14 Multimodal Dialogue Processing for Machine Translation

Alexander Waibel

14.1 Introduction

Humans converse with each other to communicate and to develop ideas interactively in the presence of imprecise and under-specified information. In an increasingly multicultural world, such communication of ideas necessitates communication across language boundaries. With more than 7,000 languages spoken on our planet, however, such boundaries cannot be overcome by language learning or human translation effort alone and require technical solutions that can help mediate between humans and machines. To be effective, such mediation cannot be accomplished by text translation alone, as human communication expresses itself in several modalities. Speech, discourse, dialogue, handwritten or texted text, road signs, gestures, eye gaze, and facial expressions all participate in human communication and complement text as an expression of thoughts and ideas, so that our messages must be transmitted multimodally across language barriers as well.

Among those modalities, speech may perhaps be the most important (next to text), because we express ourselves in multiple languages and that requires us to translate language in its spoken as well as textual form. Technologies that aim to take on such cross-lingual interpretation duties of speech are commonly known as speech-to-speech translators. In the following, we will begin with a discussion of speech-translators and their underlying technology. We will then show how their design and realization must be closely matched to their intended use case and how they must be field-able and adaptive to respond to the needs of their deployment.

Glossary

Automatic speech recognition: the signal spoken in language is recorded by microphone, processed, and converted to text (speech to text).

Code switching: mixing words from different languages, declination rules and compounding.

Consecutive interpretation typically interprets a few sentences, one at a time, before giving the dialogue partner a chance to respond.

Cross-lingual subtitling: a mixture of consecutive and simultaneous interpretation where interpretation is performed on media content and delivered textually as subtitles.

Earplugs and pixel-buds: a set of ear-plugs provides input and output for a speaker attempting to dialogue with others.

Electromyography: electrodiagnostic medicine technique for recording the electrical activity produced by muscles.

JANUS system was the first speech translation system presented to the public in the USA and Europe in 1991.

Linguistic scalability/portability. Implement the technologies developed not only in one or two languages, but extend it to cover communication among all languages and cultures on our planet.

Neural machine translation: greater abstraction and greater ease of integration is obtainable through neural translation approaches, where internal ("hidden") abstractions are generated as a side-effect of training many layers of neural structures.

Out-of-vocabulary words (OOVs): when words are missing in the pronunciation dictionary of a recognizer, leading to one or more substitution errors. Named entities and specialty terms are particularly prone to this type of problem.

Simultaneous interpretation attempts to recognize and translate spoken language in parallel to the input speech without making the speaker pause.

Speech synthesis: text in the target language is output in spoken language (text to speech).

Speech translation goggles translate output from a simultaneous (lecture) translation system delivered textually via heads-up display goggles.

Statistical machine translation: greater speed of learning and better performance and generalization to broader topics, but still requires collections of large parallel corpora.

Targeted audio: synthetic speech output in a speech translation system delivered selectively by directional loudspeakers.

Text-to-speech synthesis (TTS): TTS makes translated sentences audible in the target language and thus permits full speech-to-speech dialogues between two participants.

We will then also consider additional modalities and flexibilities between them in view of developing such seamless and language-transparent communication tools.

In its most direct form, a speech-to-speech translator could be constructed by combining a speech recognition engine (speech-to-text (STT)) with a machine translation (MT) engine, so as to translate a spoken sentence from language A to language B. If a response from a speaker in language B is to be translated into language A, we will also need recognition and translation engines in the reverse direction. Decomposing the problem in this fashion, however, vastly oversimplifies the problem of cross-lingual communication.

If we recall that the goal of cross-lingual communication and dialogue is to effectively *communicate* ideas, several orthogonal dimensions emerge that we must carefully consider to achieve a thoughtful and effective design.

1. **Spoken Language**. The first set of such issues pertains to the problems associated with translation of *spoken language*.
   - *Errors*. Speech recognizers make errors and translation engines must be robust against such errors or attempt to correct for them.
   - *Spoken language*. Speech is disfluent and hardly corresponds to syntactically well-formed text. Machine translation must therefore be adapted and trained for spoken language instead of text.
   - *Punctuation, casing, and disfluencies*. Human speech misses punctuation markers and casing, which otherwise provide important clues for translation. Instead, speech contains an array of potentially confusing disfluencies (hesitations (aeh, hum, uhm, er, etc.), false-starts, and fragments).
   - *Prosody*. Speech (unlike text) encodes additional information by way of pitch, intensity and rhythm, which transmit meta-level signals, such as emotion, gender of the speaker, emphasis, discourse information, social cues, degree of formality, etc.
2. **Interaction Style**. The second dimension pertains to the type and style of translation that depends on the situation and use case.
   - *Consecutive interpretation* typically interprets a few sentences, one at a time, before giving the dialogue partner a chance to respond, again with a short utterance of one or a small number of sentences. Processing can be more accurate and communication more effective in a face-to-face dialogue situation, as both participants are always aware of the mediation provided by translation and are thus generally more

cooperative. Also, interactive error handling can be employed. Consecutive interpretation, however, introduces a delay that slows down communication. Typical use cases are given by pocket translators, or bi-directional dialogue translators.

