Alex Waibel and Christian Fügen

Spoken Language Translation

Enabling cross-lingual human–human communication

During the past 15 years, speech translation has grown from an oddity at the fringe of speech and language processing conferences to one of the main pillars of current research activity. The expanding interest and excitement can be explained by a convergence of emerging and powerful new technical capabilities and a growing appreciation of the needs for better cross-lingual communication in a globalizing world [31].

Governments, commercial enterprises, and academic and humanitarian organizations all face internationalization and globalization at an unprecedented scale. Security, effectiveness in trade and commerce, market size, and competitive reach all depend on global information awareness and the ability to interact and communicate globally. Increased integration (e.g., witness the integration efforts in Europe and Asia) requires natural, yet effective, international cross-lingual communication. It is true that there are common languages to communicate (English, Spanish, Mandarin) among certain language groups, but language abilities vary and often prevent true integration and equal opportunities for all. Effective solutions addressing the linguistic divide (not just the "digital divide") could therefore offer considerable practical and economic benefits.

For the research community, speech translation also presents fascinating new problems that appear solvable by the introduction of considerable computing resources, seemingly unlimited Web data, and promising new machine learning techniques. Despite the promise and potential, considerable improvements are still needed in the component technologies: speech recognition, machine translation (MT), and speech synthesis. Moreover, to achieve effective cross-lingual human–human communication in practice, not only do recognition and translation error rates matter but also the user interface and overall system design in each communication setting.

In the following, we present an overview of the field of speech translation. We review the history of the field, the main achievements, and remaining challenges. We discuss the main approaches and the most promising applications. We also address the human factors of delivering and deploying speech translation in different human communicative scenarios and discuss issues regarding scaling the technology across domains, speaking styles, and languages.

Digital Object Identifier 10.1109/MSP.2008.918415

1053-5888/08/$25.00©2008IEEE

TECHNOLOGY

Speech translation systems typically consist of three components (see Figure 1): automatic speech recognition (ASR), MT, and text-to-speech synthesis (TTS). The underlying technologies for these three components have been developed independently and many of their performance issues and techniques are used to apply to speech translation as well. Clearly, better ASR, MT, or TTS performance makes for better speech translation performance. However, a speech translation system is not only the cascade of its parts. Since the goal is to produce output in a target language, the correctness of the components' output is of secondary concern. Uncertainty at the component level can be addressed by being noncommittal at their interface, linking components via near-miss hypothesis lattices [1]. Such a view also offers the possibility of jointly optimizing components based on the overall output rather than each component independently [2].

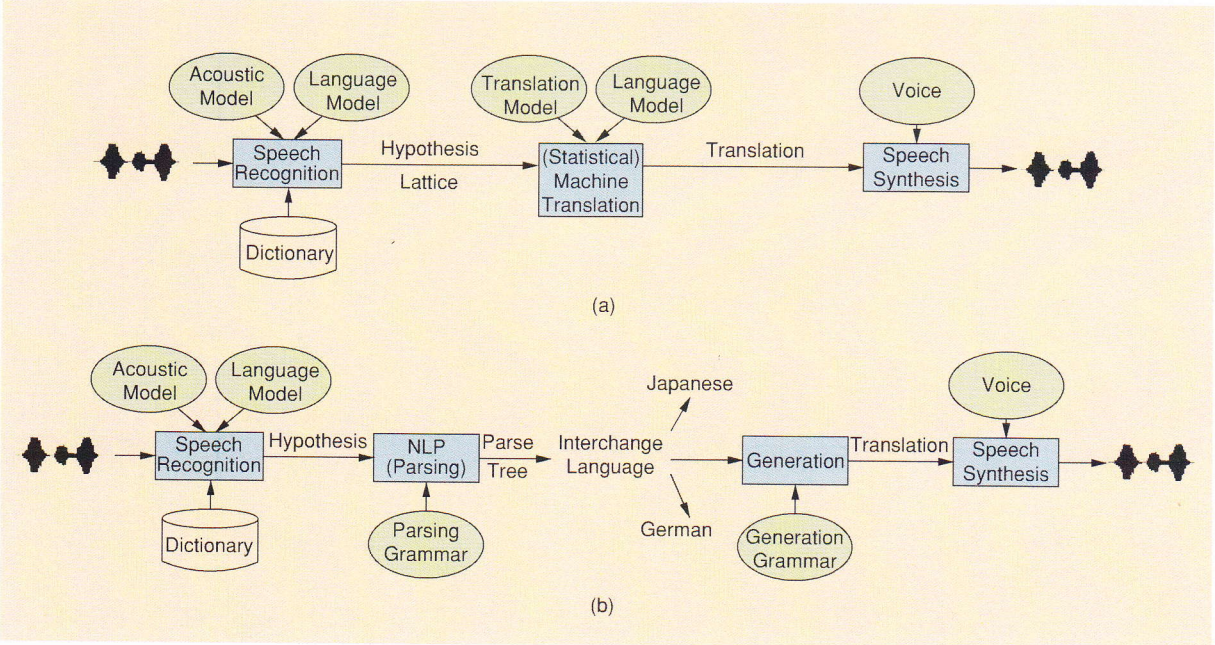
TO ERR IS HUMAN, AND USEFUL SYSTEMS MUST ACCOMMODATE A SPEAKER'S MISTAKES.

AUTOMATIC SPEECH RECOGNITION

The ASR component of a speech translation system, of course, faces all the challenges that are typically for ASR in general: noise, disfluencies, vocabulary size, and language perplexity, which complicate recognition and increase recognition errors. Recognition quality is generally measured in terms of word error rate (WER) as compared with a reference transcription. In addition to WER, other factors must be included to judge the capabilities of a recognizer: the language model perplexity (a measure of information/surprise provided by the word sequence), speed, memory usage, processor requirements, and microphone positioning. For speech translation a few of these challenges are particularly noteworthy.

