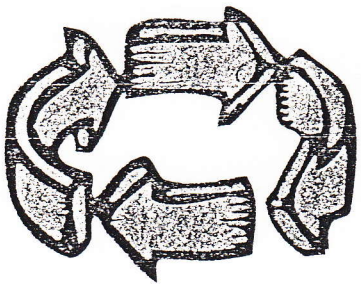
Interactive Translation of Conversational Speech

Alex Waibel
Carnegie Mellon University and University of Karlsruhe



As communication becomes increasingly automated and transnational, the need for rapid, computer-aided speech translation grows. The Janus-II system uses paraphrasing and interactive error correction to boost performance.

0018-9162/96/$5.00 © 1996 IEEE

Not very long ago, the possibility of being able to carry on a telephone conversation with someone who spoke a different language appeared remote. Speech recognition and machine translation were rudimentary, and no one expected these two technologies to deliver acceptable performance.

The past 10 years, however, have seen tremendous advances in speech recognition performance. The technology has progressed from speaker-dependent, single-utterance, small-vocabulary recognizers (for example, spoken digit strings, as in telephone numbers or zip codes) to speaker-independent, continuous-speech, large-vocabulary dictation systems with word error rates of about 10 percent. Similar advances in machine translation have resulted in commercially available text translation products.

Advances in this area will have far-reaching effects. As information services extend beyond national boundaries, database vendors will have to provide speech access in multiple languages to serve customers from different language groups. Public service operators (for emergencies, police, directory assistance, and so on) frequently receive requests from immigrants or visitors unable to speak the native language. Multilingual spoken-language services are growing to meet this need. Telephone companies in the US, Europe, and Japan now staff human operators who offer language translation. AT&T's Language Line is an example. Movies and television broadcasts are routinely translated and delivered either by dubbing, subtitles, or multilingual transcripts.

The drive to automate information services has revealed the need for automated multilingual speech processing. Few commercial multilingual speech services exist, but research in the field is intense. The major goals are

- spoken-language identification,
- multilingual speech recognition and understanding for human-machine interaction, and
- speech translation for communication between humans.

Speech translation is the most ambitious goal because it requires greater accuracy and detail during analysis than either of the other two, and it must be able to track highly *disfluent* and colloquial conversational speech. Translation of conversational speech therefore represents the ultimate frontier for both speech and language processing and offers many other potential benefits for speech and language applications.

THE CHALLENGE OF SPEECH TRANSLATION

Although speech recognition and machine translation have both improved, it has become increasingly clear that merely combining them

cannot produce good speech translation. Continuous-speech recognition has become possible without attempting to achieve perfect phoneme recognition (in fact, phoneme accuracy still ranges between 50 and 70 percent). Obviously, other factors enter into the solution, which suggests that the problem must be attacked in its entirety. Closer inspection of actual spoken dialogue verifies this intuition. Consider this fragment from a conversation between two Spanish speakers trying to agree on a time to meet. A careful manual transliteration of the utterance as actually spoken is

> *. . . si si 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 (assuming perfect speech recognition) produces

> 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? In the real world, people's spoken sentences are hardly ever well-formed and seldom obey rigid syntactic constraints. They contain disfluencies, including hesitations ("um," "hmm"), repetitions (". . . so I, I, I guess, what I was saying"), and false starts (". . . how about we meet on Tue . . . um . . . on Wednesday . . ."). Yet in context they are perfectly understandable to 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.

Recognition errors and environmental noises—coughs, laughter, a telephone ring, a door slam—exacerbate this problem. Without proper treatment, these noises may be mistaken for part of the vocabulary and thereby greatly degrade the translation.

The dramatic variation in speaking rate is another challenge. Fast speech causes considerably higher error rates because it involves more coarticulation, reduction, or elisions between words.

Spoken dialogue does not consist of sentences in the classical sense, nor are there punctuation markers to delimit sentences and clauses. 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 there were punctuation markers, attempts to translate such fragmentary utterances would result in awkward output.

To provide useful speech translation, we must attempt more than a sentence-by-sentence translation: We must *interpret* an utterance or extract its main intent. This often involves summarizing. Thus, we wish to "translate" the Spanish example above as ". . . I'm available on Friday the nineteenth . . ." Only through semantic and pragmatic interpretation within a domain of discourse can we hope to produce culturally appropriate expressions in another language.