- Consecutive interpretation *in combination with dialogue processing* aims to emulate the ability of human interpreters and to carry out monolingual dialogues in addition to interpreting between the languages.[1] In this way, certain transactions can be handled in a more compressed manner monolingually and some are communicated via interpretation [Oviatt and Cohen 1992]. A system design involves maintaining two linked dialogues with an interpreter, one in each language. The interpreter is a dialogue participant, who translates some of what is said, but might also answer questions directly (i.e., without translation), since they may already have been told the answer. In this fashion, repeated requests or clarifications can be handled by monolingual dialogue, and do not require the full round-trip to the other language.
- *Simultaneous interpretation* attempts to recognize and translate spoken language in parallel to the input speech without making the speaker pause. This mode of interpretation can be faster, and generates less interference to the speaker. It is more challenging, however, as speakers tend to be less aware of the interpretation efforts, and cannot participate in resolving errors. It must also trade-off context (and thus accuracy) against the latency between the spoken and translated words. Typical use cases are the interpretation of lectures or speeches.
- *Cross-lingual subtitling* is a mixture of the above where interpretation is performed on media content and delivered textually as subtitles. The input speech is typically less disfluent (prepared speech) than lectures or speeches; latency may be less of a concern and in some instances error processing may be possible.

3. **The third dimension is concerned with the form of delivery**. As we aim for *effective communication*, we cannot limit ourselves to recognizing and translating spoken sentences. If the goal is to get one's point across with minimal interference and minimal delay, we must also be concerned with a

1. Such as, for example, human interpreters on AT&T Language Line (see [Oviatt and Cohen 1992])

most effective human interface design and multimodal strategies. Thus, we must also consider the following.

- *Multimodal input and output*. To optimize efficiency of *communication*, it is often more effective to switch or combine multiple modalities, such as speaking, texting, typing, images, handwriting, gesturing, pointing. Output can also be produced alternatively by synthetic speech or as text depending on situation and delivered on smartphones, tablets, in heads-up display goggles or by targeted audio speakers.
- *Error handling and multimodal error repair*. Speech recognition and machine translation will always produce errors and so it is essential for effective communication to detect and correct errors in the most effective manner. Errors can, for example, be flagged visually on a screen or articulated verbally and corrected by dialogue or multimodal repair.
- *Field-adaptable and extendable systems*. Languages and vocabularies change, and interpreting dialogue systems must evolve alongside such changing languages and vocabularies and adapt to any given dialogue scenario. The situations are too numerous to predefine vocabularies and language use once and for all a priori. Effective systems must provide mechanisms that allow (non-expert) users to perform such adaptations in the field and during use.

In this chapter, we begin with an introduction to the technology of interpreting systems. We then review use-cases and deployed systems in use today. Finally, we discuss the science and art of multimodal interface design that make such systems effective in the field.

14.2 Technology

The components technologies of speech-to-speech translators and their performance are subject of much research in computer speech and machine translation communities. While different use cases (as discussed in the previous section) require different configurations (see Section 14.4), let us first consider a typical two-way speech-to-speech interpretation system (see Figure 14.1). For a human being, speaking in one language to understand another human being speaking in another language (depending on use-case, up to), three partial tasks have to be solved (possibly in two or more language directions).

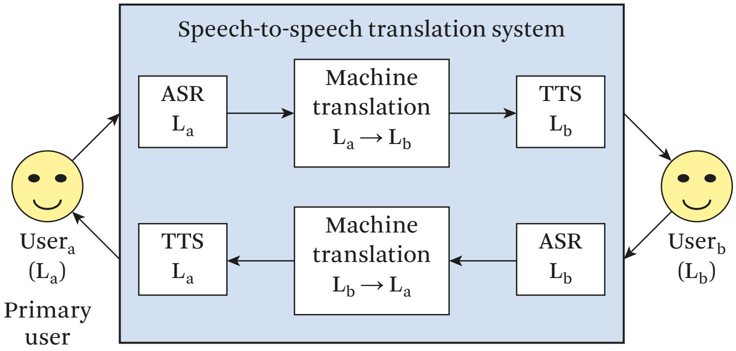

Figure 14.1 Translation of spoken language (speech to speech translation)—overview.

1. **Automatic speech recognition**. Here, the signal spoken in language (L_a) is recorded by microphone, processed, and converted to text (speech to text);
2. **Machine translation** Here, text in language L_a is translated into text in the other language (L_b) (text to text).
3. **Speech synthesis** (L_b). Here, text in the target language Lb is output in spoken language (text to speech). For a dialogue between persons speaking two languages, this process also has to be possible in the other direction (from L_b to $_a$) and, hence, requires analogous subsystems in the other language. A final integration of these subsystems with a comfortable user interface then has to be operable easily in real communication situations.

Each of these partial tasks represents an area of research, which over the years has been harder to solve than might be expected by the casual observer due to the complexity and ambiguity of human language. For this reason, they have been studied by scientists for several decades and are still challenging in spite of the considerable progress achieved. The most important lessons learned are that (1) because of inherent ambiguity and errors, we can never make hard decisions, but only soft probabilistic ones for every source of knowledge in human language, and(2) because of their complexity, we cannot encode these statements and their interactions manually, but must learn them from data.

