JANUS-III: SPEECH-TO-SPEECH TRANSLATION IN MULTIPLE LANGUAGES

Alon Lavie *Alex Waibel* *Lori Levin* *Michael Finke* *Donna Gates* *Marsal Gavaldà*
Torsten Zeppenfeld *Puming Zhan*

Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, USA.
University of Karlsruhe, Karlsruhe, Germany
Email: lavie@cs.cmu.edu

ABSTRACT

This paper describes JANUS-III, our most recent version of the JANUS speech-to-speech translation system. We present an overview of the system and focus on how system design facilitates speech translation between multiple languages, and allows for easy adaptation to new source and target languages. We also describe our methodology for evaluation of end-to-end system performance with a variety of source and target languages. For system development and evaluation, we have experimented with both push-to-talk as well as cross-talk recording conditions. To date, our system has achieved performance levels of over 80% acceptable translations on transcribed input, and over 70% acceptable translations on speech input recognized with a 75-90% word accuracy. Our current major research is concentrated on enhancing the capabilities of the system to deal with input in broad and general domains.

1. INTRODUCTION

JANUS-III is the most recent version of JANUS, a speech-to-speech translation system, designed to translate spontaneous dialogs between multiple speakers. JANUS is developed at the Interactive Systems Laboratories at Carnegie Mellon University and the University of Karlsruhe. The current system is designed for the Scheduling domain, in which two parties are participating in a negotiation dialog in an attempt to schedule a meeting.

A component diagram of our system can be seen in Figure 1. The main system modules are speech recognition, parsing, discourse processing, and generation. Each module is language independent in the sense that it consists of a general processor that can be loaded with language specific knowledge sources. This allows the easy adaptation of the system to new languages and domains. In an attempt to achieve both robustness and translation accuracy when faced with speech disfluencies and recognition errors, we use two different parsing strategies: a GLR parser designed to be more accurate, and a Phoenix parser designed to be more robust. Both modules follow an interlingua-based approach.

Speech translation in the JANUS system is guided by the general principle that spoken utterances can be analyzed and translated as a sequential collection of semantic dialog units (SDUs), each of which roughly corresponds to a speech-act. SDUs are semantically coherent pieces of information. The interlingua representation in our system was designed to capture meaning at the level of such SDUs. Each semantic dialog unit is analyzed into an interlingua representation. For both parsers, segmentation into SDUs is achieved in a two-stage process, partly prior to and partly during parsing.

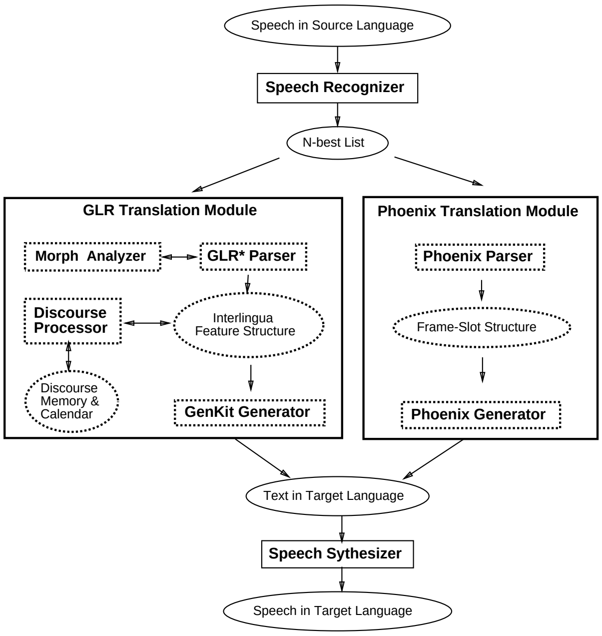

Figure 1. The JANUS System

In order to disambiguate among multiple interpretations, our strategy has been to apply a late stage disambiguation, which utilizes knowledge from all the machine translation components - acoustic and language models, parser scores, and contextual information obtained from discourse analysis. Each of these components provides a score for each possible analysis of an ambiguous input. One current research topic is the development of methods for combining these scores in a way that achieves optimal performance.

