Initial Considerations in Building a Speech-to-Speech Translation System for the Slovenian-English Language Pair

J. Žganec Gros[1], A. Mihelič[1], M. Žganec[1], F. Mihelič[2], S. Dobrišek[2], J. Žibert[2], Š. Vintar[2], T. Korošec[2], T. Erjavec[3], M. Romih[4]

[1]Alpineon R&D, Ulica Iga Grudna 15, SI-1000 Ljubljana, Slovenia
[2]University of Ljubljana, SI-1000 Ljubljana, Slovenia
[3]Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia
[4]Amebis, Bakovnik 3, SI-1241 Kamnik, Slovenia
E-mail: jerneja@alpineon.com

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 multimodal 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 (Lavie et al., 1997), (Wahlster, 2000), (Lavie et al., 2002). 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 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 co-funded by the Slovenian Ministry of Defense. The aim of the project is to build a robust multimodal speech-to-speech translation communicator, similar to Phraselator (Sarich, 2001) or Speechalator (Waibel, 2003), 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 (Seneff et al, 1998). The Galaxy Communicator open source architecture was chosen to provide inter-module communication support as 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. It is a distributed, message-based, hub-and-spoke infrastructure optimized for constructing spoken dialogue systems.

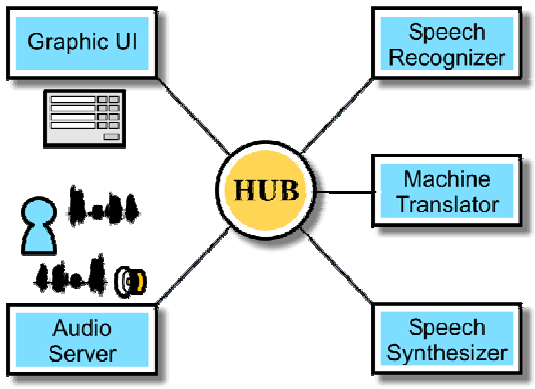

Figure 1. VoiceTRAN system architecture

The VoiceTRAN Communicator is composed of a number of servers that interact with each other through the Hub as shown in Figure 1. 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.

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

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 a 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 then input word strings into speech and prepares them for the audio server.

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.

2.2. Speech Recognizer

The speech recognition server receives the input audio stream from the audio server and provides at its output a word graph and 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 (Dobrišek, 2001). It will be upgraded to perform large vocabulary speaker (in)dependent speech recognition on a wider application domain. 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 standalone speech communicator used by a specific user, the speech recognizer can be additionally trained and adapted to the individual user 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 prosodic module will be investigated in order to link the source language to the target language, but also to enhance speech recognition proper. Besides syntactic and semantic information, properties such as dialect, sociolect, sex and attitude etc are signaled by prosody, (Wahlster, 2000). The degree of linguistic information conveyed by prosody varies between languages, from languages such as English, with a relatively low degree of prosodic disambiguation, via tone-accent languages such as Swedish, to pure tone languages (Eklund et al., 1995). Prosody information will help to determine proper punctuation and sentence accent information.

2.3. Machine Translator

The machine translator 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 appropriate 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 (Lavie et al., 2002) will be investigated and promising directions pointed out in (Ney, 2004) will be pursued.

The Presis translation system will be used as our baseline system (Romih et al., 2002). 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.

The speech recognition server sends to the hub each postprocessed sentence hypothesis from the N-best hypotheses list as a separate token. All of the different tokens for various sentence hypotheses of a given utterance can be jointly considered by a MT postprocessing module, after frame construction, which takes into consideration the quality of each hypothesis well-formedness that is evaluated by the machine translation server, along with a special criterium that looks for any salient words, such as named

entities. Early decisions can be made when a hypothesis is perfect; otherwise the final decision is delayed until the last hypothesis has been processed.

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 (Žganec Gros et al., 2004). 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.

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.

The speech synthesizer produces an audio stream for the utterance. The audio is stream is finally sent to the speakers by the audio server. After the synthesized speech has been transmitted to the user, the audio server is freed up in order to continue listening for the next user utterance.

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

Some of the multilingual language resources needed to set up STTS systems and include the Slovenian language are presented in (Verdonik et al., 2004).

For building the speech components of the VoiceTRAN system, existing speech corpora will be used (Mihelič et al., 2003). We do not have an actual in-domain speech database for the chosen application domain. However, as demonstrated by Lefevre (Lefevre et al., 2001) and others, out-of-domain speech training data do not cause significant degradation of the system performance. Since the speech corpora have been collected from different sources, adaptations will be carried out (Tsakalidis et al., 2005). 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 (Žganec Gros et al, 2004) 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 (Korošec, 2002) and various existing aligned parallel corpora will be used: (Erjavec, 2002) and (Erjavec et al., 2005).

4. 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, re-coding to XML TEI P4, sentence alignment, word-level syntactic tagging and lemmatisation (Erjavec and Džeroski 2004). 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.

5. Planned Evaluation

The intended experiments will be performed under quite clean conditions that are far different from real application environments where the speech signal is often mixed with noise and distorted by room reverberation or communication channels.

5.1. End-to-End System 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;
- 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 (Rossato et el., 2002).

5.2. Single Component Evaluation

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 (MT Evaluation Kit, 2004), (Akiba et al., 2004), having in mind that objective evaluation methods evaluate the translation quality in terms of the capacity of the system to mimic the reference text.

6. Conclusion

The VoiceTRAN project provides an attempt to build a robust multimodal 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.

7. Acknowledgements

The authors of the paper wish to thank the Slovenian Ministry of Defense and the Slovenian Research Agency for co-funding the project.