14.2.1 Automatic Speech Recognition (ASR)

For the unaware observer, the problem of speech recognition may not appear very difficult at first, as we human beings manage it well and easily. However, several ambiguities occur in spoken language already: The English acoustic sequence of

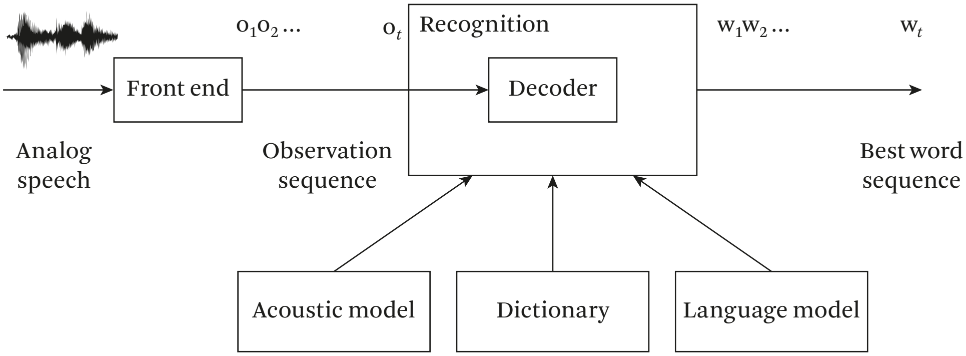

Figure 14.2 A typical speech decoder (speech to text).

sounds “yu-thu-nā-zhu” may mean both “Euthanasia” and “youth in Asia.” Sentences like “This machine can recognize speech” are pronounced in the same way as “This machine can wreck a nice beach.” Speech recognition requires an interpretation as to which of several similar alternatives is the more meaningful or more probable one in a given context. In modern speech recognition systems, this is achieved by a combination of acoustic models that assign a probability to every sound, a pronouncing dictionary (that assigns a pronunciation to every word), and a language model that evaluates the probability of every possible word sequence “w1, w2, . . . ” of the sentence. Figure 14.2 shows such a typical decoder. Evaluation of these models during recognition and settings of the best parameters of these models, however, cannot be determined manually, but require automatic search and optimization algorithms.

Parameters of acoustic and linguistic models are determined with the help of machine learning algorithms using huge databases of speech samples, whose transcriptions are known.

Algorithms work with statistical optimization methods or neural networks and learn the best match between signals and symbols (context-dependent phonemes and words) based on known exemplary data. Today’s systems use neural networks in each of these models with several millions of neural links optimized by the learning algorithm.

14.2.2 Machine Translation (MT)

First attempts to translate texts by machines (MT = machine translation) were made as early as during the second World War, but early systems attempted to encode all requisite knowledge by rules and failed due to the ambiguity of language and

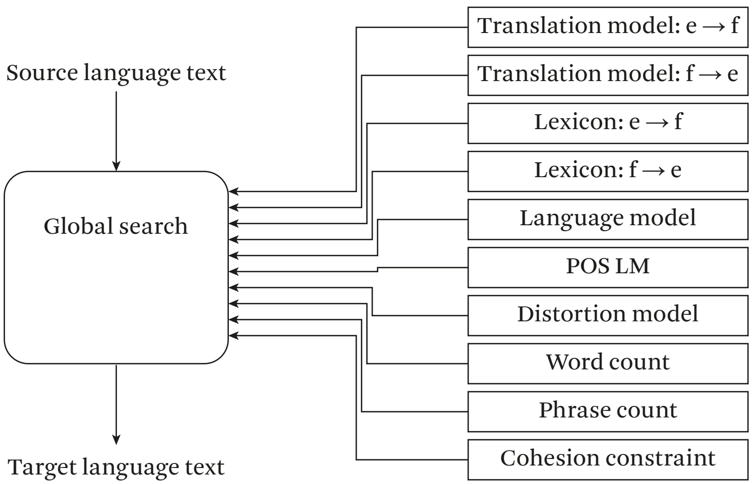

Figure 14.3 Statistical machine translation (text to text).

the complexity of required associated context knowledge. Nearly every word (skate, row, mouth) has several meanings and, hence, translations can only be interpreted correctly in context. MT folklore recounts that the sentence from the bible "The spirit is willing but the flesh is weak" was supposedly translated into Russian by an early machine translator as "The vodka is good but the flesh is rotten."[2] Also, language structure is frequently ambiguous. For instance, what does the pronoun "it" refer to in "If the baby doesn't like the milk, boil it"? Most likely the author meant boiling the milk (not the baby!) and hence the pronoun should be translated into German as "sie" and not "es."

The attempt to manually encode all required syntactic, semantic, and lexical knowledge with the help of rules would generally not scale (beyond well-defined contained domains). With the arrival of faster and more powerful computing platforms and larger data-resources on the internet, rule-based approaches eventually gave way to automatic learning systems. Modern MT system now use system architectures that optimally trained statistical knowledge sources (see Figure 14.3), or arrangements of recurrent neural network encoder/decoder structures [Kalchbrenner and Blunsom 2013, Sutskever et al. 2014, Bahdanau et al. 2015].

2. The example is due to an early article on MT from the *New Yorker* but it is uncertain if this confusion ever actually occurred in an actual machine translation system.

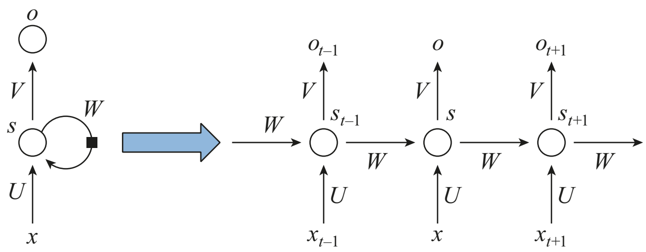

Figure 14.4 Recurrent neural nets, unfolded in time.

