Interactive Translation of Conversational Speech

Alex Waibel
waibel@cs.cmu.edu
http://www.is.cs.cmu.edu

Interactive System Laboratories

Language Technology Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Fakultät für Informatik
Universität Karlsruhe
D-76131 Karlsruhe
Germany

We present JANUS-II, a large scale system effort aimed at interactive spoken language translation. JANUS-II now accepts *spontaneous* conversational speech in a limited domain in English, German or Spanish and produces output in German, English, Spanish, Japanese and Korean. The challenges of coarticulated, disfluent, ill-formed speech are manifold, and have required advances in acoustic modeling, dictionary learning, language modeling, semantic parsing and generation, to achieve acceptable performance. A semantic "interlingua" that represents the intended meaning of an input sentence, facilitates the generation of culturally and contextually appropriate translation in the presence of irrelevant or erroneous information. Application of statistical, contextual, prosodic and discourse constraints permits a progressively narrowing search for the most plausible interpretation of an utterance. During translation, JANUS-II produces paraphrases that are used for interactive correction of translation errors. Beyond our continuing efforts to improve robustness and accuracy, we have also begun to study possible forms of deployment. Several system prototypes have been implemented to explore translation needs in different settings: speech translation in one-on-one video conferencing, as portable mobile interpreter, or as passive simultaneous conversation translator. We will discuss their usability and performance.

1.0 Introduction

Multilinguality will take on spoken form when information services are to extend beyond national boundaries or across language groups. Database access by speech will need to handle multiple languages to service customers from different language groups. Public service operators (emergency, police, telephone operators and others) frequently receive requests from foreigners unable to speak the national language. Already multilingual spoken language services are growing. Telephone companies in the US (AT&T Language Line), Europe and Japan now offer language translation services over the telephone, provided by human operators. Movies and television broadcasts are routinely translated and

delivered either by dubbing, subtitles or multilingual transcripts. With the drive of automating information services, therefore, comes a growing need for automated multilingual speech processing. While few commercial multilingual speech services yet exist, intense research activities are underway. The major aims are: (1) Spoken Language Identification, (2) Multilingual Speech Recognition and Understanding for human-machine interaction, (3) Speech Translation for human-to-human communication. Speech translation is the most ambitious of the three, as it requires greater accuracy and detail during the analysis, and potentially needs to track highly disfluent and colloquial conversational speech.

2.0 Background

In the not too distant past, the possibility of one day being able to carry out a telephone conversation with a speaker, with whom you share no common language, appeared remote. With the state of the art in speech recognition and machine translation still far short of perfection, the combination of the two technologies could not be expected to deliver acceptable performance.

The late ’80s and early ’90s, however, have seen tremendous advances in speech recognition performance, propelling the state of the art from speaker dependent, single utterance, small vocabulary recognizers (e.g., digits) to speaker independent, continuous speech, large vocabulary dictation systems at around 10% word error rate. Similarly, machine translation has advanced considerably, and a number of text translation products are now commercially available.

2.1 The Problem of Spoken Language Translation

Beyond improving each component, however, it has become increasingly clear, that good speech translation cannot be achieved by mere combination of better speech recognition and machine translation components. Just as continuous speech recognition has become possible without attempting to achieve perfect phoneme recognition performance (In fact phoneme accuracy still ranges between 50% and 70%), the problem must be attacked in its entirety. Closer inspection of actual human spoken dialogs verifies this intuition. Consider an actual spoken dialog between two Spanish speakers trying to agree on a time for an appointment. The following example shows a manually produced careful transliteration of the utterance, the way it was actually spoken by the speaker:

“...sí sí el viernes diecinueve puedo sí porque sabes me voy de viaje d hoy la verdad así es que este mes es muy viajero me voy el día seis de viaje y estoy hasta el doce así que el día diecinueve me viene muy bien francamente...”

Running this utterance through a commercial text translation system, the following translation results was obtained. (Note, that this would even assume perfect speech recognition):

yes yes on friday nineteen can yes because know I go me of trip D today the truth such is that this month is very traveler I go me the day six of trip and I am until the twelve as soon as the day nineteen comes me very well outspokenly

What went wrong? The fact is humanly spoken sentences are hardly ever well-formed in the sense that they seldom obey rigid syntactic constraints. They contain disfluencies, hesitations (um, hmm, etc.), repetitions ("*.... so I, I, I guess, what I was saying.*"), and false starts ("*..how about we meet on Tue.. um.. on Wednesday.....*"). Yet put in the context of discussion they are still perfectly understandable for a human listener. A successful speech translation system therefore cannot rely on perfect recognition or perfect syntax. Rather, it must search for a semantically plausible interpretation of the speaker's intent while judiciously ignoring linguistically unimportant words or fragments.

The problem described is exacerbated by recognition errors and environmental noises that occur during speech recording, such as coughs, laughter, telephone rings, door slams, etc.. Without proper treatment, these noises may be recognized as one of the words in the vocabulary, potentially causing great damage in the translation process. The dramatic variation in speaking rate is another problem to be accounted for in human-to-human dialog recognitions. In fast speech, considerably higher error rates are observed due to coarticulation, reduction or elisions between the words.