First off, ASR error rates should generally be lower for translation to make sense (~10%) than for other ASR applications that can tolerate higher error rates (e.g., retrieval). Since many speech translation applications involve free spontaneous human dialog, however, such low error rates are more difficult to obtain: Spontaneous dialogs tend to be disfluent, containing false starts, hesitations, repetitions, and spontaneous speech, which is less articulated, leading to higher error rates. Speech translation dialogs also often involve accents, as regional variations and cross-language expressions enter into use. Furthermore, many speech translation tasks are further affected by environmental noise, or by issues deriving from microphone positioning and type. In a two-way system for doctor–patient dialogs, for example, it may be feasible for the owner of a system to wear a close speaking microphone but not for the patient.

The ASR component must also provide a useful indication of sentence boundaries so that a subsequent translation engine can translate a sentence or fragment into another language. In many tasks, a continuous stream of voice (broadcast news, speeches, lectures, etc.) is presented so that such sentence-level segments must be inferred automatically. The resulting segmentation algorithms use natural pauses, language model statistics, and prosodic cues to infer such segments. Optimizing subsequent translation



[FIG1] Schematic overview of a speech-to-speech translation system and its models: (a) direct approach using, e.g., statistical MT, and (b) interlingua approach using an interchanged language.

quality and minimizing latency are both important considerations for an effective design.

To improve ASR modeling, more training and/or adaptation data are required, but conversational dialog data are hard to collect and "natural" dialogs through a speech translation device are difficult to simulate. Similarly, text data for language modeling and dictionary construction may be available for certain speech translation tasks (broadcast news, parliamentary speeches) but not for others (dialogs, lectures). Construction of a suitable recognition "word" lexicon can also be a problem if a language provides for many inflections of its root forms (morphology). Depending on language, text data for language model training and dictionary constructions may in fact not even be available at all if the language is a spoken language or a dialect.

DOMAIN LIMITATION MAY BE ACCEPTABLE IN CERTAIN TASKS AND ENVIRONMENTS (TOURISM, MEDICAL ASSISTANCE, ETC.), BUT FOR OTHERS IT IMPOSES TOO GREAT A RESTRICTION TO BE USEFUL.

ASR has a number of added practical requirements that are of special importance to speech translation. These include speed requirements when dialog completion is at stake or proper handling of named entities (city names, person names, food, medication, symptoms, etc.) as they vary in the field and application and are essential for translation. Since effective end-to-end communication is the goal, ASR components frequently output hypothesis lattice structures (confusion networks) and confidence measures to pass multiple near-miss alternatives to subsequent translation components. These allow for better integration and overall system optimization.

SPOKEN LANGUAGE TRANSLATION

For MT of text, the choice of technology and design remains a topic of discussion. Three different approaches have been popular in MT: the direct approach, the transfer approach, and the interlingua approach. In the first, a direct mapping between source language and target language is attempted, while transfer and interlingua approaches attempt to extract deeper linguistic structures first. Most commonly, transfer approaches will perform a syntactic analysis and transfer the derived structures from the source to the target language for generation of the target language sentence. Interlingua approaches [3] attempt to derive a semantic representation of an input sentence first and then generate a sentence in the target language from those semantic concepts. Direct approaches bypass this analysis and map input sentences directly onto a target language sentence. While early attempts at direct translation were rejected due to the high ambiguity of language, they have regained considerable following and popularity with the advent of statistical data-driven approaches, such as example based and statistical MT [4], [23].

All three MT approaches have been used for speech translation as well, each with notable advantages and disadvantages [5], [9]–[11], [13]. The interlingua approach has the advantage that it can connect N languages in any combination through its common semantic representation and therefore does not require the development of $O(N^2)$ language translators. It also permits regeneration of a paraphrase in a speaker's own language for verification. A semantic representation also strips the input surface realization from all its disfluencies and colloquialisms and can lead to a clean and semantically equivalent utterance in the target language. The biggest drawback of the interlingua approach is the manual development of semantic parsers and the complications in designing a semantic representation common to all languages. Statistical MT, by contrast, can handle the ambiguities of language by a stochastic source channel model, much like today's speech recognizers do. With it, the most likely target language word string $\hat{e}$ given a source language word string f is estimated by way of Bayes' rule as the product of a translation model $p(f|e)$ and a language model $p(e)$:

$$\hat{e} = \arg\max_{e}\{p(e|f)\} = \arg\max_{e}\{p(e)p(f|e)\}.$$

Effectively, the model combines the probabilities of different translations of words in a sentence with the monolingual likelihood of each resulting word sequence to determine the most likely translation of that sentence. Instead of just modeling this as a noisy channel approach, current SMT systems use a log linear combination of a number of feature functions that model important aspects including a language model, a word reordering model, word penalties, and various other translation models [24], [25], [27]. The SMT approach has the advantage that it requires no manual development of grammars or representations but is trained on large amounts of translated reference texts (parallel corpora). Its drawback is its need for large parallel corpora, its lack of a common representation to connect multiple languages, and challenges in view of highly disfluent input.

For domain-limited translation systems (see discussion below) the design of an interlingua has been shown to be possible and helpful, but for domain-unlimited applications (due to their unrestricted semantic coverage) SMT methods have been generally preferred. A number of hybrid techniques have been proposed to retain some of the advantages of both, including statistically trained semantic analyzers in an interlingua framework [6], or using a natural language (e.g., English) as an intermediate "pivot" language [7], [8] to connect multiple languages.

OUTPUT GENERATION (SPEECH, TEXT)

The output of a speech translation system most typically is synthetic speech in the target language. Alternative outputs, however, are possible depending on the purpose and ultimate use of

the speech translator (see discussion below). They include target language text, or summaries from translations. In human–human cross-lingual speech dialogs a speech synthesis component generates audible output from a translated text string [30]. Commonly, full TTS is used for convenience and modularity, even though one could arguably also synthesize speech based on conceptual or syntactic structures if they are provided by the MT component. Special concerns in TTS for speech translators involve generating appropriate emotion, style, and voices, so that the output speaking style corresponds to the input speech in the source language [30]. Voice conversion, in particular, attempts to generate speech in the output language with the voice of the speaker of the input language.

