[from: *Survey of the state of the art in human language technology*, ed. Giovanni Battista Varile & Antonio Zampolli (Pisa: Giardini, 1997); pp. 266-272]

8.6 Multilingual Speech Processing

Alexander Waibel
Carnegie-Mellon University, Pittsburgh, Pennsylvania, USA
and Universität Karlsruhe, Germany

Multilinguality need not be textual only, but 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 within a country or travelers from abroad. Public service operators (emergency, police, department of transportation, telephone operators, and others) in the US, Japan and the EU frequently receive requests from foreigners unable to speak the national language (see also section 8.7.1).

Multilingual spoken language services is a growing industry, but so far these services rely exclusively on human operators. Telephone companies in the United States (e.g., AT&T Language Line), Europe and Japan now offer language translation services over the telephone, provided by human operators. Movies and foreign television broadcasts are routinely translated and delivered either by lipsynchronous speech (dubbing), subtitles or multilingual transcripts. The drive to automate information services, therefore, produces a growing need for automated multilingual speech processing.

The difficulties of speech processing are compounded with multilingual systems, and few if any commercial multilingual speech services exist to date. Yet intense research activity in areas of potential commercial interest are underway. These are aiming at:

- **Spoken Language Identification** By determining a speaker's language automatically, callers could be routed to human translation services. This is of particular interest to public services such as police, government offices (immigration service, drivers license offices, etc.) and experiments are underway in Japan and some regions of the US. The technical state of the art will be reviewed in the next section;
- **Multilingual Speech Recognition and Understanding** Future Spoken Language Services could be provided in multiple languages. Dictation systems and spoken language database access systems, for example, could operate in multiple languages, and deliver text or information in the language of the input speech.
- **Speech Translation** This ambitious possibility is still very much a research area, but could eventually lead to communication assistance in the form of portable voice activated dictionaries, phrase books or spoken language translators, telephone based speech translation services and/or automatic translation of foreign broadcasts and speeches. There is a wide spectrum of possibilities, but their full realization as commercial products still requires considerable research well into the next decade and beyond.

8.6.1 Multilingual Speech Recognition and Understanding

The last decade has seen much progress in performance of speech recognition systems from cumbersome small vocabulary isolated word systems to large vocabulary continuous speech recognition (LV-CSR) over essentially unlimited vocabularies (50,000 words and more). Similarly, spoken language understanding systems now exist that process spontaneously spoken queries, although only in limited task domains under benign recording conditions (high quality, single speaker, no noise). A number of researchers have been encouraged by this state of affairs to extend these systems to other languages. They have studied similarities as well as differences across languages and improved the universality of current speech technologies.

Large Vocabulary Continuous Speech Recognition (LV-CSR)

A number of LV-CSR systems developed originally for one language have now been extended to several languages, including systems developed by IBM (Cerf-Danon, DeGennaro, et al., 1991), Dragon Systems (Bamberg, Demedts, et al., 1991), Philips and Olivetti (Ney & Billi, 1991) and LIMSI. The extension of these systems to English, German, French, Italian, Spanish, Dutch and Greek illustrates that current speech technology does generalize to different languages, provided sufficiently large transcribed speech databases are available. The research results show that similar modeling assumptions hold across languages with a few interesting exceptions. Differences in recognition performance are observed across languages, partially due to greater acoustic confusability (e.g., English), greater number of homonyms (e.g., French) and greater number of compound nouns and inflections (e.g., German). Such differences place a different burden on acoustic modeling vs. language modeling, vs. the dictionary, or increase confusability, respectively. Also, a recognition vocabulary is not as easily defined as a unit for processing in languages such as Japanese and Korean, where pictographs, the absence of spaces, and large numbers of particles complicate matters.

Multilingual Spoken Language Systems

While LV-CSR systems tackle large vocabularies, but assume benign speaking styles (read speech), spoken language systems currently assume smaller domains and vocabularies, but require unrestricted speaking style. Spontaneous speech significantly degrades performance over read speech as it is more poorly articulated, grammatically ill-formed and garbled by noise. ARPA's Spoken Language projects have attacked this problem by focusing increasingly on the extraction of the semantic content of an utterance rather than accurate transcription. One such system, that has recently been extended to other languages is MIT's Voyager system (Glass, Goodine, et al., 1993). It was designed to handle information delivery tasks and can provide directions to nearby restaurants in Cambridge and also for airline travel information (ATIS). It has recently been extended to provide output in languages other than English. Researchers at LIMSI have developed a similar system for French (also airline travel information), thereby providing extension to French on the input side as well. Availability of recognition capabilities in multiple languages have also recently led to interesting new language, speaker and gender identification strategies (Gauvain & Lamel, 1993). Transparent language identification could enhance the application of multilingual spoken language systems (see also section).