2. SPEECH RECOGNITION

The first major component in our speech-to-speech translation system is the speech recognizer. Its job is to decode the speech of a user and turn it into text to be passed to the parsing/translation modules. The accuracy of our end-to-end translation is greatly dependent on the word accuracy of our recognition components. While speech recognition systems readily achieve word accuracies of 90+% on read speech, conversational speech poses a much more difficult problem, and generally results in higher word error rates. Our JANUS-III recognition system has been applied to various conversational speech tasks, and now achieves Word Error rates below 10% on the Japanese, 23% on the English, 14% on the German [1], and 17% [11, 12] on the Spanish Spontaneous Scheduling Task. On the broad domain telephone quality, spontaneous speech task of the Switchboard corpus, our system performs at a WER of 36% [10]. This is a state-of-the-art performance result which illustrates the difficulty inherent in spontaneous speech tasks.

Porting speech recognition systems to a variety of languages requires attention to various language specific issues. These issues include the selection of the appropriate set of phonetic models, the choice of a set of word units (based on a language's morphology), and a phonetic transcription of these word units into a dictionary.

Especially in languages with a large number of inflections and compound words (like German, Spanish, Korean, Japanese) vocabulary growth is immense when unrestricted speech recognition is desired. In order to limit this large vocabulary growth, base units other than simple words have been used as new recognition units and for language model training.

Our JANUS-III recognizers are based on the Janus Recognition Toolkit (JRTk) [1], a flexible architecture for experimenting with language specific phenomena. The general configuration of our systems uses one or more streams of input features derived from Mel-scale, cepstral or PLP filters processed using Linear Discriminant Analysis (LDA). The acoustic units are context dependent, modeled via continuous density HMMs. Explicit noise models are added to help the system cope with breathing, lip-smack, and other human and non-human noises inherent in a spontaneous speech task.

Some of the recent improvements that have been introduced into our system include:

- **Speaker Normalization** - One major source of interspeaker variability is the variation in their vocal tract shape. In order to normalize for the vocal tract length, a maximum likelihood scaling in the frequency axis of the speech signal is performed for each speaker.
- **Polyphonic Modeling** - We allow questions in the allophonic decision tree to not only refer to the immediate neighboring phones but also to phones further away. This increases the degree of context-dependency.
- **MLLR Model Adaptation** - Based on the first pass recognition, we allow our models to adapt to specific speakers. The more data is available for a speaker, the more specific the models can become.
- **Dictionary Learning** - Due to the variability, dialect variations, and coarticulation phenomena found in spontaneous speech, pronunciation dictionaries have to be modified and fine-tuned for each language. To eliminate costly manual labor and for better modeling, we resort to data-driven ways of discovering such variants.
- **Morpheme Based Language Models** - For languages characterized by a richer morphology, which make wider use of inflections and compounding (compared to English), more suitable units than the 'word' are used for dictionaries and language models [3].
- **Phrase Based and Class Based Language Models** - Words that belong to word classes (such as days of the week), or frequently occurring phrases (e.g., *out-of-town, I'm-gonna-be, sometime-in-the-next*) are discovered automatically by clustering techniques and added to a dictionary as special words, phrases or mini-grammars.

2.1. Push-To-Talk versus Cross-Talk

During data collection, we have experimented with two different styles of recording. In the push-to-talk technique, only one speaker may speak at a time while a recording button is pressed. In the cross-talk technique on the other hand, both speakers are still recorded on separate channels, but may speak simultaneously. This allows a much more natural dialog. Table 1 shows some of the noticeable differences between these two recording styles for the Spanish Spontaneous Scheduling Task. Note that the cross-talk speakers use much shorter utterances. While both scenarios contain approximately the same proportion of noise, the cross-talk recordings contain more noise distorted words (e.g. words that are spoken during extraneous noises, such as laughter). We thus expect the performance of our recognizer to degrade for cross-talk recorded speech, and Table 1 shows this to be the case. However, the fact that the cross-talk utterances are much shorter helps the translation components, and thus we actually observe a slight improvement in the end-to-end performance using cross-talk recording conditions (see section 5.).