THE FIELD HAS PROGRESSED TO DATE FROM HIGHLY RESTRICTIVE DEMONSTRATION SYSTEMS TO FREE SIMULTANEOUS TRANSLATION OF SPONTANEOUS SPEECH ABOUT UNLIMITED TOPICS, PUSHING BACK ON EACH OF THE RESTRICTIONS SUCCESSIVELY.

PROGRESS IN SPEECH TRANSLATION

Based on the advances in component technologies, research on speech translation began in earnest during the late 1980s and early 1990s. In the following two decades, impressive speech translation systems have been developed. The systems and their progression can be categorized by distinct new system-level capabilities at each stage of development. These capabilities are summarized in Table 1.

Overall, each advance distinguishes itself by the levels of uncertainty that a given system can tolerate. Language is ambiguous at all levels, from signal to phonetics to syntax to semantics. Earlier systems have imposed greater constraints to control such ambiguity. For example, restrictions in speaking style, vocabulary, domain, and the use and operation of a system limit ambiguity and the search for translation hypotheses. Such constraints are inherent in the task (prerecorded announcements, limited domain, or phraseology) or recording conditions (e.g., broadcast news versus telephone conversations). Alternatively, it can be imposed as a requirement for system use. Restrictions on use, however, severely limit the usefulness of a system in many real-world situations. A domain-limited one-way system for tourists may be helpful but is limiting, as it requires the user to memorize the allowable phrases and it cannot translate back the response of the other party. It is equally limiting if the user is not allowed a hesitation (aeh, hum, etc.) while speaking, or if he/she must produce perfectly grammatical sentences to obtain useful output. To err is human, and useful systems must accommodate a speaker's mistakes. Finally, domain limitation may be acceptable in certain tasks and environments (tourism, medical assistance, etc.), but for others it imposes too great a restriction to be useful. Translation of broadcast news and speeches, for example, is only possible if a system can accommodate or adapt to a broad variety of topics, an unlimited vocabulary, and free speaking style.

Dimensions that increase uncertainty and ambiguity in speech and hence present challenges for speech translation systems are signal degradation/noise, vocabulary size/perplexity, spontaneity/disfluencies/speaking style, domain size, and speed requirements. Consequentially, the field has progressed to date from highly restrictive demonstration systems to free simultaneous translation of spontaneous speech about unlimited topics, pushing back on each of the restrictions successively.

RESTRICTED DOMAIN, RESTRICTED SPEAKING STYLE

The first speech translation systems date back to the late 1980s and early 1990s [5], [9], [12]. They were demonstration systems that showed the concept of speech translation and proved that speech translation was possible at all. They attracted a great deal of attention, as they showed that bridging the language divide by spoken language might indeed be possible [32]. These early systems did not permit free dialog and required speakers to act out prescribed sentence patterns or allowable sentences in a correct speaking style according to a

[TABLE 1] SUMMARY OF SYSTEM-LEVEL CAPABILITIES.

	YEARS	VOCABULARY	SPEAKING STYLE	DOMAIN	SPEED	PLATFORM	EXAMPLE SYSTEMS
FIRST DIALOG DEMONSTRATION SYSTEMS	1989–1993	RESTRICTED	CONSTRAINED	LIMITED	2–10 × RT	WORKSTATION	C-STAR-I
ONE-WAY PHRASEBOOKS	1997–PRESENT	RESTRICTED, MODIFIABLE	CONSTRAINED	LIMITED	1–3 × RT	HANDHELD	PHRASELATOR, ECTACO
SPONTANEOUS TWO-WAY SYSTEMS	1993–PRESENT	UNRESTRICTED	SPONTANEOUS	LIMITED	1–5 × RT	PC/HANDHELD DEVICES	C-STAR, VERBMOBIL, NESPOLE, BABYLON, TRANSTAC
TRANSLATION OF BROADCAST NEWS, POLITICAL SPEECHES	2003–PRESENT	UNRESTRICTED	READ/ PREPARED SPEECH	OPEN	OFFLINE	PCS, PC-CLUSTERS	NSF-STRDUST, EC TC-STAR, DARPA GALE,
SIMULTANEOUS TRANSLATION OF LECTURES	2005–PRESENT	UNRESTRICTED	SPONTANEOUS	OPEN	REALTIME	PC, LAPTOP	LECTURE TRANSLATOR

restricted syntax and/or a restricted vocabulary. Nevertheless, they were systems that were proposed at a time when the idea of speaker-independent, continuous speech was still a novelty and MT was considered close to impossible.

DOMAIN-LIMITED, SPONTANEOUS SPEECH

In 1992, it was already recognized that these early concept demonstration systems fall short of being usable, as speakers had to speak in a well-behaved manner and remember the words and sentences they would be allowed to say. The most unacceptable constraints were the vocabulary, syntax, and speaking-style limitation, as it is generally not possible for humans to speak flawlessly in a limited speaking style (effectively reading sentences) or remember a limited set of words or syntactic patterns. By contrast, it is generally possible for humans to stick to a domain of discourse when solving certain limited tasks. Many important applications are inherently domain limited, thus making spontaneous domain-limited speech translation a practically useful technology. Hotel bookings, car rentals, taxis and shopping negotiations, medical assistance, emergency relief, hotel/hospital/conference registration, force protection, military/police missions, and many more all require only dialogs in a limited domain. But they do require accuracy, speed, and an acceptable human-factors design.

More advanced technology was developed to address the limitation of speaking style: two-way dialogs handling free spontaneous speech input, both in recognition and translation. Spontaneous speech is a requirement for two-way dialog systems as the input of the respondent cannot be controlled or restricted. For spontaneous dialogs, we must relax syntactic constraints and allow for variations in expression. Two approaches have been popular: the interlingua approach and the direct statistical approach. The former semantic constraints can be exploited to extract possible interpretation in fragmented input. For limited-domain applications, this is possible where the typical concepts and arguments can be enumerated and represented. The statistical approach, by contrast, accommodates ill-formed input by using large translation and language models to compute the statistically most likely word sequence [13]. The first spontaneous speech translation systems were demonstrated in the early 1990s under the Consortium for Speech Translation Advanced Research (C-STAR) (http://www.c-star.org) [9], [10], [16]. Considerable work continued throughout the 1990s in Japan, Europe, and the United States until today, with large consortia and national projects supporting research [C-STAR, Verbmobil (http://verbmobil.dfki.de), Negotiating through Spoken Language in E-Commerce (Nespole) (http://nespole.itc.it), Enthusiast, Digital Olympics, Babylon, and Spoken Language Communication and Translation System for Tactical Use (Transtac) (http://www.darpa.mil/ipto/programs/transtac/transtac.asp)].