A spoken dialog does not consist of sentences in the classical sense, nor are we provided with punctuation markers to delimit them. Instead, each utterance is fragmentary and each speaker's turn often contains two or more sentences or concepts ("*... no, Tuesday doesn't work for me...how about...Wednesday morning...Wednesday the twelfth*"). Even if we were given punctuation markers, attempts to translate such fragmentary utterances frequently result in awkward output.

To provide useful spoken language communication across language barriers, we must therefore *interpret* an utterance, or extract its *main i*ntent, rather than attempt a sentence by sentence translation. This often involves summarization. Thus we wish to "translate" the previous Spanish example above as:

"... I'm available on Friday the nineteenth..."

Only by way of a semantic and pragmatic interpretation within a domain of discourse can we hope to produce culturally appropriate expressions in another language.

2.2 Research Efforts on Speech Translation

Speech translation research today began with systems in the late eighties and early nineties whose main goal was to demonstrate feasibility of the concept. In addition to domain constraints, these early systems had fixed speaking style, grammatical coverage and vocabulary size. Their system architecture was usually strictly sequentially, involving speech recognition, language analysis and generation, and speech synthesis in the target language. Developed at industrial and academic institutions, they represented a modest, yet significant first step toward multilingual communication. Early systems include independent research prototypes developed by ATR[1], AT&T[2], Carnegie Mellon University and the University of Karlsruhe[3], NEC[4], and Siemens AG.

Most were developed through international collaborations that provided the cross-linguistic expertise. Among these international cooperations, the Consortium for Speech Translation Advanced Research, or C-STAR, was formed as a voluntary group of institutions

committed to build speech translation systems. It arose from a partnership among ATR Interpreting Telephony Laboratories (now Interpreting Telephony Laboratories) in Kyoto, Japan, Carnegie Mellon University (CMU) in Pittsburgh, USA, Siemens AG in Munich, Germany, and University of Karlsruhe (UKA) in Karlsruhe, Germany. Additional members joined forces as partners or affiliates: ETRI (Korea), IRST (Italy), LIMSI (France), SRI (UK), IIT (India), Lincoln Labs (USA), DFKI (Germany), MIT (USA), and AT&T (USA). C-STAR continues to grow and to operate in a fairly loose and informal organizational style with each of its partners building complete systems or component technologies, thereby maximizing the technical exchange and minimizing costly software/ hardware interfacing work between partners. In addition to the activity of consortia such as C-STAR, and the industrial research described above, there are government sponsored initiatives in several countries. One of the largest is Verbmobil, an eight year effort sponsored by the BMFT, the German Ministry for Science and Technology [5] that involves 32 research groups.

3.0 JANUS-II - A Conversational Speech Translator

JANUS [3] was one of the early systems designed for speech translation. It was developed at Carnegie Mellon University and University of Karlsruhe in the late ’80s and early ’90s in partnership with ATR (Japan) and Siemens AG (Germany). Since then it has been extended at both sites to more advanced tasks. Results from these efforts now contribute to on-going spoken language translation efforts in the US (Project Enthusiast) and Germany (Project Verbmobil). While the first version, JANUS-I, processed only syntactically well-formed (read) speech over a smaller (500 word) vocabulary, JANUS-II now operates on spontaneous conversational human-human dialogs in limited domains with vocabularies of around 3000+ words. At present, it accepts English, German, Spanish, Japanese and Korean input and delivers translations into German, English, Spanish, Japanese or Korean. Further languages are under development.

Beyond translation of syntactically well formed speech, or (relatively well behaved) human-machine speech utterances, the research focus for JANUS-II has been on the translation of *spontaneous conversational human-to-human* speech. In the following we introduce a suitable database and task domain and discuss the JANUS-II spoken language translator.

3.1 Task Domains and Data Collection

To systematically explore the translation of spoken language, a database for training, testing and benchmarking had to be provided. For realism in practical situations a task domain had to be chosen that requires translation between humans trying to communicate with each other, as opposed to tasks that aim at information retrieval (human-machine). Some applications of speech translation (See section below.) will have elements of human-machine dialogs, when a computer intervenes in the communication process providing feedback to the users. In other situations, however, simultaneous translation of on-going human-to-human conversations is desired.

A symmetric negotiation dialog is chosen. As a task domain, many sites have adopted the appointment scheduling domain proposed in the Verbmobil project. To elicit natural con-

versations that are nonetheless contained and, more importantly, comparable across languages, we have devised sets of calendars with given constraints, that get progressively more complex and generate more conflicts between speakers. Subjects are simply requested to schedule a meeting with each other and do so at their own pace and in whatever fashion they wish to express themselves. The same calendars can be used (in translated form) for monolingual dialog recordings in each language. The dialogs are recorded in an office environment, typically using push-to-talk buttons to activate recording. The recordings are transcribed carefully and double-checked to ensure that *all* acoustic events (including repetitions, false starts, hesitations, human and non-human noises) are transcribed and listed in the transcripts as they occur in the signal. Several sites in Europe, the US and Asia are now collecting and transcribing data in this fashion. More than 2,000 dialogs corresponding to about half a million words have been collected for English. Somewhat smaller databases to date have been collected for German, Spanish, Korean and Japanese by various sites as well.