8. References

AKIBA, Y., FEDERICO, M., KANDO, N., NAKAIWA, H., PAUL, M., TSUJI, J. (2004). 'Overview of the IWSLT04 Evaluation Campaign'. In Proceedings of the International Workshop on Spoken Language Translation (pp. 1-9), Kyoto, Japan.

DOBRIŠEK, S. (2001). 'Analysis and Recognition of Phrases in Speech Signals'. PhD Thesis, University of Ljubljana, Slovenia.

EKLUND, R., LYBERG, B. (1995). 'Inclusion of a Prosodic Module in Spoken Language Translation Systems'. In Proceedings of the ASA 130th Meeting. St. Louis, MO.

ERJAVEC, T. (2002). 'The IJS-ELAN Slovene-English parallel corpus'. International Journal on Corpus Linguistics (pp. 1-20), Vol. 7., No. 1.

ERJAVEC, T., DŽEROSKI, S. (2004). 'Machine Learning of Language Structure: Lemmatising Unknown Slovene Words'. Applied Artificial Intelligence (pp. 17-41), Vol. 18, No. 1.

ERJAVEC, T. (2004). 'MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora'. In Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC'04 (pp. 1535-1538), Lisbon, Portugal.

ERJAVEC, T., IGNAT, C., POULIQUEN, P., STEINBERGER, R. (2005). 'Massive multi-lingual corpus compilation: Acquis Communautaire and totale'. Submitted to 2nd Language, Technology Conference, April 21-23, Poznań, Poland.

KOROŠEC, T. (2002). 'Opravljeno je bilo pomembno slovarsko delo o vojaškem jeziku'. Slovenska vojska,(pp. 12-13), Vol. 10., No. 10. (in Slovenian)

LAVIE, A., WAIBEL, A., LEVIN, L., FINKE, M., GATES, D., GAVALDÀ, M., ZEPPENFELD, T., ZHAN, P. (1997). 'Janus-III: Speech-to-Speech Trans-

lation in Multiple Languages'. In Proceedings of the ICASSP (99. 99-102), Munich, Germany.

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. (2002). 'A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System'. In Proceedings of the ACL 2002 workshop on Speech-to-speech Translation: Algorithms and Systems, Philadelphia, PA.

LEFEVRE, F., GAUVAIN, J.L., LAMEL, L. (2001). 'Improving Genericity for Task-Independent Speech Recognition'. In Proceedings of the Eurospeech (pp. 1241-1244), Aalborg, Denmark.

MIHELIČ, F., GROS, J., DOBRIŠEK, S., ŽIBERT, J., PAVEŠIĆ, N. (2003). 'Spoken language resources at LUKS of the University of Ljubljana'. International Journal on Speech Technologies (pp. 221-232), Vol. 6, No. 3.

MT Evaluation Kit (2002). 'NIST MT Evaluation Kit Version 11a'. Available at http://www.nist.gov/ speech/ tests/mt.

NEY, H. (2004). 'The Statistical Approach to Spoken Language Translation'. In Proceedings of the International Workshop on Spoken Language Translation (pp. XV-XVI), Kyoto, Japan.

ROMIH, M., HOLOZAN, P. (2002). 'Slovensko-angleški prevajalni sistem (A Slovene-English Translation System)'. In Proceedings of the 3rd Language Technologies Conference (p. 167), Ljubljana, Slovenia. (in Slovenian)

ROSSATO, S., BLANCHON, H., BESACIER, L. (2002). 'Speech-to-Speech Translation System Evaluation: Results for French for the Nespole! Project First Showcase'. In Proceedings of the ICSLP, Denver, CO.

SARICH, A. (2001). 'Phraselator, one-way speech translation system'. Available at http://www.sarich. com/translator/, 2001.

SENEFF, S., HURLEY, E., LAU, R., PAO, C., SCHMID, P., ZUE, V. (1998). 'Galaxy-II: A Reference Architecture for Conversational System Development'. In Proceedings of the ICSLP (pp. 931-934), Sydney, Australia, Available at http://communicator. sourceforge.net/.

TSAKALIDIS, S., BYRNE, W. (2005). 'Acoustic training from heterogeneous data sources: Experiments in Mandarin conversational telephone speech transcription'. In Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing, Philadelphia, PA, to appear.

VERDONIK, D., ROJC, M., KAČIČ, Z. (2004). 'Creating Slovenian language resources for development of speech-to-speech translation components'. In Proceedings of the Fourth international conference on language resources and evaluation (1399-1402), Lisbon, Portugal.

VIČIČ, J., ERJAVEC, T. (2002). 'Corpus driven machine translation'. In Proceedings of the 7th TELRI seminar Information in Corpora (p. 20,. Abstract, TELRI Association, Dubrovnik, Croatia.

Wahlster, W. (2000). 'Verbmobil: Foundation of Speech-to-Speech translation'. Springer Verlag.

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. (2003). 'Speechalator: Two-Way Speech-to-Speech Translation on a Consumer PDA'. In Proceedings of the EUROSPEECH (pp. 369-372), Geneva, Switzerland.

ŽGANEC GROS, J., MIHELIČ, A., ŽGANEC, M., PAVEŠIĆ, N., MIHELIČ, F., CVETKO OREŠNIK, V. (2004). 'AlpSynth corpus-driven Slovenian text-to-speech synthesis: designing the speech corpus'. In Proceedings of the joint conferences CTS+CIS, Computers in technical systems, Intelligent systems, (pp. 107-110), Rijeka, Croatia.