[FIG2] Phraselator (courtesy of Voxtec International, Annapolis, MD, http://www.voxtec.com).

[FIG3] A PDA two-way pocket translator (English-Thai) (courtesy of Mobile Technologies, LLC, Pittsburgh, PA, http://www.mobytrans.com).

PORTABLE, FIELDABLE SYSTEMS

More recently, portable, fieldable speech-to-speech translation systems have been developed around wearable platforms (laptops, PDAs). This may impose additional hardware-related constraints on the ASR, SMT, and TTS components. For PDAs, memory limitations and the lack of a floating-point unit require redesign of algorithms and data structures. Thus, the recognition and translation accuracy of PDA-based speech-to-speech translation systems may decrease compared to systems developed for laptops. In addition to continued attention to speed, recognition, translation, and synthesis quality, usability of the user interface, microphone type, place and number, user training, and field maintenance must be considered. Figure 2 and Figure 3 show two mobile speech translators. The first, the Phraselator (http://www.sarich. com/translator), is a pragmatic approach based on restricted-domain/restricted-speaking style technology (Figure 2). This approach does not address the problem of speaking style, but it relaxes vocabulary restrictions and provides speakable phrases on a hand-held device. Sometimes called a "one-way," it does not allow for free dialogs between two conversants (this requires spontaneous speech), but it permits speech entry of a list of useful phrases for a given situation. Figure 3 is a two-way device, the Pocket Translator, based on two-way speech translation technology described above. It runs on a standard PDA platform and permits spoken input for travel, medical, and military domains. A push-to-talk button on the device activates the system. The display shows recognition output, back-translation for verification, and translation output. A combination of using common pretranslated phrases by classifiers and look-up and performing actual translation has also

been proposed [26] and is used by several systems. Different user scenarios also do or do not prefer human– machine interaction. Textual displays and visual user feedback provide opportunities for interactive error correction and system maintenance but distract the user away from the human–human interaction. While the former may be preferable in tourist scenarios, the latter might be preferable in medical and military deployments.

DOMAIN-UNLIMITED SPEECH TRANSLATION

While numerous practical cross-lingual communication scenarios can be served by domain-limited speech translators, a large class of applications cannot be addressed by systems in this category: translations of broadcast news, parliamentary speeches, academic lectures, telephone conversations, and meetings all are open domain, as speakers may discuss any topic at any time. In 2003, work began in earnest toward removing this final limitation as well. The NSF-ITR project STR-DUST (Speech Translation for Domain-Unlimited Spontaneous Communication Tasks, National Science Foundation ITR Project, http://www.nsf.gov) in the United States (2003), the integrated project TC-STAR (Technology and Corpora for Speech to Speech Translation, Integrated Project of the European Union, http://www.tc-star.org) in Europe (2004) and the U.S.-DARPA GALE (Global Autonomous Language Exploitation, DARPA, http://www.darpa.mil/ipto/programs/gale/gale_ concept.asp) effort (2006) all aim to develop speech translators without domain limitation. Several different scenarios (broadcast news, parliamentary speeches, academic lectures) and different languages (Chinese, Arabic, Spanish) are being investigated in these projects. The lack of domain constraints has practically limited adoption of all approaches that require knowledge-based design or manually encoded linguistic representations. Instead, most emerging systems adopt data-driven learning approaches (statistical, example based) in their MT engines.

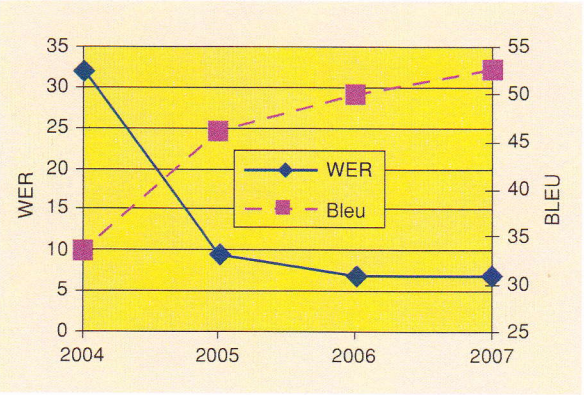
FOR SPONTANEOUS DIALOGS, WE MUST RELAX SYNTACTIC CONSTRAINTS AND ALLOW FOR VARIATIONS IN EXPRESSION.