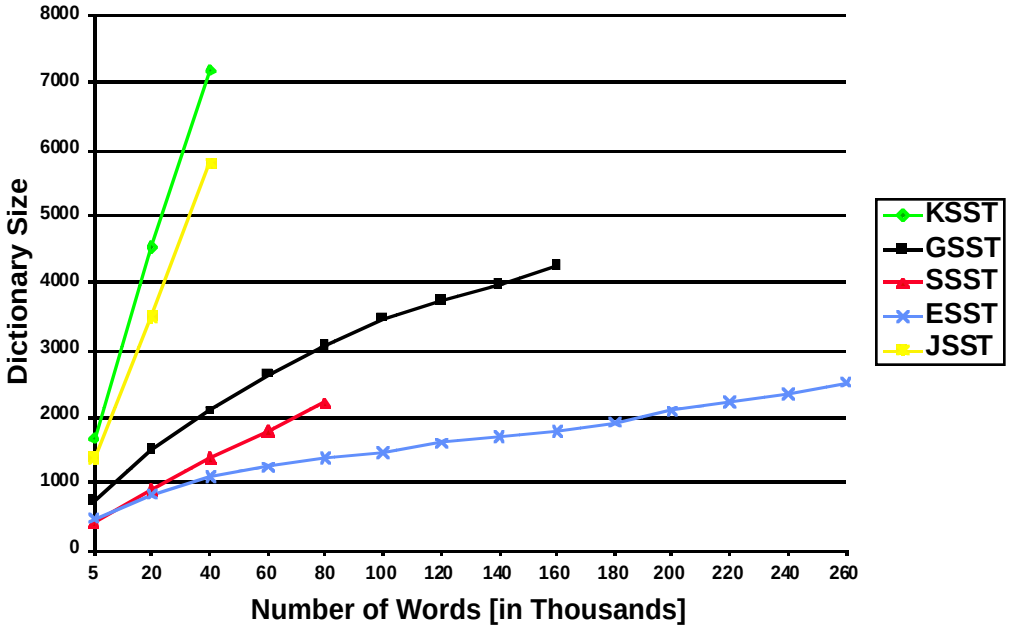

FIGURE 1. Vocabulary Growth as a Function of Database Size

Figure 1 shows for various languages the growth of the resulting vocabularies as a function of the number of words spoken by subjects. For up to a quarter of million spoken words the domain vocabulary grows to around 3000 words in English. We note that (as always in spontaneous speech) there cannot be full saturation even at that level, as there will always be new words to contend with. Interesting in this figure is also the rapid growth in vocabulary size for Japanese, Korean and even for German. This is the result of using full form ‘word’ entries in the dictionary, a strategy that is appropriate for English, debatable for German, and inappropriate for Japanese and Korean. German, characterized by large numbers of inflections and noun compounds, and Japanese/Korean by packaging entire phrases into inflected forms generate many more variants from root forms than English and have to be broken down into subunits. In Spanish we have also explored two different data collection strategies: (1) A push-to-talk button scenario on one side, which

requires the speaker to hold down a record a button while talking to the system. (2) A cross-talk scenario on the other allowing speakers to speak simultaneously and taking turns whenever they want to. The speech of each dialog partner is recorded on separate channels. Each of these two recording scenarios is evaluated in actual speech translation system tests.

3.2 System Description

The key to the problem of speech translation is finding an approach to dealing with uncertainty and ambiguity at every level of processing. A speaker will produce ill-formed sentences, and noise will surround the desired signal; the speech recognition engine will produce recognition errors; the analysis module will lack in coverage, and without consideration to dialog and domain constraints each utterance will be ambiguous in meaning.