Statistical Machine Translation offers greater speed of learning and better performance and generalization to broader topics, but still requires collections of large parallel corpora. However, they still have to be trained one language-pair at a time and cannot easily abstract across languages or include more varied information sources (e.g., prosody, meta-information, etc.) without ever more complicated combinations of individual models.

Neural Machine Translation. Greater abstraction and greater ease of integration is obtainable through neural translation approaches, where internal ("hidden") abstractions are generated as a side-effect of training many layers of neural structures. Generally, they are today implemented as recurrent networks that encode sentences by presenting words (or some compact representation of them) in sequence, and then decoding them in sequence in the other language. A recurrent neural network (RNN) and it sequential unfolding is shown in Figure 14.4. As before outputs O are generated from inputs X, but also influenced by the state of the net. In Figure 14.4 we see such a recurrent net unfolded in time. Here a neural net produces a sequence of outputs. The output at timestep t (o_t) is based on the input x_t but also from the state of the net at the previous timestep s_{t-1}. With words represented as vectors as input x_t, a recurrent neural network can remember sequential information in this recurrent state. Once an entire sentence is *encoded* in this manner, the remaining context vector can then be used to *decode* a sentence in another language, as shown in Figure 14.5. Recurrent encoder-decoder models as shown in Figure 14.5 were attempted for neural dialogue modeling and machine translation as early as the late 1980s [Miikkulainen and Dyer 1989, Jain et al. 1989] and early 1990s [Wang and Waibel 1991], and more recently [Cho et al. 2014a, Sutskever et al. 2014].

Early RNN-based encoder-decoder models had limited success for MT, however, because the recurrences in an RNN tend to remember only recent information

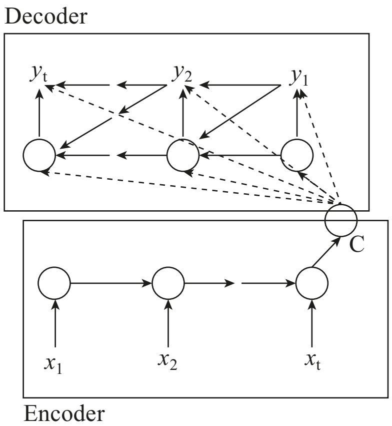

Figure 14.5 Recurrent encoder-decoder for neural MT.

(words) and forget the earlier context. This is a problem, particularly, when translation requires long-distance reordering between words. Several modifications of the models were proposed to prevent such forgetting in RNNs. The addition of an "attention" mechanism, finally, was shown to be effective to overcome this limitation. It permits different words from the input word sequence to be weighted appropriately ("attention") for each word in the output sequence (see Figure 14.5b) [Bahdanau et al. 2015]. The attention mechanism was found to yield considerable improvements in MT performance, particularly for language pairs that involve long-distance reorderings (e.g., German). Most recently, it was found that this attention mechanism is indeed sufficient for high performance even without a recurrent state model [Vaswani et al. 2017].

Using attention to represent long-distance relationships and context in language, neural machine translation (NMT) networks now achieve such dramatic improvements over statistical methods that they have all but replaced statistical machine translation (SMT) as the method of choice in MT. Moreover, in addition to superior performance, the practical advantage of NMT is that it is possible to train networks over multiple languages at the same time, so that a single network architecture can serve multiple languages and language *pairs* at the same time, by way of learning some kind of internal semantic representation for all! [Ha et al. 2016, Johnson et al. 2017]. The practical impact (in addition to the better performance) cannot be overstated. At 7,000+ languages in the world, simplifying extensions to new languages and language pairs is key to scaling machine translation, globally. Moreover, as neural abstractions are learned just from data, the input/output to such networks also does not need to be words alone. They can be acoustic features, meta-level information, or even images. Indeed, cross-modal translation from video to text and vice versa is now a subject of intense research. Automatic descriptions of visual scenes or video generation from text are likely applications.

14.2.3 Speech Synthesis (Text to Speech (TTS))

If a speech-translator is to output speech in another language, the third component is created by text-to-speech synthesis (TTS). TTS makes translated sentences audible in the target language and thus permits full speech-to-speech dialogues between two participants. In comparison to automatic speech recognition (ASR) and MT, TTS synthesis is generally considered to be a simpler problem, as only one signal has to be produced from a textual sentence to be understandable and it is not necessary to handle the great breadth of ambiguities of the other components. Still, open and important issues exist; however, that continues to be a subject of research. These include improving the language portability of TTS subcomponents through machine learning, to reduce the cost and effort to build TTS systems for more languages. Voice conversion (to adapt the output voice to an input speaker) is also a topic of interest. And critically, better prosodic control of output is needed, so that more suitable emotional emphasis, tone of voice, dialogue context, social setting, level of formality, gender, social roles of speaker and addressee, and other such factors that affect a conversation can be better situated and synthesis thus delivered.