READ/PREPARED SPEECH: PARLIAMENTARY SPEECHES AND BROADCAST NEWS

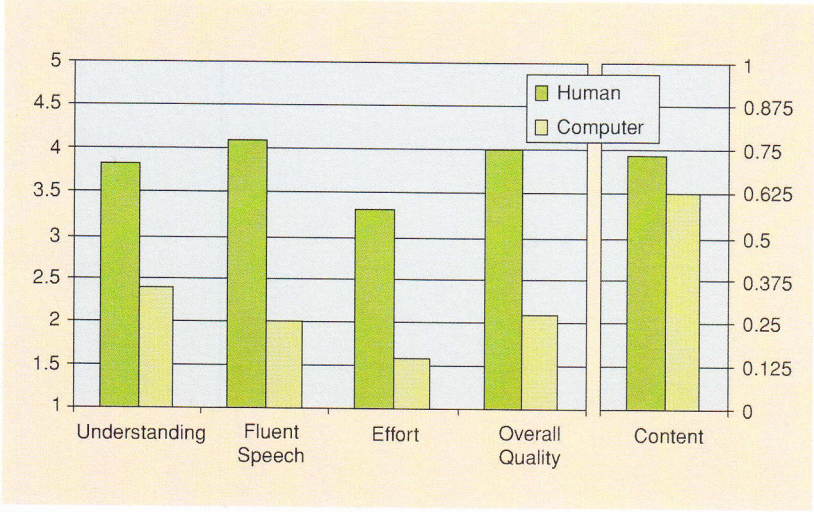
The translation of speech combines error-prone subcomponent engines, speech recognition (ASR), and MT. Errors in recognition may lead to errors in translation, and (unlike errors in recognition output) erroneous output from a combined speech translator generally has no phonetic or semantic similarity to the original input. Hence, the highest possible performance of each of the component technologies is of utmost importance for the usability of the overall resulting system. In the projects TC-STAR and GALE, extensive performance evaluations as well as manual usability tests are carried out. European parliamentary speeches (TC-STAR) and foreign broadcast news (GALE) were used as data material. Both are challenging tasks, but recording conditions are at least high quality and speaking style is relatively well articulated or read. Manual reference transcripts and translations are used in both projects to help evaluate and track performance. To evaluate performance, good metrics that can be evaluated automatically and repetitively are essential. While WER is an established method for ASR, MT is harder to evaluate as more than one translation can be correct. Yet automatic MT metrics (e.g., BLEU, NIST) have been proposed (by IBM, NIST, and others; see [22] for references) and found considerable following as they are inexpensive to run, objective, repeatable, and correlate well with human judgments. Human judgments [human translation error rate (HTER) [22]], however, are also periodically determined to assess the actual usability of MT systems.

The goal of GALE is to provide relevant information in English, where the input is derived from large amounts of speech in multiple languages (a particular focus is on broadcast news in Arabic and Chinese). To better measure the effectiveness of the technology, progress is measured by WER and BLEU, but also HTER, a measure of the human editing effort required to correct a machine-generated output. In the integrated project TC-STAR, speeches from the European parliament are automatically transcribed and translated between Spanish and English. Figure 4 shows the best recognition and translation quality results achieved in TC-STAR during the three years of project duration. It was also found that a WER of around 30% is influencing the machine translation quality significantly while a WER of 10% or better provides reasonable transcripts leading to generally understandable translations.

Figure 5 compares human and computer speech-to-speech translations on five different aspects by human judgment [17]: was the message understandable, fluent, listening effort, and overall quality; the scale ranges from 1 (very bad) to 5 (very good). The fifth result shows the accuracy [%] by which content questions could be answered by human subjects based on

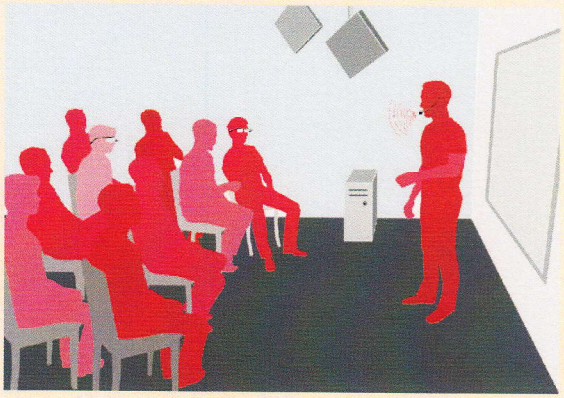


[FIG4] Improvements in speech translation and ASR over years on English European Parliament Plenary Sessions and translation into Spanish [33].

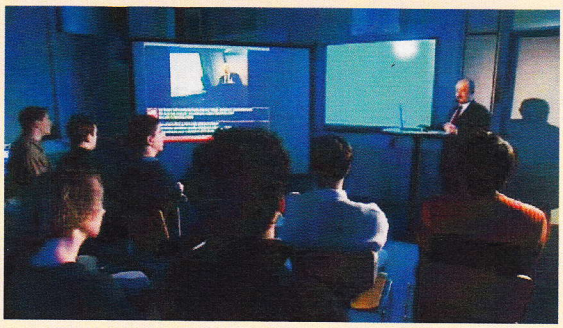


[FIG5] Human versus automatic translation quality (from [17]).

the output from human and machine translators. It can be seen that automatic translation quality lags behind human translation but reaches usable and understandable levels compared to human translations. It is also interesting that human translations also fall short of perfection. As it turns out, this is because human translators are unable to keep up with speaking rate and thus occasionally omit information [29]. This suggests an intriguing speculation: if machines are still limited by translation performance, and humans appear to be (and remain) cognitively limited, the day could come when machines may do a comparable or even better job at simultaneous translation than humans.



(a)



(b)

[FIG6] The Simultaneous Lecture Translation System at the Universität Karlsruhe [18]. The lecture room provides ceiling-mounted target audio speakers and translation goggles for its audience (left). Alternatively (right), simultaneous translation output can be displayed as text on a separate screen next to the presentation.

SPONTANEOUS DOMAIN-UNLIMITED SPEECH TRANSLATION: LECTURES

A further advance in cross-lingual communication tools may be given by a simultaneous translator that produces simultaneous real-time translation of spontaneous lectures and presentations (Figure 6). Compared to parliamentary speeches and broadcast news, lectures, seminars, and presentations of any kind present additional problems for domain-unlimited speech translation by:

- spontaneity of free speech, disfluencies, and ill-formed spontaneous natural discourse
- specialized vocabularies, topics, acronyms, named entities, and expressions in typical lectures and presentations (by definition specialized content)
- real-time and low-latency requirements, online adaptation to achieve simultaneous translation
- selection of translatable chunks or segments.

To address these problems in ASR and MT engines, changes to an offline system are introduced:

- to speed up recognition, acoustic and language models can be adapted to individual speakers (the size of the acoustic model is restricted and the search space is more rigorously pruned)
- to adapt to a particular speaking style and domain, the language model is tuned offline on slides and publications by the speaker, either by reweighting available text corpora or by retrieving pertinent material on the Internet or previous lectures by the same speakers.

As almost all MT systems are trained on sentence-aligned corpora and therefore ideally expect sentence-like segments as input, particular care has to be taken for suitable online segmentation. Deviations from sentence-based segmentation can lead to significant degradation. In view of minimizing overall system latency, however, shorter speech segments are preferred [18].