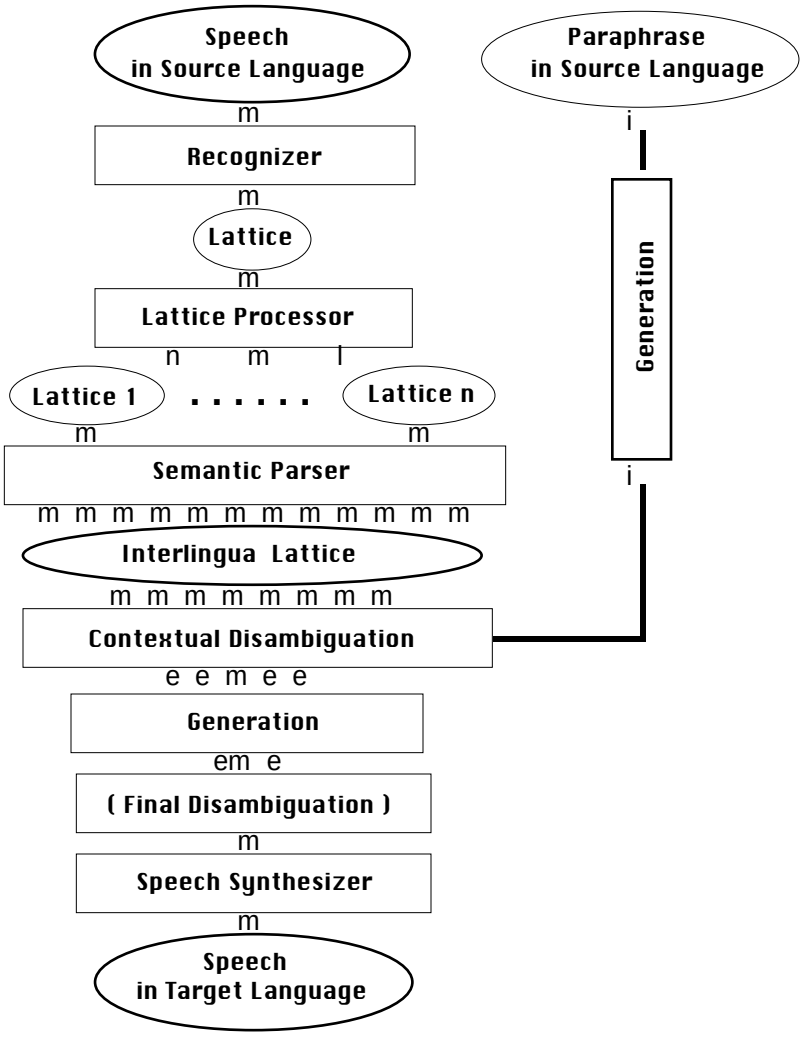

FIGURE 2. JANUS-II System Level Overview

JANUS-II was designed to deal with this problem by applying all sources of knowledge (from acoustic to discourse) successively to narrow the search for the most plausible translation. Two approaches appear possible: (1) to provide feedback (backtracking) from later knowledge sources to earlier knowledge sources, (2) to maintain a list or graph of possibilities at each stage and narrow these possibilities as each subsequent knowledge source is applied. The second approach is selected, mostly for efficiency reasons. It does not require backtracking or repeating earlier processing stages and allows, in principle, for incremental speech translation, that is, continuous recognition and translation, potentially while the speaker is speaking.

Figure 2 shows a system overview. 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.

Speech is accepted through a signal processing front-end and processed by the speech recognition module. Stationary background noises (computer hum, telephone hiss, air conditioner, microphone channel) are removed or normalized by signal enhancement techniques in the signal processing front-end. Nonstationary human and non-human noises such as coughs, lipsmacks, breathing, and telephone rings, door slams, etc. are modeled as 'garbage' models and recognized individually as noise-words. To avoid having to create models for each conceivable noise in the world, a clustering algorithm reduces these garbage words to a more manageable number of up to seven prototypical noise garbage categories.

The recognition module then generates acoustic scores for the most promising word hypotheses, given a pronunciation dictionary. It uses Hidden Markov Models (HMM) and HMM-Neural Net hybrid technologies combined with statistical language models [6] in an attempt to produce most robust recognition performance.

In lieu of the best recognition hypothesis, the JANUS-II recognition engine returns a lattice of near-miss hypothesis fragments organized as a graph. This graph is then reduced by a lattice processor that has two functions:

- Eliminate redundant or unproductive alternatives, such as arcs that differ only by different noise-word hypotheses on the assumption that a confusion between such noise alternatives (say, key-click vs. mircophone tap) have no bearing on translation accuracy.
- Break a long utterance into usable smaller sublattices according to rough prosodic cues, such as pauses and hesitations for further processing.

The resulting shorter and reduced lattices are then passed on to the language analysis.

Unlike JANUS-I that relied heavily on syntactic analysis, JANUS-II employs almost exclusively a semantic analysis. This is to obtain robust interpretation of the meaning in spite of ill-formedness of expression and recognition error from the input. Several parsing approaches are used: A semantic pattern based chart parser (Phoenix) and GLR*, a stochastic, fragment based extension of an LR parser. Both employ semantic grammars and derive a language independent representation of meaning called "Interlingua".

There are three main advantages for the Interlingua approach: First, it aims to reduce the dependence of the output sentence on the structural form of the input language. What matters is the intent of the input utterance, whatever the way it was expressed. Sentences like "*I don't have time on Tuesday*", "*Tuesday is shot*", "*I am on vacation, Tuesday*", can now all be mapped onto the same intended meaning "I am unavailable on Tuesday", and an appropriate sentence in the output language can be generated. Even culturally dependent expressions can be translated in a culturally appropriate fashion. Thus "*Tuesday's no good*" could be translated into "*Kayoobi-wa chotto tsugo-ga warui"* literally: "*As for Tuesday, the circumstance is a little bit bad"*. The second advantage of the Interlingua approach is the comparative ease by which additional languages can be added. Thus only one output generator has to be written for each new output language, as opposed to adding an analysis and an generation module for each language pair. The ease of generating output in any language constitutes the third opportunity and advantage: generating an output utterance in the input language thereby paraphrasing the input. This permits the user to verify if an input utterance was properly analyzed. This very important feature improves the usability of a speech translation, as the user most likely does not know if an output translation in an unknown language is correct or not.

Semantic representations in natural language processing have, of course, been studied extensively over the years, leading to a number of Interlingua based text translation systems (see [11][12][13] for review). We find the use of an interlingua based approach particularly advantageous for the translation of spontaneous speech, as spoken language is syntactically more ill-formed and less reliable, but the semantics typically more contained.