	push-to-talk	cross-talk
utterances	1090	7740
words	42142	73617
words/utt	38.6	9.5
percent noise	18%	19%
percent noise distorted words	10%	30%
performance	20%	23%

Table 1. Spanish Spontaneous Scheduling Task

3. THE ROBUST GLR AND PHOENIX TRANSLATION MODULES

JANUS employs two robust translation modules with complementary strengths. The GLR module gives more complete and accurate translations whereas the Phoenix module is more robust over the disfluencies of spoken language. The two modules can run separately or can be combined to gain the strengths of both.

The GLR module is composed of the GLR* parser [4][5], the LA-Morph morphological analyzer and the GenKit generator. The GLR* parser is based on Tomita's Generalized LR parsing algorithm [7]. GLR* skips parts of the utterance that it cannot incorporate into a well-formed sentence structure. Thus, it is well-suited to domains in which non-grammaticality is common. The parser conducts a search for the maximal subset of the original input that is covered by the grammar. This is done using a beam search heuristic that limits the combinations of skipped words considered by the parser, and ensures feasible time and space bounds. JANUS GLR grammars are designed to produce feature structures that correspond to a frame-based language-independent representation of the meaning of the input utterance. For a given input utterance, the parser produces a set of interlingua texts, or ILTs. The GLR* parser also includes several tools designed to address the difficulties of parsing spontaneous speech, including a statistical disambiguation module, a self-judging parse quality heuristic, and the ability to segment multi-sentence utterances. Target language generation is done using GenKit, a unification-based generation system. With well-developed generation grammars, GenKit results in very accurate translation for well-specified ILTs. We currently support GLR analysis grammars for Spanish and English, and a GenKit generation grammar for English.

The JANUS Phoenix translation module [6] is an extension of the Phoenix Spoken Language System [8]. It consists of a parsing module and a generation module. Unlike the GLR method which attempts to construct a detailed ILT for a given input utterance, the Phoenix approach attempts to only identify the key semantic concepts represented in the utterance and their underlying structure. The Phoenix parsing grammar specifies patterns which represent concepts in the domain. Each concept, irrespective of its level in the hierarchy, is represented by a separate grammar file. These grammars are compiled into Recursive Transition Networks (RTNs). The parser matches as much of the input utterance as it can to the patterns specified by the RTNs. The parser can ignore any number of words in between top-level concepts, handling out-of-domain or otherwise unexpected input. The parser has no restrictions on the order in which slots can occur. This may add to the ambiguity in the segmentation of the utterance into concepts. The parser uses a disambiguation algorithm that attempts to cover the largest number of words using the smallest number of concepts. Generation in the Phoenix module is accomplished using a simple strategy that sequentially generates target language text for each of the top level concepts in the parse analysis. Each concept has one or more fixed phrasings in the target language. The result is a meaningful but somewhat telegraphic translation. The simplicity of the Phoenix concept representation allows very rapid development of generation grammars for new languages. A generation grammar for Italian was recently developed to a reasonable level of performance within ten days. We currently support Phoenix analysis and generation grammars for German, Spanish and English, as well as additional generation grammars for Korean, Chinese, Japanese and Italian.

Although both GLR* and Phoenix were specifically designed to deal with spontaneous speech, each of the approaches has some clear strengths and weaknesses. Because each of the two translation methods appears to perform better on different types of utterances, they may be combined in a way that takes advantage of the strengths of each of them. One strategy that we have investigated is to use the Phoenix module as a back-up to the GLR module. The parse result of GLR* is translated whenever it is judged by the parse quality heuristic to be "Good". Whenever the parse result from GLR* is judged as "Bad", the translation is generated from the corresponding output of the Phoenix parser. Results of using this combination scheme are presented in Section 5.. We are in the process of investigating some more sophisticated methods for combining the two translation approaches.