Other speech translation efforts

Systems in the late '80s and early '90s were intended mainly to demonstrate the feasibility of speech translation. Along with domain constraints, they had a fixed speaking style, and vocabulary size and grammatical coverage were limited. Their system architecture usually handled speech recognition, language analysis and generation, and speech synthesis in the target language sequentially. Developed at industrial and academic institutions, these systems 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 speech translation systems were developed through international collaboration that provided cross-linguistic expertise. For example, the Consortium for Speech Translation Advanced Research, or C-STAR, arose from a partnership comprising ATR Interpreting Telephony Laboratories (now just Interpreting Telephony Laboratories) in Kyoto, Japan; Carnegie Mellon University in Pittsburgh; Siemens AG in Munich; and the University of Karlsruhe in Karlsruhe, Germany. Additional members joined as partners or affiliates: ETRI (Korea), IRST (Italy), LIMSI (France), SRI (UK), IIT (India), DFKI (Germany), and Lincoln Labs, MIT, and AT&T in the US. Still growing, C-STAR has a fairly loose and informal organizational style. Each partner builds complete systems or component technologies, thereby maximizing technical exchange and minimizing costly software/hardware interfacing efforts among members.

Governments in several countries have also sponsored initiatives. One of the largest is Verbmobil, an eight-year effort sponsored by BMFT, the German Ministry for Science and Technology, that involves 32 research groups.

References

1. T. Morimoto et al., "ATR's Speech Translation System: ASURA," *Proc. Eurospeech 93*, pp. 1,295-1,298.
2. D.B. Roe et al., "Efficient Grammar Processing for a Spoken Language Translation System," *Proc. Int'l Conf. Acoustics, Speech and Signal Processing*, IEEE Press, Piscataway, N.J., 1992, pp. 213-216.
3. A. Waibel et al., "Janus: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies," *Proc. Int'l Conf. Acoustics, Speech and Signal Processing*, IEEE Press, Piscataway, N.J., 1991.
4. K. Hatazaki et al., "INTERTALKER: An Experimental Automatic Interpretation System Using Conceptual Representation," *Proc. Int'l Conf. Speech and Language Processing*, 1992.

JANUS-II—A CONVERSATIONAL-SPEECH TRANSLATOR

Janus[1] was an early speech-translation system developed at Carnegie Mellon University and the University of Karlsruhe in the late '80s and early '90s, in partnership with ATR in Japan and Siemens in Germany. (See the sidebar "Other speech translation efforts" for additional background.) My colleagues and I at both sites have extended Janus to handle more advanced tasks. Results from these efforts now contribute to ongoing spoken-language-translation efforts in the US (Project Enthusiast) and Germany (Project Verbmobil). The first version of our spoken-language translator, Janus-I, processed only syntactically well-formed speech (speech that was read) using a small (500-word) vocabulary.

Janus-II operates on spontaneous conversational human dialogue in limited domains with vocabularies of 3,000 or more words. Current experiments involve 10,000- to 40,000-word vocabularies. It now accepts English, German, Japanese, Spanish, and Korean input, which it translates into any other of these languages.

Beyond translating syntactically well-formed speech or carefully structured human-to-machine speech utterances, Janus-II research has focused on the more difficult task of translating spontaneous conversational speech between humans. This naturally requires a suitable database and task domain.

Task domain and database

To systematically explore spoken-language translation, we needed a database for training, testing, and benchmarking. To be realistic and practical, the chosen task domain had to require translation between humans trying to communicate with each other, as opposed to tasks that involve human-machine information retrieval.

The first step is choosing a *symmetric negotiation* dialogue—conversation containing some give and take. A task domain with this kind of dialogue is appointment scheduling, as proposed in the Verbmobil project.[2] To elicit natural conversations that are nonetheless contained and, more importantly, comparable across languages, we have devised sets of calendars with constraints that get progressively more complex and generate additional conflicts between speakers. We asked subjects to schedule a meeting at their own pace and to express themselves however they wanted.

To build our database, we recorded and transcribed these dialogues. Working in an office environment, participants typically pushed buttons to activate the recording. The recordings were transcribed carefully and double-checked to ensure that all acoustic events (including repetitions, false starts, hesitations, and human and nonhuman noises) were transcribed and listed in the transcripts as they occurred in the signal. Several sites in Europe, the US, and Asia are now collecting and transcribing data in this fashion. We have collected more than 2,000 English-language dialogues encompassing about half a million words. Various sites have collected somewhat smaller databases for German, Spanish, Korean, and Japanese.