For each recognition hypothesis emerging from the recognizer, a semantic analysis is performed, resulting in a rank ordered list or a lattice of meanings. Naturally, not every recognition hypothesis will result in a different hypothesis, nor will every recognition hypothesis result in a semantically plausible hypothesis, so that a substantial reduction in remaining hypotheses can be achieved. The semantic analysis in the JANUS-II system is provided by one of several parsing schemes, the Phoenix parser, the GLR* parser (see discussion below), and several exploratory connectionist and statistical learning parsers.

After parsing, a discourse processor or contextual disambiguation can be applied to select the most appropriate meaning from the Interlingua lattice, based on the additional consideration of the context or discourse state. There are three different approaches that can be used to perform this selection or reordering: 1.) discourse plan based inference mechanisms, 2.) Interlingua N-grams (conditioning the current meaning on previous dialog states, and 3.) a dialog finite state machine. The proper weighting of each of the disambiguating strategies is obtained by training statistics over a large training database.

Following the parsing stage, generation of an appropriate expression in the output language is performed, followed by speech synthesis in the output language. For synthesis, JANUS-II resorts to commercial synthesis devices and/or builds on the speech synthesis research work of our partners.

3.2.1 The Recognition Engine

The baseline JANUS-II recognizer uses two streams of LDA coefficients derived over melscale, power and silence features. It uses a three pass Viterbi Decoder, Continuous

Density HMM's, Cross-Word Triphones and speaker adaptation. A channel normalization and explicit noise models are designed to reduce stationary background noise and human and non-human noise events.

In our effort of enhancing the overall system performance, we continue to improve the underlying speech and translation strategies. Particularly, in the light of our need to rearrange and redeploy our recognizer for different languages and different tasks, we wish to automate many aspects of the system design that might otherwise be predetermined once.

Improved results have recently been achieved through the following strategies[6]:

- **Data Driven Codebook Adaptation** - These are methods aimed at automatically optimizing the number of parameters.
- **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, use of inflections and compounding than English, more suitable units than the 'word' are used for dictionaries and language models.
- **Phrase Based and Class Based Language Models** - Words that belong to word classes (MONDAY, TUESDAY, FRIDAY...) or frequently occurring phrases (e.g., OUT-OF-TOWN, I'M-GONNA-BE, SOMETIMES-IN-THE-NEXT) are discovered automatically by clustering techniques and added to a dictionary as special words, phrases or mini-grammars.
- **Special Subvocabularies** [7] - Special Confusable Subvocabularies (e.g. Continuous Spelling for Names and Acronyms) are processed in a second classification pass using connectionist models.

3.2.2 Robust Parsing Strategies

Two main parsing strategies are used in our work: the Phoenix Spoken Language Parser, and the GLR* robust parser.

- **The Phoenix Spoken Language System** [8] was extended to parse spoken language input into slots in semantic frames and then use these frames to generate output in the target language. Based on transcripts of scheduling dialogs, we have developed a set of fundamental semantic units that represent different concepts of the domain. Typical expressions and sentence patterns in a speaker's utterance are parsed into semantic chunks, which are concatenated without grammatical rules. As it ignores non-matching fragments and focuses on important key phrases, this approach is particularly well suited to parsing spontaneous speech, that is often ungrammatical and subject to recognition errors. Generation based on conceptual frames is terse but delivers the intended meaning.

- **The GLR* Parser** [9] - As a more detailed semantic analysis we also pursue GLR*, a robust extension of the Generalized LR Parser. It attempts to find a maximal subsets of an input utterance that are parsable, skipping over unrecognizable parts. By means of a semantic grammar GLR* parses input sentences into an interlingua, a language independent representation of the meaning of the input sentence. Compared to Phoenix interlingua generated by GLR* offers greater level of detail and more specificity, e.g. different speaker attitudes and levels of politeness. Thus, translation can be more natural, overcoming the telegraphic and terse nature of concept based translation. As GLR* skips over unrecognizable parts, it has to consider a large number of potentially meaningful sentence fragments. To control the combinatorics of this search, stochastic parsing scores and pre-breaking of the incoming lattices are used to reduce the ambiguity. GLR* has greater computational requirements but produces more detailed translation.

3.3 Performance Evaluation

To assess the performance and relative progress in the development of speech translators, several evaluation measures have to be devised. Evaluations can be performed at three levels:

- Speech Recognition Rate - Measured, as usual, by counting substitution, deletion and insertion errors over an unseen test database.
- Semantic Analysis based on Transcripts - This can be measured, if a ‘desired’ interlingua representations (the reference) has been established over a new test set. The drawback of this approach is that it is subjective and requires considerable manual labor.
- End-to-End Translation Accuracy based on 1.) Transcriptions and 2.) Recognizer Input. Each clause or conceptual fragments (not each turn) represents an event for evaluation to avoid undue weighting of short confirmatory remarks (e.g., “*That’s right*”, “*OK*”). Output is then judged by a panel of three judges under the criteria “Good”, “Acceptable” and “Bad”, where Acceptable means an utterance was translated awkwardly, but still transmits the intended meaning. Utterances that were established as ‘out-of-domain’ were counted as acceptable, if they produced an acceptable translation nonetheless or rejected the utterance as ‘out-of-domain’, and they were counted as bad otherwise.