Despite the encouraging beginnings, multilingual spoken language systems still have to be improved before they can be deployed on a broad commercially feasible scale. Prototype systems have so far only been tested in benign recording situations, on very limited domains, with cooperative users, and without significant noise. Extending this technology to field situations will require increases in robustness as well as consideration of the human factors aspects of multilingual interface design.

8.6.2 Speech Translation Systems

There are no commercial speech translation systems in operation to date, but a number of industrial and government projects are exploring their feasibility. The feasibility of speech translation depends largely on the scope of the application, and ranges from applications that are well within range -such as voice activated dictionaries- to those that will remain impossible for the foreseeable future (e.g., unrestricted simultaneous translation.) Current research therefore aims at milestones between these extremes, namely limited domain speech translation. Such systems restrict the user in what he/she can talk about, and hence constrain the otherwise daunting task of modeling the world of discourse. Nevertheless such systems could be of practical and commercial interest, as they could be used to provide language assistance in common yet critical situations, such as registration for conferences, booking hotels, airlines, car rentals and theater tickets, ordering food, getting directions, scheduling meetings or in medical doctor-patient situations. If successful, it may also be possible to combine such domains to achieve translation in a class of domains (say, travel).

To be sure, spoken language translation---even in limited domains---still presents considerable challenges, which are the object of research in several large research undertakings around the world. Translation of spoken language (unlike text) is complicated by syntactically ill-formed speech, human (cough, laughter, etc.) and non-human (door-slams, telephone rings, etc.) noise, and has to contend with speech recognition errors. The spoken utterance does not provide unambiguous markers indicating the beginning or end of a sentence or phrase, and it frequently contains irrelevant information, that need not or should not be translated. Even simple concepts are expressed in quite different ways in different languages. A successful system must therefore interpret the speaker's intent -instead of translating his/her words- and deliver an appropriate message in the target language. For the speech processing components of a speech recognition system high accuracy is not the primary or only area of concern, but understanding, and understanding may be achieved by selectively extracting words of interest, and/or by occasionally prompting the user for important information. Researchers are now exploring solutions to the problem as a whole without expecting each separate part to function perfectly.

A speech translation system can also not be categorized uniquely as either ``translation for assimilation'' nor as ``translation for dissemination'', as textual translation systems are frequently described. It has some of the characteristics of both. Aiming at the interpretation of a speaker's intent, some research avenues in speech translation are attempting to extract (assimilate) the key information to interpret the gist of an utterance. Yet spoken language in many of the targeted application scenarios involves the interaction between two cooperative speakers, who can control to some extent the input to produce the desired result. This may allow for some limited domain systems to interact with the speaker of the source language until the correct interpretation can be transmitted (disseminated) in the target language(s).

A further complicating factor currently under investigation is that speech translation involves aspects of both human-to-human, as well as human-machine (the interpreting system) dialogues. This may require a system to distinguish between utterances and meta-level utterances, and to deal with code switching (change of language) in case of speakers with partial knowledge of each others' language or when making reference to objects, names or items in the other language. Experiments over several speech databases in several languages indicate that human-to-human speech contains more disfluencies, more speaking rate variations and more coarticulation resulting in lower recognition rates (Levin, Suhm, et al., 1994) than human-machine interaction. These difficulties require further technological advances, a rethinking of common speech and language processing strategies, and a closer coupling between the acoustic and linguistic levels of processing.

Early Systems:

Speech Translation research today is being developed against the background of early systems implemented in the eighties to demonstrate the feasibility of the concept. In addition to domain limitations, these early systems had also fixed speaking style, grammatical coverage and vocabulary size and were therefore too limited to be of practical value. Their system architecture is usually strictly sequentially, involving speech recognition, language analysis and generation, and speech synthesis in the target language. Developed at industrial and academic institutions and consortia, they represented a modest but significant first step and proof of concept that multilingual communication by speech might be possible. Systems include research prototypes developed by NEC, AT&T, ATR, Carnegie Mellon University, Siemens AG, University of Karlsruhe, and SRI. Most have arisen or been made possible through international collaborations that provide the cross-linguistic expertise.