Figure 1 shows how vocabularies grow as a function of the number of words spoken. A quarter of a million words spoken in English produces a domain vocabulary of about 3,000 words. In spontaneous speech, a system dictionary cannot achieve full coverage even at that level, because there will always be new words to contend with. The figure shows rapid growth in vocabulary size for Japanese, Korean, and even for German. These languages generate many more variants from root forms than English and must be broken down into subunits. Thus, a strategy of using full-form "word" entries in the dictionary is appropriate for English and Spanish, possible for German, but inappropriate for Japanese and Korean.

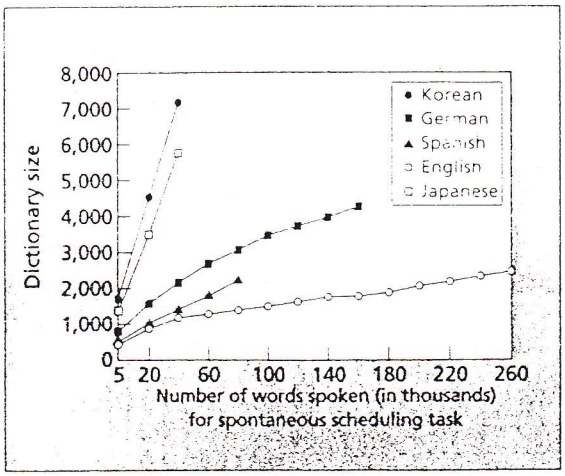


Figure 1. Domain vocabulary size as a function of database size. The rate of vocabulary growth differs by language.

System description

The key to speech translation is finding a way to deal with uncertainty and ambiguity at every level of processing. For example,

- a speaker will produce ill-formed sentences,
- noise will surround the desired signal,
- the speech recognition engine will make errors,
- the analysis module will lack coverage, and
- without dialogue and domain constraints, an utterance's meaning may be ambiguous.

Janus-II was designed to deal with these difficulties by successively applying all sources of knowledge—from acoustic to discourse—to narrow the search for the most plausible translation. Two approaches appear possible:

- Provide feedback (backtracking) from later knowledge sources to earlier knowledge sources.
- Maintain a list or a graph of possibilities at each stage and narrow these possibilities as each subsequent knowledge source is applied.

We selected the second approach, mainly for its efficiency. It does not require backtracking or repeating earlier processing stages. In principle, it allows for incremental speech translation—that is, continuous recognition and translation, potentially while the speaker is speaking.