THE USABILITY OF TRANSLATION SERVICES

Speech translation is a technology that is to improve human-to-human communication. At best, it should be completely transparent and unnoticed and quietly help us bridge the language divide. It should provide accurate, reliable translation with minimal delay and with minimal distraction. To achieve this is a human-factors challenge raising numerous design choices and trade-offs.

First off, there is the question of which platform the system is to run on. A PC or PC-cluster is acceptable for large-scale offline translation runs. Applications that fit into this category may be translation of media content, such as broadcast news, movies, radio, etc. For dialog situations the choice of platform depends on whether the dialog situation arises in a stationary installation, such as a meeting room, a classroom, a briefing room, or in teleconferencing applications. Here individual PCs or laptops may be installed or accessed in a client-server mode. For dialogs in mobile situations, a smaller platform is desirable. PDAs or pocket devices are preferable. Current hardware limitations of such devices impose compromises on performance [14], [15], if all components (ASR, MT, TTS) are to run on the device. Alternatively, client–server architectures can be chosen under which the necessary computing is provided over the network at a remote server [16].

Aside from the choice of platform there are several human-factor issues that make a speech translator more or less a cumbersome assistant. One issue is the control of the device itself. Should it be hands-free/eyes-free allowing the user to focus on the dialog partner, or should it be controlled by the user, allowing the user to inspect the output to abort faulty translations or provide interactive correction, repair, or system customization?

Another issue is the choice and use of the microphone. While headset microphones are clearly the best from a performance point of view, they may generate too much of an imposition or distraction, particularly in dialog situations, where one would not be able to mount it on a dialog partner (for example, humanitarian, military, police missions). Here a number of alternatives have been proposed: handheld microphones, telephone handsets, or remote directional microphones. For speech translators in stationary, domain-unlimited environments, such as lecture or meeting translation, lapel microphones (if not headsets) or directional table-top microphones provide a compromise between performance and user convenience.

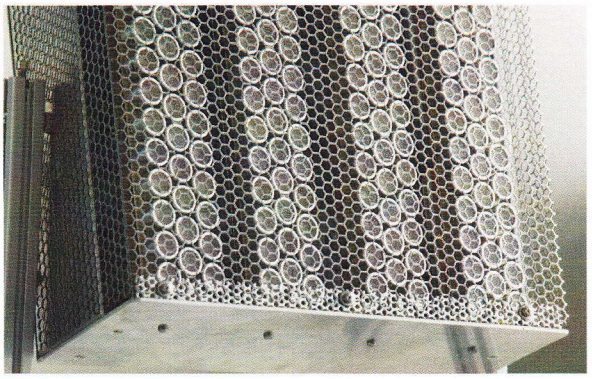
A third issue is how to present the output of speech translation services. Synthesized speech may be, in many situations, the preferred choice, but it can create delays in a dialog and it is by its very nature an audible, perhaps annoying, interference. For lecture translation, for example, a loud simultaneous translation in a lecture hall would not be acceptable, and several alternatives have been proposed:

- *Display screens:* Naturally, output can be delivered via traditional display technology, display on separate screens, or as subtitles, but all add distraction and inconvenience and it limits its output to one language.
- *Personalized headphones or PDA screens:* This allows for individual choice of output language (if several are provided) but adds inconvenience to wear/handle the device and (for headphones) masks the original speaker's voice.
- *Translation goggles:* Heads-up display goggles that display translations as captions in a pair of personalized goggles. Such a personalized visual output mode exploits the parallelism between acoustic and visual channels. This is particularly useful if listeners have partial knowledge of a speaker's language and wish to add complementary language assistance (Figure 7).
- *Targeted audio speakers:* A set of ultrasound speakers with high directional characteristics has been proposed [19] that provides a narrow audio beam to individual listeners in a small area of the audience where simultaneous translation is required. Since such speakers are only audible in a narrow area, they do not disturb other listeners, and several speakers can provide different languages to different listeners (Figure 8).

SCALING: DOMAINS, STYLE, LANGUAGES

In parallel to attention to performance, there is considerable concern (and there should be) about portability and scaling. Even if superior speech translation systems exist in one language pair and one application, how easy or difficult is it to transfer this

[FIG7] Translation goggles (from MicroOtical, http://www.microoptical.net).



[FIG8] Targeted audio [19].

ability to other domains, different speaking styles, and new languages? Modifications or adaptations still require considerable engineering effort, translating into cost that is only affordable for a very limited set of applications and languages. Currently, only a few domains (tourism, medical, military) have been seriously considered for domain-limited systems, and domain-unlimited systems exist in only a few language pairs, e.g., Spanish-, Chinese-, Arabic-English. Since all algorithms and components are data driven and apply to any language or domain, there is really no technical or linguistic reason limiting these choices. Yet data collection and development for each domain and language are still too costly to broaden the scope further. Speaking style also adds complications as conversational speech is harder to recognize and highly disfluent language is difficult to translate (indeed, it may require interpretation already in the source language!). The probable answer to all these challenges will likely come from further automation of the knowledge acquisition process. A particularly prominent example of this is the problem of language portability.

ERRORS IN RECOGNITION MAY LEAD TO ERRORS IN TRANSLATION, AND ERRONEOUS OUTPUT FROM A COMBINED SPEECH TRANSLATOR GENERALLY HAS NO PHONETIC OR SEMANTIC SIMILARITY TO THE ORIGINAL INPUT.