Perfect	Fluent translation with all information conveyed
OK	All important information translated correctly but some unimportant details missing or translation is awkward
OK tagged	The sentence or clause is out-of-domain and no translation is given.
Bad	Unacceptable translation

Figure 2. Evaluation Grade Categories

	Transcription	Output of Speech-recognition
GLR*	82.9%	54.0%
Phoenix	76.3%	48.6%
Combined	83.3%	63.6%

Figure 3. End-to-end Translation Performance Results

4. LATE-STAGE DISAMBIGUATION

An important feature of our translation approach is to allow multiple interpretations to be processed through the system, and to use context to disambiguate between alternatives in the final stage of the process, where knowledge can be exploited to the fullest. Since it is infeasible to process *all* hypotheses produced by each of the system components, context is also used locally to prune out unlikely alternatives. The final disambiguation combines all knowledge sources obtained: the acoustic score, the parser score, and information obtained from the discourse processor. The best scoring interpretation is then sent to the generation module. This interpretation is also sent back to the discourse processor so it can update its internal structures and the discourse state.

5. EVALUATION METHODS AND RESULTS

The goal of our evaluation methods is to provide a meaningful and accurate measure of the capability of our system as a whole. We accomplish this by periodically testing our system on sets of "unseen" data. The data chosen for testing consists of dialogs by speakers whose voices were not used for training or development of both the speech recognizer and the translation components. We perform evaluations on the end-to-end system from speech recognition through target language generation. A similar evaluation is conducted using transcribed input instead of speech recognized input. This allows us to isolate performance deficiencies that are solely due to speech recognition errors. The evaluations are scored by independent graders. We employ a consistent set of criteria for judging the quality of the utterances as well as their relevance to the current domain. Each SDU is assigned a separate grade. A grading assistant program helps the scorer in assigning SDU level scores, tabulates and saves the results. Figure 2 lists the possible grades and the criteria for assigning them. The translation modules attempt to detect out-of-domain SDUs (SDUs that are not about scheduling meetings) in order to avoid erroneous translations. An SDU that is recognized as out-of-domain and is not translated is given the score "OK tagged".

The results in Figure 3 show the performance of the GLR and the Phoenix Spanish-English translation modules on a recent test set of 3 dialogs (103 utterances) recorded in a cross-talk setting (see following subsection). The results shown are for in-domain SDUs only and reflect the percent of acceptable translations. The speech recognition average word accuracy on this test set was 66.8%. The results in the last row of Figure 3 reflect the combination of the GLR* and Phoenix systems as described in Section 3.. As can be seen, the combination of the two parsers results in a significant improvement in translation performance on speech recognized input. On transcribed input the improvement is much less significant.

A similar evaluation of German-to-English translation was recently conducted using only the Phoenix translation module (we do not support a GLR German analysis grammar). The test set consisted of 3 dialogs (98 utterances). System performance was 82.4% acceptable translations on transcribed input, and 70.3% acceptable translations on speech recognized input, where the average word accuracy of the test set was 86.2%. Phoenix generation into other target languages achieved similar level of performance.

6. CONCLUSIONS AND FUTURE WORK

In this paper we described the methods we employ in the JANUS system for integrating speech recognition and translation in multiple languages. Our end-to-end evaluation procedures allow us to assess the overall performance of the system, using each of the translations methods separately or both combined.

Our current and future research efforts concentrate on extending the design of the system to enable handling more general domains. We are focusing our attention on the "Travel Planning" domain, on which we collaborate with other C-STAR (Consortium for Speech Translation Research) member groups. Our speech recognition system already achieves state-of-the-art performance on the broad domain Switchboard corpus, and will be further developed for the Travel Planning domain. We are also experimenting with several approaches for adapting our translation modules to the travel domain. These include more general semantic grammars and interlingua representations, as well as methods for combining grammars for limited sub-domains. Our significant progress in dealing with speech translation for multiple languages in the Scheduling domain leads us to believe that multi-lingual speech translation in broad domains is an achievable near-future goal.