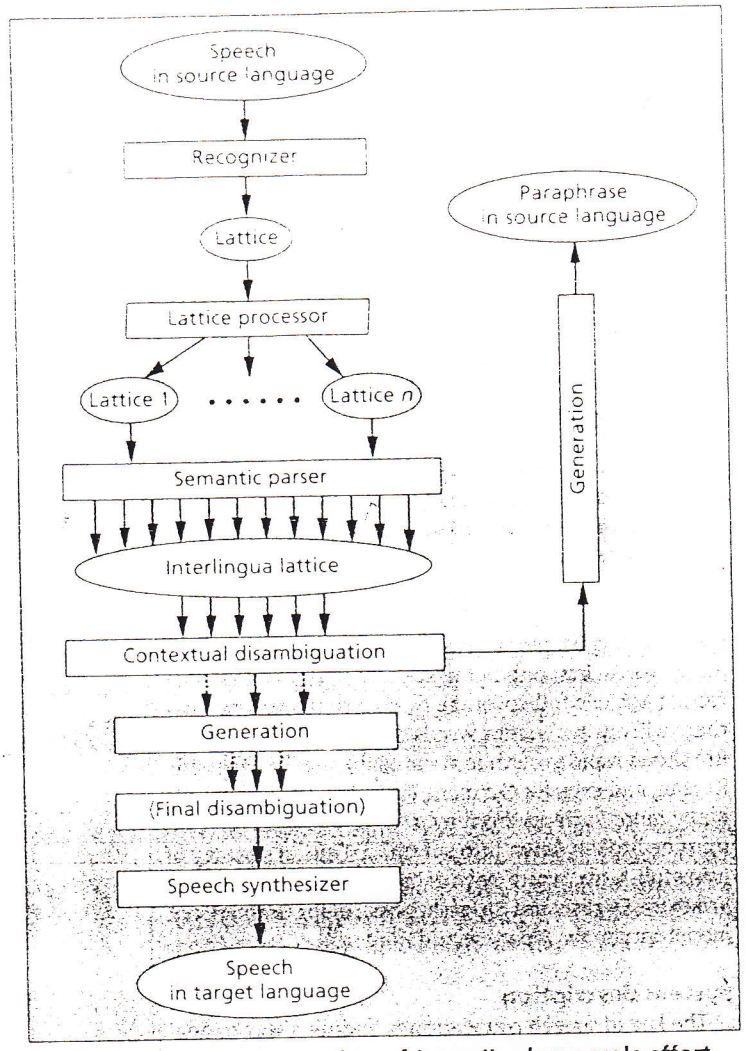


Figure 2. Overview of Janus-II, a large-scale effort aimed at interactive spoken-language translation. The system uses language-independent modules.

Figure 2 shows a Janus-II system overview. The main system modules are speech recognition (the recognizer), parsing, discourse processing (contextual disambiguation), 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.

Speech is accepted through a signal processing front end, which uses signal-enhancement techniques to filter or normalize stationary background noises. Nonstationary human and nonhuman noises, such as lip smacking, coughs, doors slamming, or telephones ringing, must be explicitly modeled as "garbage" words and then removed. So that we do not have to create a model for every conceivable noise, a clustering algorithm reduces these noises to seven prototypical noise-garbage categories.

Given a pronunciation dictionary, the recognition module then generates acoustic scores for the most promising word hypotheses. It uses hidden Markov models (HMMs) and HMM-neural net hybrid technologies combined with statistical language models[3] to try to produce the most robust recognition performance.

In lieu of the best recognition hypothesis, the Janus-II recognition engine returns a *lattice* (a rank-ordered list) 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. (We assume that confusion between such noise alternatives—say, key click versus microphone tap—has no bearing on translation accuracy.)
- Break a long utterance into usable smaller sublattices, using rough prosodic cues such as pauses and hesitations.

The shorter, reduced lattices are then passed to the language analysis module (semantic parser in Figure 2). Unlike Janus-I, which relied heavily on syntactic analysis, Janus-II employs semantic analysis almost exclusively. This approach obtains a robust interpretation of meaning in spite of poorly formed expression and input recognition errors. Janus-II uses several parsing approaches: A semantic pattern-based chart parser (Phoenix), and GLR* (generalized LR*), a stochastic, fragment-based extension of an LR parser. Both employ semantic grammars and derive a language-independent representation of meaning—an *Interlingua*.

The Interlingua approach has three main advantages: First, it aims to reduce the output sentence's dependence on the input language's structure. What matters is the intent of the input, regardless of how it's expressed. Janus-II can now map sentences like "I don't have time on Tuesday," "Tuesday is shot," or "I am on vacation Tuesday" to the same intended meaning—"I am unavailable on Tuesday"—and generate an appropriate sentence in the output language. Even culturally dependent expressions can be translated in a culturally appropriate fashion. Thus "Tuesday's no good" could be translated into Japanese as *"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 with which additional languages can be added. Thus, only one output generator must be written for each new output language, as opposed to adding an analysis and a generation module for each language pair.

The third advantage is that it is easy to generate output in any language. Being able to generate an output utterance in the input language lets us paraphrase the input. With this very important feature, the user can verify whether an input utterance was properly analyzed. Speech translation thus becomes more practical, because users probably don't know if an output translation in an unknown language is correct.

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.[4-6] We find the use of an Interlingua-based approach particularly advantageous for translating spontaneous speech, because spoken language is less well formed syntactically and less reliable; however, the semantics are typically more contained.

For each recognition hypothesis the recognizer generates, the semantic parser performs a semantic analysis, resulting in a lattice of meanings. Naturally, not every recogniti

hypothesis will result in a different semantic hypothesis, nor will every recognition hypothesis be semantically plausible. Thus, after semantic analysis fewer hypotheses will remain. Semantic analysis in the Janus-II system is provided by one of several parsing schemes, as I describe later.

After parsing, the system can apply a discourse processor or contextual disambiguation to the remaining semantic hypotheses. This makes it possible to incorporate additional consideration of the context or discourse state when selecting the most appropriate meaning from the Interlingua lattice. To select this meaning, Janus-II can use

- discourse-plan-based inference mechanisms,
- statistics of turn-taking patterns (conditioning the current meaning on previous dialogue states), and/or
- a dialogue finite-state machine.

We can obtain the proper weighting of each disambiguating strategy by training statistics over a large training database.