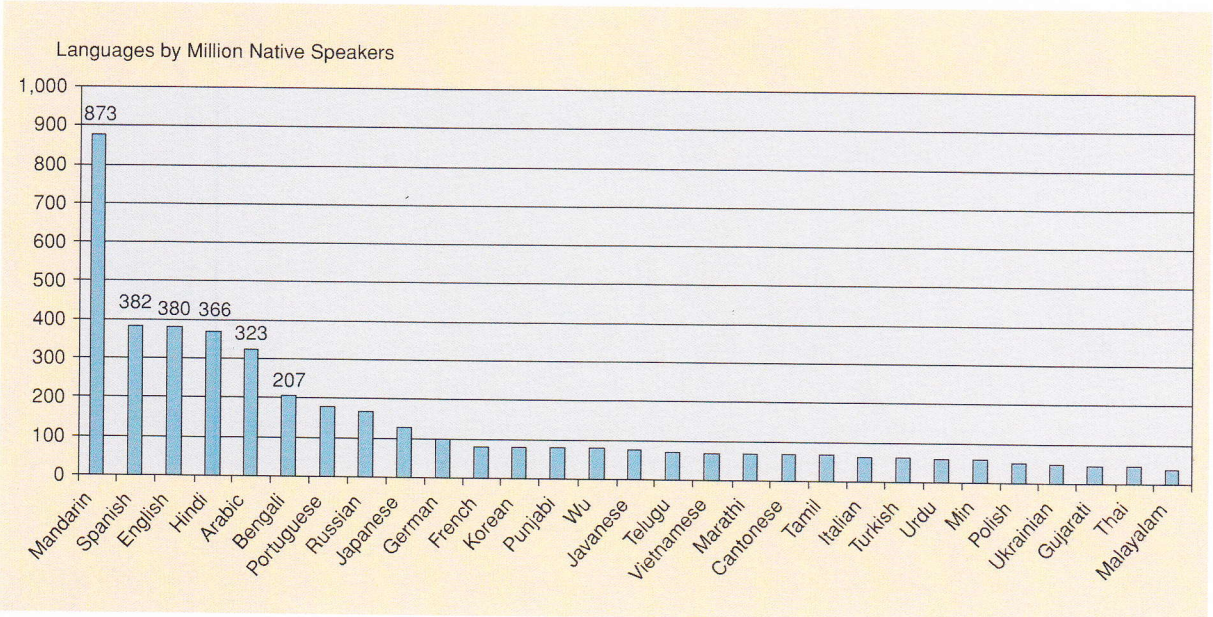
THE LONG TAIL OF LANGUAGE

By current estimates, there are more than 6,000 languages in the world, and only a few (perhaps less than ten) are currently seriously considered for speech translation development. Language needs are certainly given in other languages (perhaps even more pronounced), but their volume (market share) and their available data resources are considerably smaller than for the top four. As a result, all but a few languages [he long tail of language, see (Figure 9)] remains unaddressed. In response to this problem, language portability has emerged as a research concern in its own right, independent and orthogonal to the ongoing quest for better performance.

The greatest problem that remains is the acquisition of speech and language information, which requires a collection of large databases. While research exclusively aimed at performance uses ever more impressively large and massive data volumes [34], other research moves in an orthogonal direction, attempting to make do with less at lower cost. Techniques that have been proposed include:

- design of language-independent or adaptive system components (this was demonstrated for acoustic modeling [20] and could potentially be expanded)
- more selective use of available data and minimum-cost data collection
- interactive and implicit training by the user
- training from spoken simultaneous translation eliminating the need for text corpora [21]
- the use of pivot languages [7], [8]
- Web crawlers and the use of comparable corpora instead of parallel corpora [27], [28].

With further advances in portability, the cost of developing new languages is expected to come down, hopefully leading to a proliferation of cross-language communication tools.



[FIG9] The long tail of languages.

SUMMARY

In this article we have reviewed state-of-the-art speech translation systems. We have discussed issues of performance as well as deployment, and we reviewed the history and technical underpinnings of this growing and challenging research area. The field provides a plethora of fascinating research challenges for scientists as well as opportunities for true impact in the society of tomorrow.

AUTHORS

Alex Waibel (ahw@cs.cmu.edu) is a professor of computer science at Carnegie Mellon University and at the University of Karlsruhe. He directs the international Center for Advanced Communication Technologies (InterACT) at both locations with research interests in multimodal and multilingual human communication systems. His team pioneered many of the first domain-limited and domain-unlimited speech translators. Dr. Waibel was one of the founders and chairmen (1998–2000) of C-STAR, the consortium for speech translation research. He published extensively in the field, received several patents and awards, and launched several successful companies. He received his B.S., M.S., and Ph.D. degrees at MIT and CMU, respectively.

Christian Fügen (fuegen@ira.uka.de) received his diploma degree in computer science from the University of Karlsruhe in December 1999. Since then, he has been working as a research assistant at the Interactive Systems Labs with research interests in the field of automatic speech recognition and especially in the fields of acoustic and language modeling and adaptation in the context of a simultaneous speech-to-speech translation system, for which he is currently the maintainer at the University of Karlsruhe and Carnegie Mellon University. He was involved in several speech-to-text evaluations on conversational telephone, meeting, lecture and parliamentary speech, and he is currently the maintainer of the Janus Speech Recognition Toolkit.

REFERENCES

[1] N. Bertoldi, R. Zens, and M. Federico, "Speech translation by confusion network decoding," in *Proc. ICASSP*, Honolulu, Hawaii, vol. 4, pp. 1297–1300, Apr. 2007.

[2] S. Bangalore and G. Ricardi, "A finite-state approach to machine translation," in *Proc. North American ACL*, Pittsburgh, PA, May 2001, pp. 1–8.

[3] L. Levin, D. Gates, A. Lavie, and A. Waibel, "An interlingua based on domain actions for machine translation of task-oriented dialogues," in *Proc. ICSLP'98*, Sydney, Australia, Nov. 1998, pp. 1155–1158.

[4] F.J. Och and H. Ney, "The alignment template approach to statistical machine translation," *J. Computat. Ling.*, vol. 30, no. 4, pp. 417–449, 2004.

[5] A. Waibel, A.N. Jain, A.E. McNair, H. Saito, A.G. Hauptmann, and J. Tebelskis, "JANUS: A Speech-to-speech translation using connectionist and symbolic processing strategies," in *Proc. ICASSP*, Toronto, Canada, pp. 793–796, May 1991.

[6] Y. Gao, B. Zhou, Z. Diao, J. Sorensen, and M. Picheny, "MARS: A statistical semantic parsing and generation-based multilingual automatic translation system," *J. Mach. Translat.*, vol. 17, no. 3, pp. 185–212, 2002.