14.2.4 Machine Learning, Statistics, and Neural Networks

All three components of a speech translation system are now powered by systems that are built by exploiting machine learning, both to deal with ambiguity as well as to learn automatically from data as opposed to a developer's writing rules following introspection. Crucial to their success: the dramatic growth in available data resources (mostly over and through the internet) and available computing resources. These resources have led to a rethinking and replacement of the dominant learning paradigm from statistical modeling back to neural network models, i.e., the models that had already been explored in the 1980s. Neural models that are almost identical to those developed during the late 1980s [Waibel et al. 1989, Waibel 1989, Bourlard and Wellekens 1989], now show their advantage fully as they are trained over several orders of magnitude larger databases and they now deliver up to 30% relative performance gains in speech recognition and MT performance. At the time of this writing, neural "deep learning" models are rapidly replacing statistical models as the dominant approach for speech recognition, MT, and speech synthesis

[Zenkel et al. 2017, Zweig et al. 2016, Miao et al. 2015, Sennrich et al. 2016, Cho et al. 2016, Neubig 2016]. Systems that include multimodal signals, generalize across many languages, and systems that could train directly end-to-end, from speech to speech, may be possible and become reality soon.

14.3 Evolution of System Prototypes and Deployments

The development of automatic spoken language translation systems started in the early 1990s when ASR, MT, and TTS systems first reached a minimum degree of maturity required to attempt a first integration [Waibel et al. 1991, Morimoto et al. 1993, Wahlster 1993]. In the course of the following two decades, major limitations in technology were overcome in a number of research and development phases. Today, speech translators have entered commercial and public usage and can be used by everyone. In the following, we review the different technological milestones, phases of maturity, and the use cases and key deployments the enabled (see Table 14.1 for an overview of system qualifications).

14.3.1 First Demonstrations

The JANUS system was the first speech translation system presented to the public in the U.S. and Europe in 1991 (see Figure 14.6). JANUS was developed for German, Japanese, and English by Universität Karlsruhe in Germany and Carnegie Mellon University in Pittsburgh, PA, USA. It was a result of cooperation with the ATR Interpreting Telephony Laboratories in Japan, which developed similar systems for the Japanese language in parallel. The systems together were presented in the first translating video conferences [Waibel et al. 1991, Handelsblatt 1991, Morimoto et al. 1993].

These systems represented first steps, managed an initially small vocabulary (< 1000 words), required a relatively restricted syntax, and covered a limited domain (e.g., registration for a conference). They were too large and slow to really be of assistance in field situations, e.g., to a traveler. Similar demonstration systems were presented by other research groups—AT&T [Roe 1992] and NEC [Hatazaki et al. 1992].

14.3.2 Research Systems and Prototypes

For these systems to be used in practice, other important phases of development followed to successively master difficult problems:

Spontaneous Speech, Domain-limited Research Systems. To implement practical systems, the assumption of syntactic correctness has to be eased or eliminated. People

Table 14.1 Development phases of speech translation systems.

	Years	Vocabulary	Speaking Style	Domain	Speed	Platform	Example Systems
First Dialogue Demonstration Systems	1989–1993	Restricted	Constrained	Limited	2–10× RT	Workstation	JANUS-1 (ATR, CMU), C-STAR-I, NEC, ATT
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	JANUS-III, C-STAR, Verbmobil, Nespole, Babylon, Transtac
Translation of Broadcast News, Political Speeches	2003–Present	Unrestricted	Read/Prepared Speech	Open	Offline	PC's, PC-Clusters	NSF-STRDUST, EC TC-STAR, DARPA GALE,
Simultaneous Translation of Lectures	2005–Present	Unrestricted	Spontaneous	Open	Realtime	PC, Laptop	KIT/CMU-Lecture Translator
Commercial Consecutive Translators on a Mobile Phone	2009–Present	Unrestricted	Spontaneous	Open	Online and Offline	Smartphone	Jibbigo, Google, Microsoft,
Simultaneous Interpretation *Services*	2012–Present	Unrestricted	Spontaneous	Open	Realtime, Online	Server, Cloud-Based	KIT-Lecture Translator, EU-BRIDGE, Microsoft
Consecutive Interpreting Telephony *Services*	2015–present	Unrestricted	Spontaneous	Open	Realtime	Server, Cloud-Based	Microsoft Skype

Figure 14.6 First speech translation prototypes in video conferences (1991).

rarely speak syntactically correct and complete sentences. They rather speak fragmentary segments with stammering, repetitions, filler words, and hesitations (er, hum, aeh, etc.). These fragments first have to be identified correctly and then filtered out or corrected by processing before translation takes place. First, spontaneous speech translation systems were developed from 1993–2000 [Morimoto et al. 1993, Takezawaet al. 1998]. These systems were still slow and required extensive hardware. Their domain continued to be too limited to extract the fragments relevant to translation by modeling the semantics. JANUS-III, C-STAR Systems, VERBMOBIL, and other projects made considerable progress, but still remained unusable in practice [Lavie et al. 1997]. Domain limitation and vocabulary restrictions had to be overcome first and systems had to be accelerated and readied for mobile use. In due course, manually programmed approaches (possible in limited domains) were replaced by automatically learned, statistic subsystems that scaled better to larger domains, and improved robustness and accuracy [Brown et al. 1993, Och and Ney 2004, Wang and Waibel 1997, Koehn et al. 2007]. Smartphones and cloud computing offered platforms that could perform these tasks in real-time and were accessible by a broad audience of users.

Two types of applications, serving different use cases began to emerge.