After disambiguation, an appropriate expression is generated in the output language, followed by speech synthesis in the output language. For synthesis, Janus-II resorts to commercial synthesis devices or builds on the speech synthesis research work of partners in the project.

RECOGNIZER. The baseline Janus-II recognizer uses two streams of coefficients derived by performing a linear discriminant analysis over mel-scale spectral features and power and silence features. It uses a three-pass Viterbi decoder, continuous-density HMMs, cross-word triphones, and speaker adaptation. Channel normalization and explicit noise models reduce stationary background noise or nonstationary human noises (breathing, smacking of lips) and nonhuman noises (doors slamming, phones ringing).

In trying to enhance overall system performance, we continue to improve the underlying speech recognition and translation strategies. Especially because we need to rearrange and redeploy our recognizer for different languages and tasks, we wish to automate many system design aspects to minimize the experimental effort when tasks or languages are changed.

We've recently achieved improved results through the following strategies[2]:

- *Data-driven codebook adaptation.* These are methods aimed at automatically optimizing the number of modeling parameters.
- *Dictionary learning.* Because of the variability in pronunciation, dialect variations, and coarticulation phenomena found in spontaneous speech, pronunciation dictionaries must be modified and fine-tuned for each language. We use data-driven methods to save time and effort and to improve performance.
- *Morpheme-based language models.* For languages characterized by a richer morphology and greater use of inflections and compounding than occur in English, more suitable units than the "word" are used for dictionaries and language models.
- *Phrase- and class-based language models.* Words that belong to word classes (Monday, Tuesday, Friday) or frequently occurring phrases (for example, "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 minigrammars.
- *Special subvocabularies.* To avoid confusion, special subvocabularies (for example, continuous spelling for names and acronyms) are processed in a second classification pass using connectionist models.

PARSER. We use two main parsing strategies in our work: the Phoenix spoken-language parser and the GLR* robust parser.

- *Phoenix spoken-language system.*[3] This system was extended to parse spoken-language input into slots in semantic frames and then use those frames to generate output in the target language. On the basis of scheduling-dialogue transcripts, 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. Because it ignores nonmatching fragments and focuses on important key phrases, this approach is particularly well suited for parsing spontaneous speech, which is often ungrammatical and subject to recognition errors. Generation based on conceptual frames is terse but delivers the intended meaning.
- *GLR* Parser.*[9] For a more detailed semantic analysis, we also pursue GLR*, a robust extension of the Generalized LR Parser. This strategy attempts to find maximal subsets of an input utterance that are parsable, skipping over unrecognizable parts. Using a semantic grammar, GLR* parses input sentences into an Interlingua, a language-independent representation of the input sentence's meaning. Compared with Phoenix, the Interlingua generated by GLR* provides a greater level of detail and more specificity—for example, different speaker attitudes and levels of politeness. Thus, translation can be more natural, overcoming the telegraphic and terse nature of concept-based translation. Because GLR* skips over unparsable parts, it must consider a large number of potentially meaningful sentence fragments. To control the combinatorics of this search, GLR* uses stochastic parsing scores and prebreaking of the incoming lattices to reduce the ambiguity. GLR* has greater computational requirements but produces more detailed translation.

Performance evaluation

We devised several measures to evaluate the performance and relative progress of speech translator development. Thus, we evaluate Janus-II on three levels:

- *Speech recognition rate.* We measure this by counting substitution, deletion, and insertion errors over a previously unseen test database.
- *Semantic analysis based on transcripts.* For this to be measured, a "desired" Interlingua representation (the reference) must have been established over a new test set. This approach is subjective and requires considerable manual labor.