Figure 3a shows the recognition results obtained over the course of recent development on a Spanish conversational translator for the scheduling domain. As can be seen, the initial recognition accuracy was quite low, which is explained in part by insufficient data in the initial stages of development for a new language. In other parts, however, the results reflect the difficulty of processing human-to-human conversational dialogs. As other research teams have found (see ICASSP’95, for example) on similar tasks (e.g., the Switchboard corpus, where, due to higher perplexity and the additional difficulty of telephone bandwidth, the results of only 50+% word accuracy have so far been achieved), human-to-human dialogs are highly disfluent, heavily coarticulated, vary enormously in speaking rate, and contain many more short poorly articulated words than read or human-machine speech. Indeed, better accuracies (exceeding 80%) *can* be observed in the scheduling domain, when speakers are not conversing with each other but are cognizant of the fact that they are talking to a computer.

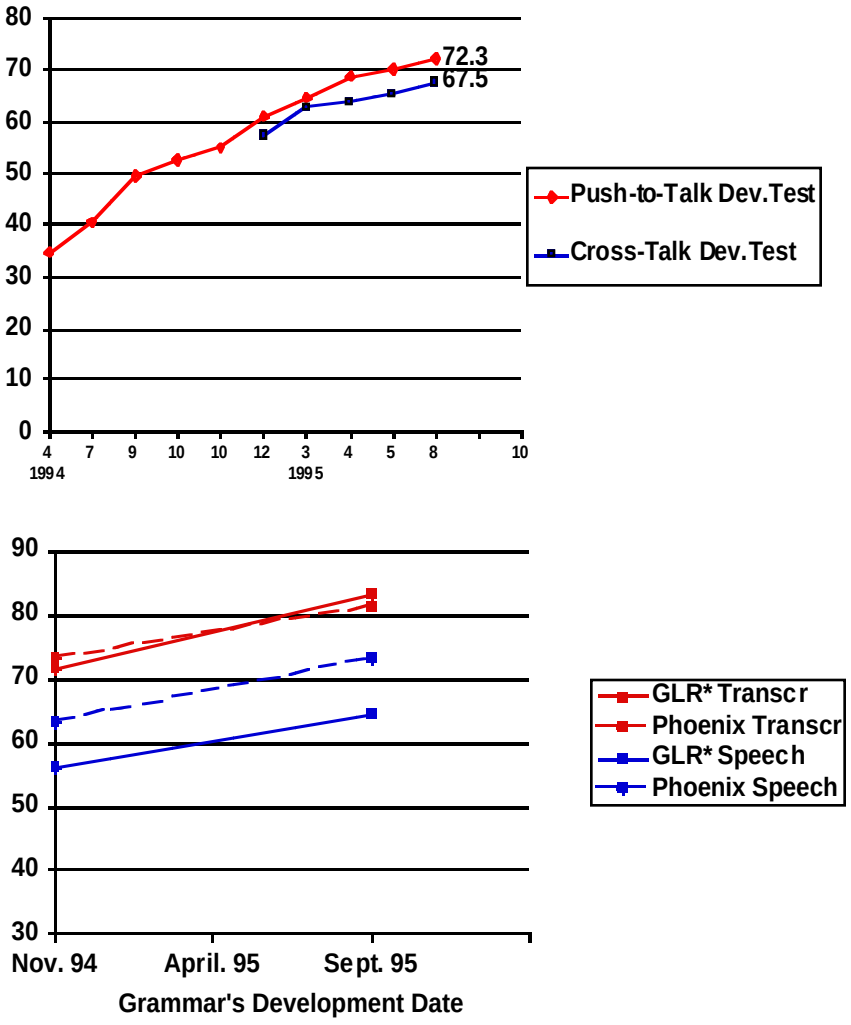

FIGURE 3. Performance Results for the Recognition (a) and Translation (b) of Spanish Conversational Dialogs

Figure 3a also shows a comparison between speech collected using a push-to-talk switch and free cross-talk dialogs. While both are human-to-human, cross-talk appears to result in even less well behaved speech and thus is more difficult than push-to-talk speech. For other languages (English, German, Japanese), JANUS-II currently delivers similar word accuracies of 70+%. In recent evaluations carried out by the Verbmobil project using five different recognition engines recognition these accuracies up to 70% were found to be the best achievable for conversational German so far.

Figure 3b shows the result of end-to-end speech translation performance over a set aside test set. The results were obtained by scoring the translations produced by three different grammars from three different moments in the development cycle. The same test set was used to test all three grammars (of course, without any development in the interim). Reassuringly, translation accuracy was found to improve with grammars of greater coverage. It

can be seen that translation accuracies up to 85% can be achieved based on transcribed spoken language, and up to 74% using the two parsers, Phoenix and GLR*.

Table 1 finally shows an interesting comparison between cross-talk and push-to-talk conditions. It was carried out using the Phoenix parser in both cases over several unseen test sets. Human translators report translating the rapid-fire turn-taking in spontaneous dialogs as unacceptably difficult. Based on these reports, we predicted that cross-talk speech would be much harder to recognize and to translate by machine as well. Since we have to compare results from different test dialogs (with considerable variability in performance) to check this prediction, we note that a precise comparison under equal conditions is not possible. Within our task domain and over multiple tests, however, a surprising trend appears to emerge. While cross-talk speech is indeed generally harder to recognize than push-to-talk, it results in shorter turns that were found to translate as well or better. Thus, translation of uninhibited human conversational dialogs appears to be no more problematic than controlled turn taking.The difficulties human translators experience with rapid cross talk dialogs might be related to the human cognitive load of tracking two parallel speech channels, rather than any intrinsic translation difficulty of the material.