- The first is given by mobile devices that provide consecutive interpretation in human-human interactive dialogues. Here, speakers converse through an interpretation system that translates sentences consecutively. A speaker says one or more sentences in one language, followed by the system's translation. Then the other party responds in another language followed by translation back to the first speaker's language. Consecutive translation slows the flow of a conversation (because speakers have to wait for a translation to complete), but they make interpretation controllable and observable, and (in case of errors) speakers can intervene to make themselves understood. For most applications (tourism, medical uses, humanitarian aid, etc.) a vocabulary of about 40,000 words is sufficient to cover most conversational needs. But systems have to run on small mobile devices, which requires either fast cloud-based operation via telephone networks or compact-efficient implementations on the device.
- The second is given by simultaneous interpretation for stationary use: in many deployments of speech translation, a dialogue between two conversation partners is actually not needed but rather a fast interpretation of a stream of speech (or monologue) is desired. For example, TV broadcasts, internet videos, lectures, speeches, and addresses all require no response. In most of these deployments mobility is less of a concern, as the actual processing can be performed in the cloud on powerful servers. Simultaneous interpretation, however, is complicated by a broader range of vocabularies and special terms, and by the absence of obvious sentence markers. The system itself has to determine the beginnings and end of translatable units or sentences, and—in the case of simultaneous interpretation—must deliver translation output with little delay, before a speaker is finished speaking. Segmentation into units or translatable fragments have to be performed automatically and punctuation (full stops, commas, question marks) inserted automatically based on partial context (The Economist 2006). Statistical and neural models perform these predictions, and display interfaces must manage updates when further context requires revision.

14.3.3 Translation of Deployments and Services

Early research systems (1990–2005) solved technical problems and paved the way for the sales and real use of speech translation systems in society.

14.3.3.1 Mobile Consecutive Interpretation Systems

Interpretation systems were first tested in the field during humanitarian and logistic exercises of the U.S. government. Although network-based solutions were proposed, fieldable speech-translators usually required off-line operation, as network access could not be assured (or might be prohibitively expensive) in most humanitarian, logistic, and—indeed—tourist/travel deployments. The resulting systems resorted to laptops, and later PDAs and smartphones with all their speech translation software running on device. Computation was kept within manageable bounds, initially, by limiting the domain of speech-to-speech translation systems to transactional tasks of limited scope (e.g., hotel reservations, scheduling, healthcare interviews) or, alternatively, by the use of simple phrase books that would be accessed by voice. Either solution required only smaller vocabularies and could anticipate a more limited language use and thus restrict computation and memory requirements. [Eck et al. 2010, Stüker et al. 2006, Voxtec none, Ectaco 1989]. Early models of such systems offered commercially by VOXTEC and MOBILE TECHNOLOGIES are shown on the left of Figure 14.5. Due to the limitations in vocabulary and hardware, and -in the case of phrase-books- due to the inflexibility of expression, such early systems could achieve adoption only in special situations, were restricted phraseology and limited tasks are acceptable. For the wider use of speech translators by the wider public during travel and communication, further advances were necessary.

With the emergence of smartphones, both the general availability of a suitable platform as well as the necessary computational performance reached the critical capacities that made speech recognition and translation of open unlimited (> 40,000 words) vocabularies embedded on a device in near real time possible.

In 2009, Mobile Technologies (a startup of Carnegie Mellon researchers) launched, Jibbigo, the first domain-unlimited speech-to-speech translation system fully embedded on a phone in 2009 [Eck et al. 2010]. The system found quick adoption and distribution through the simple sales mechanisms of the Apple iTunes app stores and with the growing use of smartphones worldwide, Jibbigo quickly expanded to 15 languages and reached worldwide distribution. Other similar products followed suit, such as systems by Google and Microsoft. While Jibbigo offered a downloadable off-line solution (for a fee—a network-based solution was also available for free), many other entries were and still are exclusively network based. Although network-based solutions can access more powerful computational resources and connect with related internet resources, off-line systems require no roaming fees nor existing infrastructure and are thus preferable in many

(a) (b) (c)

Figure 14.7 First commercial systems: (A) Phraselator, (B) iPaq PDA-based speech translator (2005), (C) Jibbigo, the world's first speech-to-speech translator on a phone (2009). (Phraselator™ by VOXTEC LLC and Speech Translator™ by Mobile Technologies LLC)

humanitarian and travel situations. Jibbigo has thus been used in a number of humanitarian missions and government deployments, where an existing network infrastructure cannot be relied upon (Figure 14.8 A-D, show healthcare initiatives in Thailand, Cambodia, and Honduras for translation between English-speaking physicians and patients speaking other languages).

Typical system configurations may run on iPhones, Android smartphones, or on tablet computers. Tablets were found to be particularly well-suited for face-to-face interaction between partners sitting opposite to each other in medical missions. After five years of development in field situations, the systems were evaluated to perform well in humanitarian missions (MEDCAP—Medical Civil Action Program, Thailand, in 2013 [Hourin et al. 2013]). It was found that 95% of the interactions during the registration of patients, the conversation could be managed with the sole aid of the automatic tablet interpreter (Jibbigo).

Google and Microsoft followed suit (2013) with translation capability of their own that could be downloaded for off-line use, while broadening the number of languages on offer, making smartphone translators a common tool for today's travelers.

Figure 14.8 Medical operations in Thailand, Cambodia, and Honduras: (A) translingual dialogues between American physicians and patients in Thailand; (B) medical care with help of the JIBBIGO-speech to speech translator in Thailand; (C) medical operations in Cambodia; and (D) humanitarian operations with Jibbigo in Honduras.