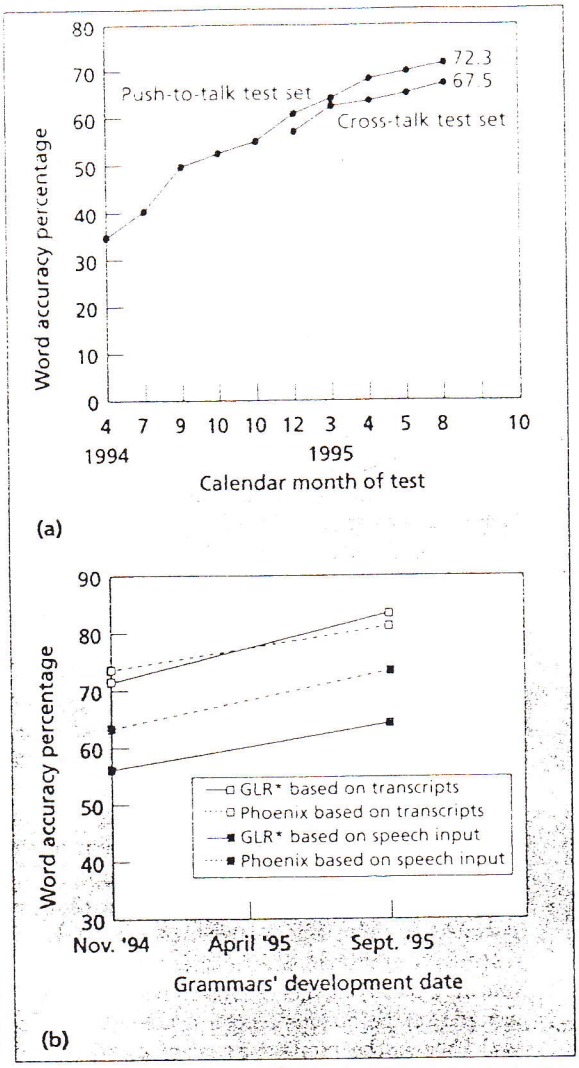


Figure 3. Performance results shown as percentages for the recognition (a) and translation (b) of Spanish conversational dialogues within a domain.

- *End-to-end translation accuracy based on transcriptions and recognizer input.* Each clause or conceptual fragment (not each turn) represents an event for evaluation. This avoids undue weighting of short confirmatory remarks ("That's right," "OK," and so forth). Three (human) judges then determine whether the output is "good," "acceptable," or "bad." Acceptable means an utterance was translated awkwardly but still transmits the intended meaning. Utterances established as "out-of-domain" are counted as acceptable if they produce an acceptable translation nonetheless, or if they are rejected as out-of-domain; otherwise, they are counted as bad.

Figure 3a shows recognition results obtained during recent development of a Spanish conversational translator for the scheduling domain. We see that initial recognition accuracy was quite low; this is partly explained by insufficient data in the initial stages of development for a new language. Additionally, the results reflect the difficulty of processing conversational dialogue between people. Such dialogue is highly disfluent and heavily coarticulated, varies considerably in speaking rate, and contains many more short, poorly articulated words than human-machine speech or speech that is read. Other research teams have also experienced this on similar conversational tasks. For example, on the Switchboard task database, higher perplexity and the additional difficulty of telephone bandwidth result in current word accuracies of only 60 percent. But when speakers know they are talking to a computer rather than conversing with each other, better than 80 percent accuracy can be observed in the scheduling domain.

Figure 3a also compares speech collected using a *push-to-talk switch* versus *free cross-talk dialogues*. While both represent human conversation, cross-talk appears to result in even less well articulated speech and thus is more difficult to recognize and translate than push-to-talk speech. For other languages (English, German, Japanese), Janus-II currently delivers similar word accuracies of 70-plus percent. In recent evaluations carried out by the Verbmobil project using five different recognition engines, 70 percent accuracy was the best achievable for conversational German.

Figure 3b shows the result of end-to-end speech translation performance over a set-aside test set (data not used to train the system). 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. Figure 3b shows that by using the two parsers, Phoenix and GLR*, translation accuracies approaching 85 percent can be achieved for transcribed spoken language and up to 74 percent when using the output from the recognizer (recognition errors included).

Table 1 compares cross-talk and push-to-talk conditions. In both cases the test was carried out using the Phoenix parser over several previously unseen test sets. Human translators report that translating rapid-fire, turn-taking spontaneous dialogue is unacceptably difficult. On the basis of these reports, we predicted that cross-talk speech would also be much harder to recognize and translate by machine. Since we must compare results from different test dialogues (with considerable variability in performance) to check this prediction, we note that precise comparison under equal conditions is not possible. Within our task domain and over multiple

Table 1. Performance comparison for push-to-talk and cross-talk dialogues between humans.

	Speech recognition accuracy	Translation of transcript	Speech-to-speech translation
Push-to-talk data	71%	74%	52%
Cross-talk data	70%	81%	73%

tests, however, a surprising trend appears to emerge. Although cross-talk speech is indeed generally harder to recognize than push-to-talk speech, it results in shorter turns that were found to translate as well or better. Thus, it appears to be no more problematic to translate uninhibited human conversation than controlled turn-taking conversation. Human translators' difficulties with rapid cross-talk dialogue might be related to the human cognitive load of tracking two parallel speech channels rather than any intrinsic translation difficulty in the material.

