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., Zue, V.: Galaxy-II: A Reference Architecture for Conversational System Development. In Proceedings of the ICSLP, Sydney, Australia (1998) pp. 931-934
7. Dobrišek, S.: Analysis and Recognition of Phrases in Speech Signals. PhD Thesis, University of Ljubljana, Slovenia. (2001)
8. Ney, H.: The Statistical Approach to Spoken Language Translation. In Proceedings of the International Workshop on Spoken Language Translation, Kyoto, Japan (2004) pp. XV-XVI
9. Romih, M., Holozan, P.: Slovensko-angleški prevajalni sistem (A Slovene-English Translation System). In Proceedings of the 3rd Language Technologies Conference, Ljubljana, Slovenia (2002) p. 167
10. Žganec Gros, J., Mihelič, A., Žganec, M., Pavešić, N., Mihelič, F., Cvetko Orešnik, V.: AlpSynth corpus-driven Slovenian text-to-speech synthesis : designing the speech corpus. In Proceedings of the joint conferences CTS+CIS, Rijeka, Croatia (2004), pp. 107-110
11. Mihelič, F., Gros, J., Dobrišek, S., Žibert, J., Pavešić, N.: Spoken language resources at LUKS of the University of Ljubljana. International Journal on Speech Technologies, Vol. 6. No. 3 (2003) pp. 221-232
12. Erjavec, T.: The IJS-ELAN Slovene-English parallel corpus. International Journal on Corpus Linguistics, Vol. 7 No. 1 (2002) pp. 1-20
13. Erjavec, T., Džeroski, S.: Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. Applied Artificial Intelligence, Vol. 18, No. 1 (2004) pp. 17-41
14. MT Evaluation Kit. NIST MT Evaluation Kit Version 11a. (2002) Available at http://www.nist.gov/speech/tests/mt.