[7] A. Waibel, T. Schultz, S. Vogel, C. Fügen, M. Honal, M. Kolss, J. Reichert, and S. Stüker, "Towards language portability in statistical machine translation," in *Proc. ICASSP*, Montreal, Canada, vol. 3, pp. 765–768, May 2004.

[8] H. Wu and H. Wang, "Pivot language approach for phrase-based statistical machine translation," in *Proc. ACL 2007*, Prague, pp. 856–863, Jun. 2007.

[9] T. Morimoto, T. Takezawa, F. Yato, S. Sagayama, T. Tashiro, M. Nagata, and A. Kurematsu, "ATR's speech translation system: ASURA," in *Proc. Eurospeech'93*, Geneva, Italy, pp. 1291–1294, Sept. 1993.

[10] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto, "A Japanese-to-English speech translation system: ATR-MATRIX," in *Proc. ICSLP'98*, Sydney, Australia, Nov. 1998, pp. 2779–2782.

[11] Y.-Y. Wang and A. Waibel, "Modeling with structures in statistical machine translation," in *Proc. COLING-ACL 1998*, Quebec, Canada, Aug. 1998, pp. 1357–1363.

[12] D.B. Roe, F.C. Pereira, R.W. Sproat, and M.D. Riley, "Efficient grammar processing for a spoken language translation system," in *Proc. ICASSP*, San Francisco, CA, vol. 1, pp. 213–216, Mar. 1992.

[13] S. Vogel, F.J. Och, C. Tillmann, S. Nießen, H. Sawaf, and H. Ney, "Statistical methods for machine translation," in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed. Berlin: Springer, pp. 377–393, 2000.

[14] Y. Gao, L. Gu, B. Zhou, R. Sarikaya, M. Afify, H.-K. Kuo, W.-Z. Zhu, Y. Deng, C. Prosser, W. Zhang, and L. Besacier, "IBM MASTOR System: Multilingual automatic speech-to-speech translator," in *Proc. Workshop on Medical Speech Translation, HLT-NAACL-2006*, New York, pp. 57–60, June 2006.

[15] T. Schultz, A.W. Black, S. Vogel, and M. Woszczcyna, "Flexible Speech translation system," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 2, pp. 403–411, Mar. 2006.

[16] E. Sumita, T. Shimizu, and S. Nakamura, "NICT-ATR speech-to-speech translation system," in *Proc. 45th Annu. Meeting Association for Computational Linguistics (Companion Volume Proceedings of the Demo and Poster Sessions)*, Prague, pp. 25–28, June 2007.

[17] O. Hamon, D. Mostefa, and K. Choukri, "End-to-end evaluation of a speech-to-speech translation system in TC-STAR," in *Proc. MT-Summit*, Copenhagen, 2007, pp. 223–230.

[18] C. Fügen, M. Kolss, M. Paulik, and A. Waibel, "Open domain speech translation: from seminars and speeches to lectures," in *Proc. TC-STAR Workshop Speech-to-Speech Translation*, Barcelona, Spain, 2006, pp. 81–86.

[19] D. Olszewski, F. Prasetyo, and K. Linhard, "Steerable highly directional audio beam loudspeaker," in *Proc. Interspeech*, Lisboa, Portugal, Sept. 2006, pp. 137–140.

[20] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, no. 1–2, pp. 31–51, Aug. 2001.

[21] M. Paulik, S. Stüker, C. Fügen, T. Schultz, T. Schaaf, and A. Waibel, "Speech translation enhanced automatic speech recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, San Juan, Puerto Rico, pp. 121–126, Dec. 2005.

[22] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. Association for Machine Translation. in the Americas*, 2006, pp. 223–231.

[23] D. Chiang, "Hierarchical phrase-based translation," *J. Computat. Ling.*, vol. 33, no. 2, pp. 201–228, 2007.

[24] F.J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. 40th Annu. Meeting Association for Computational Linguistics*, Philadelphia, PA, pp. 295–302, July 2002.

[25] P. Koehn, "Pharaoh: A beam search decoder for phrase-based statistical machine translation models," in *Proc. 6th Conf. Association for Machine Translation in the Americas*, Washington, DC, 2004, pp. 115–124.

[26] S. Narayanan, S. Ananthakrishnan, R. Belvin, E. Ettaile, S. Ganjavi, P. Georgiou, C. Hein, S. Kadambe, K. Knight, D. Marcu, H. Neely, N. Srinivasamurthy, D. Traum, and D. Wang, "Transonics: A speech to speech system for English-Persian interactions," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, St. Thomas, U.S. Virgin Islands, pp. 670–675, 2003.

[27] D. Wu, and P. Fung, "Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora," in *Proc. 2nd Int. Joint Conf. Natural Language Proceedings*, Jeju Island, Korea, 2005, pp. 257–268.

[28] D. Munteanu and D. Marcu, "Improving machine translation performance by exploiting comparable corpora," *J. Comput. Ling.*, vol. 31, no. 4, pp. 477–504, Dec. 2005.

[29] R. Al-Khanji, S. El-Shiyab, and R. Hussein, "On the use of compensatory strategies in simultaneous interpretation," *Meta : Journal des Traducteurs*, vol. 45, no. 3, pp. 544–557, 2000.

[30] A.W. Black, "Multilingual speech synthesis," in *Multilingual Speech Processing*, T. Schultz, and K. Kirchhoff, Eds., New York: Academic, 2006.

[31] V. Steinbiss (2006, Apr.), "Human language technologies for Europe" [Online]. Available: http://www.tc-star.org/pubblicazioni/D17_HLT_ENG.

[32] A. Pollack (1993, Jan.), "Computer translator phones try to compensate for babel," The NY Times [Online]. Available: http://query.nytimes.com/gst/fullpage.html?res=9F0CE7D7113AF93AA15752C0A965958260.

[33] H. Ney, "TC-Star: Statistical MT of text and speech," presented at TC-Star Review Workshop, Luxembourg, May 2007.

[34] F.J. Och, "Statistical machine translation: Foundations and recent advances," presented at MT-Summit 2005, Phuket, Thailand. SP