APPLICATIONS

The need for speech translation arises in different situations, each posing its own challenges and opportunities. We've begun experimenting with three different applications: spoken-language interpretation in an interactive videoconferencing environment, portable speech translation, and simultaneous dialogue translation.

Interactive dialogue translation

Figure 4 shows a prototype videoconferencing station with a spoken-language interpretation facility. There are two displays, one facing the user and another embedded in the desk. The user operates the station via the touch-sensitive display in the desk. A *record* button activates speech acquisition and displays both the recognition result and a paraphrase of the machine-analyzed utterance. This is accomplished by having Janus-II perform a generation from the (language-independent) Interlingua back into the user's language. The user can then determine whether the paraphrase reflects the intended meaning of the input utterance. If it does, pressing a *send* button replaces the paraphrase with the translation into the selected output language and sends it on to the other videoconferencing site, where the translation appears in subtitles under the transmitted video image of the user. It is also synthesized in the target language for speech output. The translation display can be used to run collaborative virtual environments such as joint white-boards or applications both parties make reference to. Translation can be delivered in about real time.

The videoconferencing station is a cooperative translation environment in which both parties are trying to be understood and can verify the system's understanding of a spoken utterance. The station can therefore benefit from user feedback and can more easily ensure correctness. It also offers alternative modes for user input as well as for error recovery. Input options include handwriting, typing, or speech. In case of error, these alternative modalities can be applied to generate a new paraphrase and translation.[3] In this way, effective communication results despite imperfect recognition and translation. In addition to offering various recovery mechanisms, the translation station elicits a somewhat more benign user speaking style than conversational speech between humans.

To exploit this opportunity for error correction, we are exploring several strategies for recovering from human and machine errors.[4] These include repair by respeaking, repair by spelling, and repair by handwriting as alternative redundant modes of human-computer interaction. One or two tries are typically enough for recovery. The Janus-II system offers other simple forms of assistance, such as letting the user type over erroneous recognitions.

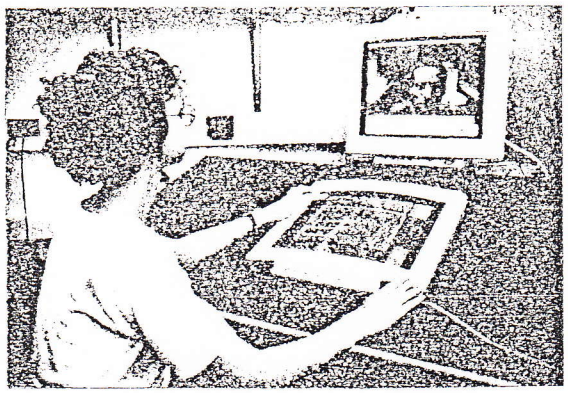


Figure 4. Janus-II speech translator in a videoconferencing environment. Translation, appearing visually as subtitles as well as by synthetic acoustic output, is obtained in about real time. The user controls the system through buttons on a touch-sensitive display. Vocabulary size is between 1,000 and 3,000 words per language.

The interface lets the user select different output languages with language buttons on the translator screen. The user sets the input language at system start-up, or it can be set automatically by a language identification module as a preprocessor. In effect, the system begins by using recognizers for several languages to process an incoming speech utterance, and goodness of match then determines the most probable language.

Many opportunities remain to further study the human factors of interactive spoken-language translation. Researchers continue to investigate the best trade-off between processing speed and accuracy, the role of repair and multimodality, how to deal with out-of-domain utterances, and how to learn and integrate new words or concepts.

Figure 5. Instead of a camera, this "tourist" is wearing a speech translator with a microphone and a head-mounted heads-up display. See-through goggles facilitate an overlaid display of translation output. Alternatively, acoustic output can be presented by earpiece. Current speed is still four times slower than real-time, and the system's vocabulary had to be reduced to 500-plus words per language from a limited domain.

Portable speech translation

The Janette system, a small version of Janus-II, runs on a 75-MHz Pentium laptop with 32 Mbytes of memory. Janette takes about twice as long to translate an utterance as Janus-II. As Figure 5 shows, Janette can be carried in a knapsack or bag. Translation is presented either through an earpiece or on a wearable display. The heads-up display shows the translation in text form on see-through goggles, thereby

allowing the user to see subtitles under the face of the person he or she is talking to. Presenting translation results this way allows greater throughput, as the translation can be viewed without interrupting the speaker. While acoustic output may allow for feedback with the system, a simultaneously presented visual translation may provide greater communication speed. The human factors of such new devices await further study in actual field use.