Among these international cooperations, the Consortium for Speech TrAnslation Research (C-STAR) was formed as a voluntary group of institutions committed to building speech translation systems. Its early members, ATR Interpreting Telephony Laboratories (now ``Interpreting Telephony Laboratories'') in Kyoto, Japan, Siemens AG in Munich, Germany, Carnegie Mellon University (CMU) in Pittsburgh, USA, and University of Karslruhe (UKA) in Karlsruhe, Germany, developed early systems, that accepted speech in each of the members' languages (i.e., English, German and Japanese) and produced output text in all the others (Morimoto, Takezawa, et al., 1993; Waibel, Jain, et al., 1991; Woszczyna, Aoki-Waibel, et al.,

1994). The system modules allowed for continuous speaker-independent (or adaptive) input from a 500 word vocabulary in the domain of conference registration. The systems' modules operated strictly sequential, did not allow for feedback, and only accepted syntactically well formed utterances. After speech recognition, language analysis and generation, output text could then be transmitted to each of the partners sites for synthesis there. Translation was performed by an Interlingua approach in JANUS, the CMU/UKA system, while a transfer approach was used in ATR's ASURA and Siemens's systems. In early '93, they were shown to the public in a joint demonstration using video conferencing. Given the restrictions on speaking style and vocabulary, the systems performed well and provided good translation accuracy.

Early industrial speech-translation efforts are illustrated by AT&T's VEST (Roe, Pereira, et al., 1992) and NEC's Intertalker systems. VEST resulted from a collaboration between AT&T and Telefonica in Spain and translated English and Spanish utterances about currency exchange. It uses a dictionary of 374 morphological entries and an augmented phrase structure grammar that is compiled into a finite state grammar used for both language modeling and translation. The system was demonstrated at EXPO'92 in Seville, Spain. NEC's Intertalker system also used finite state grammars to decode input sentences in terms of prescribed sentence patterns. The system ran on two separate tasks: reservation of concert tickets and travel information, and was successfully demonstrated at GlobCom'92. SRI in collaboration with Swedish Telecom recently reported on another system (Rayner et al., 1993), that is based on previously developed system components from SRI's air travel information system. The ATIS speech understanding component is interfaced with a generation component. The system's input language is English and it produces output in Swedish. It represents an early attempt at extending spontaneous multilingual human-machine dialogues to translation.

Translation of Spontaneous Speech:

To develop more practical, usable speech translation, greater robustness in the face of spontaneous ill-formed speech has to be achieved. A number of research activities aiming at the translation of spontaneous speech have since been launched. Several industrial and academic institutions, as well as large national research efforts in Germany and in Japan are now working on this problem. Virtually all of these efforts aim at restricted domains, but now remove the limitation of a fixed vocabulary and size, and also no longer require the user to speak in syntactically well-formed sentences (an impossibility in practice, given stuttering, hesitations, false starts and other disfluencies found in spontaneous speech).

The C-STAR consortium was extended to translate spontaneous speech. In addition to the partners of the first phase, it includes presently ETRI (Korea), IRST (Italy), LIMSI (France), SRI (UK), IIT (India), Lincoln Labs (USA), MIT (USA), and AT&T (USA). Each C-STAR partner builds a complete system that at the very least accepts input in the language of this partner and produces output in one other language of the consortium. In a multinational consortium, building full systems thereby maximizes the technical exchange between the partners while minimizing costly software/hardware interfacing work. C-STAR continues to operate in a fairly loose and informal organizational style. Present activity has shifted toward a greater emphasis on interpretation of spoken language, i.e., the systems ability to extract the intent of a speakers utterance. Several institutions involved in C-STAR therefore stress semantic parsers and an interlingual representation (CMU, UKA, MIT, ATT, ETRI, IRST), more in line with message extraction than with traditional text translation. Other approaches under investigation include Example Based Translation (ATR), with its potential for improved portability and reduced development cost through the use of large parallel corpora. Robust Transfer Approaches (ATR, Siemens) are also explored, with robust and stochastic analysis to account for fragmentary input. System architectures under investigation are no-longer strictly sequential, but begin to involve clarification or paraphrase in the speaker's language as first attempts at the machine's feedback of its understanding. At the time of this writing, such feedback is still very rudimentary and does not yet involve more elaborate confirmatory meta-level dialogues or repair mechanisms. Current research also begins to actively exploit discourse and domain knowledge, as well as prosodic information during turn taking, for more robust interpretation of ambiguous utterances.

Verbmobil is a large new research effort sponsored by the BMFT, the German Ministry for Science and Technology (Wahlster, 1993). Launched in 1993 the program sponsors over 30 German industrial and academic partners who work on different aspects of the speech translation problem and are delivering system components for a complete speech translation system. The system components (e.g., speech recognition components, analysis, generation, synthesis, etc.) are integrated into a research prototype, available to all. The initial task is appointment scheduling with possible extensions to other domains. Verbmobil is aimed at face-to-face negotiations, rather than telecommunication applications and assumes that two conversants have some passive knowledge of a common language, English. It is to provide translation on demand for speakers of German and Japanese, when they request assistance in an otherwise English conversation. Verbmobil is therefore concerned with code switching and the translation of sentence fragments in a dialog. Verbmobil is an eight-year project with an initial four-year phase.