TABLE 1. Comparison between Push-to-Talk and Cross Talk Dialogs (Human-Human)

	SpeechRecognition Accuracy	**Translation of Transcript**	**Speech-to-Speech Translation**
Push-to-Talk Data	71%	74%	52%
Cross-Talk Data	70%	81%	73%

4.0 Applications and Forms of Deployment

The need for Spoken Language Interpretation arises in different situations, each posing different challenges and opportunities. We have begun experimenting with three such different application scenarios: 1.) Spoken language interpretation in an interactive video conferencing environment, 2.) Portable Speech Translator, and 3.) Simultaneous Dialog Translation.

4.1 Interactive Dialog Translation

Figure 4 shows a prototype video conferencing station with spoken language interpretation facility. There are two displays: One facing the user and another, touch sensitive display, embedded in the desk. The user operates his own station by way of the desk screen. A record button activate speech acquisition and displays both the recognition result and a paraphrase of the analyzed utterance. This is accomplished by performing a generation from the (language independent) interlingua back into the user's language. The user can now verify if the paraphrase reflects the intended meaning of the input utterance. If so, he presses a send button, which replaces the paraphrase by the translation into the selected output language and sends it on to the other video conferencing site. At the other site, the

translation appears as subtitles under the transmitted video image of our user. It is also synthesized in the target language for speech output. The translation display can also be used to run collaborative virtual environments such as joint white-boards or applications that the conversants make reference to. Translation can be delivered in about two times real time.

FIGURE 4. JANUS-II Speech Translator in a Video Conferencing Environment. Translation appears visually as subtitles as well as by synthetic output. The system is controlled by buttons on a touch sensitive display. Vocabulary size is between 1000 and 3000 words per language. Translations are obtained in 2-3 times real-time

The video conferencing station is a cooperative translation environment, where both conversants are trying to be understood and can verify the systems understanding of a spoken utterance. It can therefore benefit from user feedback and can more easily assure correctness. It also offers alternative modes for user input as well as for error recovery: Input can be provided by handwriting or typing in addition to speech. In case of error these alternative modalities can be applied to generate a new paraphrase and translation [10]. In this way, effective communication can be provided despite imperfect recognition and translation. In addition to offering a variety of recovery mechanisms, the translation station also elicits somewhat more benign user speaking style than human-to-human conversational speech.

Work is in progress that exploits this opportunity for error correction. To recover from human and machine error, a number of strategies have been explored [11], including repair by respeaking, spelling, and handwriting as alternative redundant modes of human-computer interaction. Recovery can typically be achieved within one or two tries. The JANUS-

II system also offers other simple forms of assistance, such as letting the user simply type over erroneous recognitions.

The interface allows the user to select different output languages by language buttons on the translator screen. The input language is either set by the user at system start-up or can be set automatically by a language identification module as a preprocessor. In effect, the system begins by processing an incoming speech utterance via recognizers of several languages, and the most probable language is selected based on the goodness of match.

The environment still offers many opportunities for further study of the human factors of interactive spoken language translation. The best trade-off between processing speed and accuracy, the role of repair and multimodality in the translation process, how to deal with out-of domain utterances, how to learn and integrate new words or concepts, are all issues for continuing investigation.

4.2 Portable Speech Translation Device

JANETTE is a down-sized version of JANUS-II. The system runs on a Laptop PC (a 75 MHz Pentium) with 32 MB of memory. In this configuration the system currently still takes about twice as long per utterance to translate than on our video stations. The system can be carried in a knapsack or a carrying bag (Figure 5). Translation is presented either by an acoustic earpiece, or by a wearable heads-up display. The wearable heads-up display displays the translation in text form on see-through goggles, thereby allowing the user to see subtitles under the face of the person he/she is talking to. This alternate presentation of translation result allows for greater throughput, as the translation can be viewed without interrupting the speaker. While acoustic output may allow for feedback with the system, a simultaneously displayed translation may therefore provide greater communication speed. The human factors of such new devices still await further study in actual field use.

4.3 Passive Simultaneous Dialog Translation

The language interpreting systems described so far offer the opportunity for feedback, verification and correction of translation between two cooperative conversants who want to cooperate with each other. Not every situation affords this possibility, however. In N-party conference situations, foreign TV or radio broadcasts, or simultaneous translation of speeches or conversations, a passive un-cooperative translation situation is encountered. Here, the speaker cannot be involved in the communication process for verification of the translation. Also, in the case of conversational speech, this kind of translation is likely to be particularly difficult as it requires processing of human-to-human speech, greater coarticulation, and potentially more difficult turn taking phenomena. Indeed, the rapid succession of sometimes overlapping turns makes the cognitive planning of a translation particularly difficult for humans attempting to translate conversational dialog.