Passive simultaneous dialogue translation

The language interpreting systems described so far offer the opportunity for feedback, verification, and correction of translation between two willing and cooperative parties. Not every situation affords this possibility, however. Conferences among many users, foreign TV or radio broadcasts, or simultaneous translation of speeches or conversations are passive, uncooperative translation situations in which the speaker cannot verify the translation. Also, with conversational speech, this kind of translation may be particularly difficult, as it requires processing of speech between people, greater coarticulation, and potentially more difficult turn-taking protocols. Indeed, the rapid succession of sometimes overlapping turns makes the cognitive planning of a translation particularly difficult for humans trying to translate conversational dialogue.

THE RESULTS REPORTED in Table 1 for cross-talk and push-to-talk dialogues suggest that the cognitive limitations human translators experience do not hold for machines: Two separate speech-translation processes can operate easily in separate dialogue channels and produce translations that keep up with the speakers. Our lab has installed a conversational translator that slices turns at major breaking points and sends the corresponding speech signals to an array of five processors that incrementally generate translations during a human conversation (again, two subjects negotiating a meeting). Despite the disfluent nature of such an interactive and rapid conversation, conversational dialogues within this domain can be translated accurately more than 70 percent of the time. ■

Acknowledgments

I thank my 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. I also thank the partners and affiliates in C-STAR, who have helped define speech translation today.

References

1. A. Waibel et al., "Janus: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies," *Proc. Int'l Conf. Acoustics, Speech and Signal Processing*, IEEE Press, Piscataway, N.J., 1991.
2. W. Wahlster, "First Results of Verbmobil: Translation Assistance for Spontaneous Dialogues," *Proc. ATR Int'l Worksho Speech Translation*, Kyoto, Japan, Nov. 1993.
3. B. Suhm et al., "Janus: Towards Multilingual Spoken La guage Translation," *Proc. ARPA Spoken Language Technolog Workshop*, Austin, Tex., Jan. 1995. Also published by Morga Kaufmann, San Francisco.
4. A.E. McNair and A. Waibel, "Improving Recognizer Acce tance through Robust, Natural Speech Repair," *Proc. Int Conf. Speech and Language Processing*, Acoustical Soc. Japan, Yokohama, Japan, 1994, pp. 1,299-1,302.
5. S. Nirenburg et al., *Machine Translation: A Knowledge-Bas Approach*, Morgan Kaufmann, San Mateo, Calif., 1992.
6. E.H. Hovy, "How MT Works," *Byte*, Jan. 1993, pp. 167-176
7. H. Hild and A. Waibel, "Integrating Spelling into Spoken Di logue Recognition," *Proc. Eurospeech 95*, Vol. 2, Imperial Pre Ltd., Mississauga, Canada, pp. 1,977-1,980.
8. W. Ward, "Understanding Spontaneous Speech: The Phoen System," *Proc. Int'l Conf. Acoustics, Speech and Signal Pr cessing*, 1991, pp. 365-368.
9. A. Lavie and M. Tomita, "GLR*—An Efficient Noise-Skippi Parsing Algorithm for Context-Free Grammars," *Proc. Th Int'l Workshop Parsing Technologies*, 1993, p. 123.
10. M.T. Vo et al., "Multimodal Learning Interfaces," *Proc AR Spoken Language Technology Workshop*, Morgan Kaufman San Francisco, 1995.

Alex Waibel *is a professor of computer science at Karlsru University, Germany, and a principal research computer s entist in the School of Computer Science at Carnegie Mell University in Pittsburgh. He directs the Interactive Syste Laboratories at Carnegie Mellon and at the University Karlsruhe. At Carnegie Mellon he also serves as director the PhD program in language and information technolog and as associate director of the Language Technologies In tute. He also holds joint appointments in the Robotics In tute and the Human-Computer Interaction Institute, both Carnegie Mellon.*

Waibel received a BS from the Massachusetts Institute Technology, and an MS in electrical engineering and compu science and a PhD in computer science from Carnegie Mell

Contact Waibel at the School of Computer Science, Carne Mellon University, Pittsburgh, PA 15213, or at the Fakul für Informatik, Universität Karlsruhe, D-76131 Karlsru Germany.

Web sites

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