8.6.3 Future Directions

To meet the challenges in developing multilingual technology, an environment and infrastructure must be developed. Contrary to research fostered and supported at the national level, multilingual research tends to involve cooperations across national boundaries. It is important to define and support efficient, international consortia, that agree to jointly develop such mutually beneficial technologies. An organizational style of cooperation with little or no overhead is crucial, involving groups who are in a position to build complete speech translation systems for their own language. There is a need for common multilingual databases and data involving foreign accents. Moreover, better evaluation methodology over common databases is needed to assess the performance of a speech translation systems in terms of accuracy and usability. Research in this direction needs to be supported more aggressively across national boundaries.

Beyond improvements in component technologies (speech and language processing), innovations in language acquisition are badly needed to achieve greater portability across domains. While acoustic models can be reused to a certain extent (or at least adapted) across domains, most language work still requires inordinate amounts of resources. Grammar development requires considerable development work for each domain. Language models have to be retrained and require large amounts of transcribed data within each domain. Continued research on language acquisition may provide better domain adaptation, and/or incrementally improving language models, grammars and dictionaries.

The limitation to restricted domains of discourse must be lifted, if broader usage is to be guaranteed. Short of universal and reliable speech translation (as could be needed for example, for automatically translated captions in movies, or simultaneous translation), intermediate goals might be given by large domains of discourse, that involve several subdomains. Integration of subdomains will need to be studied.

Last, but not least, better human-computer interaction strategies have to be developed, as multilingual spoken language translation becomes a tool to broker an understanding between two humans rather than a black box that tries to translate every utterance. A useful speech translation system should be able to notice misunderstandings and negotiate alternatives. Such ability requires better modeling of out of domain utterances, better generation of meta-level dialogues and handling of interactive repair.

References

Bamberg, P., Demedts, A., Elder, J., Huang, C., Ingold, C., Mandel, M., Manganaro, L. and van Even, S. (1991). Phoneme-based training for large-vocabulary recogntition in six european languages. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, volume 1, pages 175-181, Genova, Italy. European Speech Communication Association.

Cerf-Danon, H., DeGennaro, S., Ferreti, M., Gonzalez, J., and Keppel, E. (1991). Tangora – a large vocabulary speech recognition system for five languages. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, volume 1, pages 183-192, Genova, Italy. European Speech Communication Association.

Gauvain, J-L. and Lamel, L.F. (1993). Identification of non-linguistic speech features. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*, page Session 6, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Glass, J., Goodine, D., Phillips, M., Sakai, S., Seneff, S., and Zue, V. (1993). A bilingual voyager system. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann. Session 6.

Levin, L., Suhm, B., Coccaro, N., Carbonell, J., Horiguchi, K., Isotani, R., Lavie, A., Mayfield, L., Rose, C.P., Van Ess-Dykema, C., and Waibel, A. (1994). Speech--language integration in a multi-lingual speech translation system. In *Proceedings of the 1994 AAAI Conference*, Seattle, 1994. American Association for Artificial Intelligence.

Morimoto, T., Takezawa, T., Yato, F., Sagayama, S., Tashiro, T., Nagata, M., and Kurematsu, A. (1993). ATR's speech translation system: ASURA. In *Proceedings of the Third Conference on Speech Communication and Technology*, pages 1295-1298, Berlin, Germany, September 1993.

Ney, H. and Billi, R. (1991). Prototype systems for large-vocabulary speech recognition: Polyglot and Spicos. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, volume 1, pages 193-200, Genova, Italy. European Speech Communication Association.

Rayner, M. et al. A speech to speech translation system built from standard components. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Roe, D.B., Pereira, F.C., Sproat, R.W., and Riley, M.D. (1992). Efficient grammar processing for a spoken language translation system. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 213--216, San Francisco, March 1992. Institute of Electrical and Electronic Engineers.

Wahlster, W. (1993). Verbmobil, translation of face-to-face dialogs. In *Proceedings of the Fourth Machine Translation Summit*, pages 127--135, Kobe, Japan, 1993.

Waibel, A., Jain, A., McNair, A., Saito, H., Hauptmann, A., and Tebelskis, J. (1991). JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 793-796, Toronto. Institute of Electrical and Electronic Engineers.

Woszczyna, M., Aoki-Waibel, N., Buo, F.D., Coccaro, N., Horiguchi, K., Kemp, T., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C.P., Schultz, T., Suhm, B., Tomita, M., and Waibel, A. (1994). Towards spontaneous speech translation. In *Proceedings of the 1994 International Conference on Acoustic, Speech, and Signal Processing*, volume 1, pages 345-349, Adelaide, Australia. Institute of Electrical and Electronic Engineers.