14.3.3.2 Consecutive Interpreting Telephony

Mobile speech translators on smartphones or tablets offer effective and flexible consecutive translation in face-to-face field situations. Of course, consecutive translation can also be used for remote communication over telecommunication networks. Indeed, the earliest prototypes and research projects had envisioned translated video chat services as their use case. For example, the ATR-Interpreting Telephony Laboratories in Osaka, Japan, were already established in 1986 to investigate this possibility, and subsequent public demos together with partners in the U.S. (CMU) and Germany (Siemens, Karlsruhe) demonstrated such video chat sessions as early as 1991. Commercialization of consecutive translation services (human and automatic) followed in the decade since. In the U.S., AT&T established a human interpreting service (AT&T Language Line)[3] to fill consecutive interpreting needs over telephone lines, followed by software-driven services. Consecutive translation

3. http://www.languageline.com.

(human or mechanical) as a fee-based service has only been moderately successful, however, and so it was frequently packaged as a feature for video chat service providers, where translation provides broader network reach and contributes to consistent service expansion (a language on/off-ramp of sorts) for operators of communication services. In this manner, an early commercial video chat room enhanced by speech translation was introduced in 2010 by Hewlett-Packard's in its MyRoom Video Chat product and other communication services followed suit.

The broadest and largest telephone and video communication provider today is Skype (by Microsoft), a free voice-over-IP telephony service. In its continuing drive to expand its network, Skype now offers language interpretation through "Skype Translator" (see Figure 14.9), an automatic speech-to-speech interpretation service for human dialogues. Due to its massive and growing user base, Skype Translator represents one of the largest deployments of speech translation. The system accepts speech from speakers in two languages, interprets their messages, and outputs results as speech or text in the other language. A "TrueText" facility cleans up the disfluencies of spontaneous speech and turns them into more readable text. Synthetic output after translation is also overlaid on top of the original speech (at reduced volume) much like voice-overs in TV reporting. The approach helps reduce delays in the consecutive translation of dialogues (Skype calls this approach "ducking"). Skype (as well as other) research teams also experimented with robotic mediators, but Skype found this approach—or at least its implementation—somewhat awkward. Given the text messaging features of Skype, cross-lingual communication is also improved multimodally by allowing the participants to resort to complementary communication modalities, including speech, text, and video. Given the large number of users, a Skype translator[4] can then learn and improve from continued use [Lewis 2015].

14.3.3.3 Simultaneous Interpretation

In a multi-lingual environment, dialogue between conversation partners speaking different languages is not the only challenge. When thinking of TV news, films, presentations, lectures, speeches, road signs, transparencies for lectures, and short messages, we see many other challenges, where translingual technologies are required.

An important area of application is the interpretation of lectures. In spite of excellent scientific equipment and funding, German universities, for example, are often disadvantaged in international competition for talents, simply because many

4. https://www.skype.com/en/features/skype-translator.

Figure 14.9 The Skype translator: consecutive interpretation on a telephone network.

foreign students or scientific employees and academics do not want to learn another new language (especially such a difficult language as German). How are German universities or German companies to react? Is a German university supposed to have all courses and lectures presented in English? The author of this article does not consider this desirable or practicable. A hybrid solution with the help of modern language technologies that supports linguistic and cultural diversity and tolerance (and does not suppress one or the other direction) appears to be far more promising, as it fosters and improves internationalization and international understanding.

At Karlsruhe Institute of Technology (KIT), such a system is being used for students in the main auditorium [Cho et al. 2013]. Speech translation continues to be the subject of research, as not all problems are solved. But thanks to continuing benchmarking and competitive evaluations (the IWSLT campaign, EU funded programs like EU-BRIDGE, etc.) consistent improvements and advances can be observed. While output is far from perfect (and falls short of expert human interpreters) for a listener at a University or conference, who does not speak the language of the lecturer, an imperfect computer-based interpreter is better than nothing.

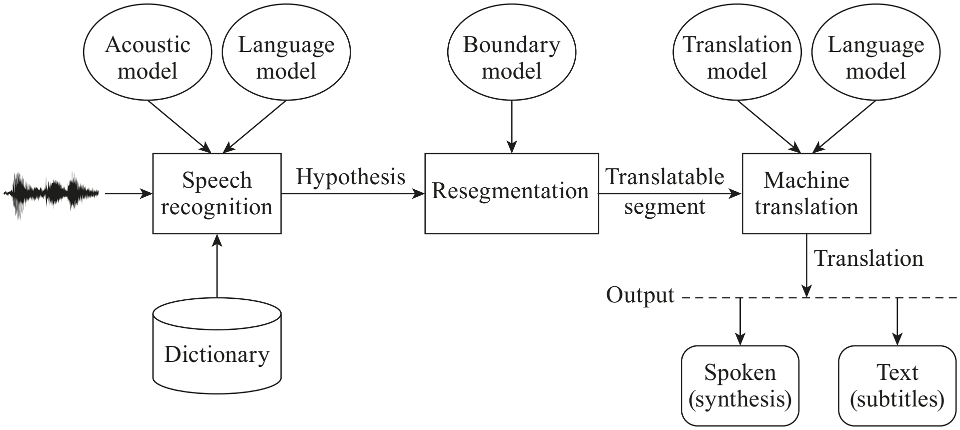

Figure 14.10 Speech translation of lectures.