Our results reported above for cross-talk and push-to-talk dialogs, however, suggest that the same cognitive limitations experience by human translators do not hold for machines: two separate speech translation processes can easily process separate channels of a dialog and produce translations that keep up with the conversants. In our lab, a conversational translator has been installed that slices turns at major breaking points and sends the corre-

sponding speech signals to an array of 5 processors, that incrementally generate translations during the course of a human conversation (here, once again, two subjects negotiating a meeting). Despite the disfluent nature of such an interactive and rapid conversation, translation of conversational dialogs within this domain can be performed accurately more than 70% of the time.

FIGURE 5. Wearable Speech Translator
Shown with Microphone and Head-Mounted Heads-Up Display. The Heads-Up- Display shows translation output overlaid using see-through goggles. Alternatively, acoustic output can be presented by earpiece. Current speed is still 7 times real-time and the system's vocabulary had to be reduced to 500+ words per language from a limited domain.

Acknowledgments

The author wishes to express his gratitude to his collaborators, Jaime Carbonell, Wayne Ward, Lori Levin, Alon Lavi, Carol VanEss Dykema, Michael Finke, Donna Gates, Marsal Gavalda, Petra Geutner, Thomas Kemp, Laura Mayfield, Arthur McNair, Ivica Rogina, Tanja Schultz, Tilo Sloboda, Bernhard Suhm, Monika Woszczyna and Torsten Zeppenfeld.

This Research would not have been possible without the support of the BMBF (project Verbmobil) for our work on the German recognizer, the US Government (project Enthusiast) for Spanish components, and ATR Interpreting Telecommunications Laboratories for English speech translation collaboration. Thanks are also due to the partners and affiliates in C-STAR, who have helped define speech translation today.

References

[1] T.Morimoto, T.Takezawa, F.Yato, S.Sagayama, T.Tashiro, M.Nagata, and A.Kurematsu, “ATR's Speech Translation System: ASURA”, EUROSPEECH 1993, pp. 1295.

[2] D.B.Roe, F.C.N.Pereira, R.W.Sproat, and M.D.Riley, ”Efficient Grammar Processing for a Spoken Language Translation System”, ICASSP 1992, Vol. 1, pp. 213.

[3] A.Waibel, A.M.Jain, A.E.McNair, H.Saito, A.G.Hauptmann, and J.Tebelskis “JANUS: A Speech-To-Speech Translation System Using Connectionist and Symbolic Processing Strategies”, ICASSP’91, 1991.

[4] K.Hatazaki, J.Noguchi, A.Okumura, K.Yoshida, T.Watanabe, “INTERTALKER: An Experimental Automatic Interpretation System Using Conceptual Representation”, ICSLP 1992.

[5] W.Wahlster “First Results of Verbmobil: Translation Assistance for Spontaneous Dialogues” ATR International Workshop on Speech Translation, November 8-9, 1993.

[6] B.Suhm, P.Geutner, T.Kemp, A.Lavie, L.Mayfield, A.E.McNair, I.Rogina, T.Schultz, T.Sloboda, W.Ward, M.Woszczyna, and A.Waibel, “JANUS: Towards Multilingual Spoken Language Translation” in Proceedings of the ARPA Spoken Language Technology Workshop, Austin, TX, January 1995.

[7] H.Hild, and Alex Waibel, “Integrating Spelling into Spoken Dialogue Recognition”, EUROSPEECH, Vol. 2, pp. 1977.

[8] W.Ward, “Understanding Spontaneous Speech: The Phoenix System”, ICASSP 1991, Vol. 1, pp. 365.

[9] A.Lavie and M.Tomita, “GLR* - An Efficient Noise-Skipping Parsing Algorithm for Context-Free Grammars”, Proceedings of Third International Workshop on Parsing Technologies, 1993, p. 123.

[10] M.T.Vo, R.Houghton, J.Yang, U.Bub, U.Maier and A. Waibel, “Multimodal Learning Interfaces”, in Proceedings of the ARPA Spoken Language Technology Workshop, Austin, TX, January 1995.

[11] A.E.McNair and A.Waibel, “Improving Recognizer Acceptance through Robust, Natural Speech Repair”, ICSLP 1994, Vol. 3, pp. 1299.

[12] Hutchins, W.J. and H. Somers. “An Introduction to Machine Translation, Academic Press, San Diego, 1992.

[13] Hovy, E.H., “How MT Works”, Special Feature on Machine Translation, Byte Magazine, (167--176), January, 1993.

[14] S. Nirenburg, J.C. Carbonell, M. Tomita, and K. Goodman, “Machine Translation: A Knowledge-Based Approach”, Morgan Kaufmann, San Mateo, 1992.

Web Pages to Probe Further:

Interactive Systems Lab:	http://www.is.cs.cmu.edu
C-STAR:	http://www.is.cs.cmu.edu/cstar
Verbmobil:	http://www.dfki.uni-sb.de/verbmobil

Key Words:

Speech translation, speech-to-speech translation, voice translation, portable speech translators, spoken language understanding, spontaneous speech, multilingual speech processing, speech recognition, semantic representation, interlingua, machine translation, human-to-human speech dialogs, discourse, conversational speech, machine translation evaluation, parsing, machine learning, user interfaces, human computer interaction.