ACKNOWLEDGEMENTS

The work reported in this paper was funded in part by grants from ATR - Interpreting Telecommunications Research Laboratories of Japan, the US Department of Defense, and the Verbmobil Project of the Federal Republic of Germany.

REFERENCES

[1] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries and M. Westphal. *The Karlsruhe-Verbmobil Speech Recognition Engine,* To appear in ICASSP-97, Munich, Germany

[2] D. Gates, A. Lavie, L. Levin, A. Waibel, M. Gavaldà, L. Mayfield, M. Woszczyna and P. Zhan. *End-to-end Evaluation in JANUS: a Speech-to-speech Translation System,* To appear in Proceedings of ECAI Workshop on Dialogue Processing in Spoken Language Systems, Budapest, Hungary, August 1996.

[3] P. Geutner. *Using Morphology towards better Large-Vocabulary Speech Recognition Systems,* in Proceedings of ICASSP-95, Detroit, Michigan, 1995

[4] A. Lavie and M. Tomita. *GLR* - An Efficient Noise Skipping Parsing Algorithm for Context Free Grammars,* Proceedings of the third International Workshop on Parsing Technologies (IWPT-93), Tilburg, The Netherlands, August 1993.

[5] A. Lavie. *An Integrated Heuristic Scheme for Partial Parse Evaluation,* Proceedings of the 32nd Annual Meeting of the ACL (ACL-94), Las Cruces, New Mexico, June 1994.

[6] L. Mayfield, M. Gavaldà, Y-H. Seo, B. Suhm, W. Ward, A. Waibel. *Parsing Real Input in JANUS: a Concept-Based Approach,* In Proceedings of TMI 95.

[7] M. Tomita. *An Efficient Augmented Context-free Parsing Algorithm,* Computational Linguistics, 13(1-2):31–46, 1987.

[8] W. Ward. /em Extracting Information in Spontaneous Speech, In Proceedings of International Conference on Spoken Language, 1994.

[9] M. Woszczyna, N. Aoki-Waibel, F. D. Buo, N. Coccaro, T. Horiguchi, K. and Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. P. Rosé, T. Schultz, B. Suhm, M. Tomita, and A. Waibel. *JANUS-93: Towards Spontaneous Speech Translation,* In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'94), 1994.

[10] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal and A. Waibel. *Recognition of Conversational Telephone Speech using the JANUS Speech Engine,* To appear in ICASSP-97, Munich, Germany

[11] P. Zhan, K. Ries, M. Gavalda, D. Gates, A. Lavie and A. Waibel. *JANUS-II: Towards Spontaneous Spanish Speech Recognition,* Proceedings of ICSLP-96, Philadelphia, PA, October 1996

[12] P. Zhan and M. Westphal *Speaker Normalization Based on Frequency Warping,* To appear in ICASSP-97, Munich, Germany

JANUS-III: SPEECH-TO-SPEECH TRANSLATION IN MULTIPLE LANGUAGES

Alon Lavie , Alex Waibel , Lori Levin , Michael Finke , Donna Gates , Marsal Gavaldà , Torsten Zeppenfeld and Puming Zhan

Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, USA.
University of Karlsruhe, Karlsruhe, Germany
Email: `lavie@cs.cmu.edu`

This paper describes JANUS-III, our most recent version of the JANUS speech-to-speech translation system. We present an overview of the system and focus on how system design facilitates speech translation between multiple languages, and allows for easy adaptation to new source and target languages. We also describe our methodology for evaluation of end-to-end system performance with a variety of source and target languages. For system development and evaluation, we have experimented with both push-to-talk as well as cross-talk recording conditions. To date, our system has achieved performance levels of over 80% acceptable translations on transcribed input, and over 70% acceptable translations on speech input recognized with a 75-90% word accuracy. Our current major research is concentrated on enhancing the capabilities of the system to deal with input in broad and general domains.