The first such system was proposed by researchers at CMU, KIT and Technologies in 2005 (see Figure 14.10)[Fügen et al. 2007]. It manages such a simultaneous interpretation use-case uni-directionally as a monologue to be translated into a target language. Such a system does not have to be run on a mobile device, but may be operated on servers in a cloud-based manner and accessed via the internet. Contrary to a translating system for dialogues, a lecture translator requires just a speech recognition component and a translation machine, if only subtitles are desired. Speech synthesis can take place afterward, but it is optional. In addition, a segmentation component is required to decide explicitly or implicitly when the end of a sentence or at least of a translatable fragment is reached in the stream of words. Segmentation can also be performed incrementally with multiple segmentation hypotheses to be explored in parallel, during execution. Vocabularies containing many technical terms and jargon, foreign words, and expressions, formulas and acronyms present an additional range of problems for lectures.

A lecture translator may be operated in two modes: as a simultaneous interpretation system *during* a lecture and also *afterward* over recorded archival lectures in a post-processing mode for retrieval and review. Simultaneous interpretation is required when a listener wishes to follow along while present in a lecture, and both the input language (transcript) and/or translations can be presented. Simultaneous use requires real-time recognition and translation (i.e., the system has to keep up with the speech). Latency (i.e., the time lag between the spoken word and the translated word) also has to be minimized. Otherwise, the listener will lose track of the lecture and of what is happening in the lecture hall. For some languages (especially German, as it turns out), these requirements are a challenge, when verbs or important parts of the verb appear at the end of a sentence (or sometimes even later), thus introducing substantial uncertainty when decisions have to be made before a sentence is completed. The verb "vor-schlagen" means "to propose," and "schlagen" (without the prefix "vor") means "to hit." But in a sentence such as "Ich schlage Ihnen nach eingehender Prüfung Ihres Antrags, der uns gestern . . . eine neue Vorgehensweise . . . vor" (translation: I propose to you after considerable review of your proposal a new approach . . . .), German syntax strips the leading prefix of "vor-schlagen" off and moves it to the end of the sentence, after potentially many words and minutes of speech later. Appropriate interpretation of German in a low latency mode thus keeps us guessing how the story might end and forces an early translation decision before all the information is in.

In many application scenarios of academic teaching and multimedia broadcasting, offline processing of speech and translation are acceptable and sometimes desirable. Offline operation does not necessarily require real-time capability, although an excessively long processing time may become a relevant cost factor. Furthermore, the system can make a better transcription and translation when taking into account a longer context. A lecture translator, for instance, may be run online in the lecture hall and the output may be reprocessed in offline mode later for storing an improved version for listeners in the archive.

Such a lecture translation system was taken into operation at KIT in 2012 as an internet service in several of its main lecture halls (Figure 14.11) [Cho et al. 2013, Greve-Dierfeld 2012]. Students, who wish to have translation support, connect their phones, tablets or PCs to a course-website via a normal internet web browser and are provided with a simultaneous transcription of the text in German (useful in case of hearing problems) and a translation into English. Output languages include French, Arabic, Spanish, and further languages are under development.

Even though the system is already deployed and in actual use, many linguistic and machine learning problems remain. They continue to be subject of ongoing research. In addition to the problems of word order and verbs discussed above, the following difficulties are encountered (particularly in the German language).

- **Compound words**. German words like "Fehlerstromschutzschalterprüfung" (examination of the protective electric currency malfunction switch) switch first have to be decomposed before they can be translated into English. Algorithms for compound word decomposition have to be developed. Due to the ambiguities of language, however, this, too, is not necessarily easy. While de-

Figure 14.11 The lecture translator in use in the main auditorium of KIT.

composition into *Fehler-Strom-Schutz-Schalter-Prüfung* in our example may be straight forward, decomposition of "dramatisch" (dramatic) into "Drama-Tisch" (drama table) or of "Asiatisch" into "Asia-Tisch" (asia table) are inappropriate in the context or even change the intended meaning [Koehn and Knight 2003].

- "**Agreement**". Suffixes in the German language have to be consistent and agree with the nouns: "in der wichtigen, interessanten, didaktisch gut vorbereiteten, heute und gestern wiederholt stattfindenden Vorlesung" [in the important, interesting, well prepared today, . . . lecture]. The suffixes of each adjective depends on the gender and case of the final noun.
- **Technical terms, jargon and unknown words**. This is a major problem, in particular when processing lectures at a university, because every lecture has its own technical terms and linguistic features. What are "Cepstral-Koeffizienten" (cepstral coefficients), "Walzrollenlager" (roller bearings), and "Würfelkalküle" (cube calculi), and how do we translate them? In order to avoid major dictionary maintenance efforts usable systems must seek out the necessary information from other complementary resources across multiple modalities by themselves. As one effective solution to this problem, automatic algorithms can be devised that search the video and presentation materials of a lecturer for clues to the most likely interpretation of a speaker's

speech during a lecture (see Waibel, US patents 2013–2018). Technical terms are automatically extracted from the slides and related terms identified on the internet. Unknown words are then added to the recognition vocabulary and translations derived from internet sources, such as Wikipedia [Niehues and Waibel 2011]. In addition to finding unknown words, performance is improved by cross-referencing spoken language with words and concepts from the corresponding slides.

A second alternative to the problem of unknown words (beyond technical terms these typically also include names, foreign words and abbreviations) is to include human assistance, either by professional editors or the crowd-sourced spontaneous edits from student users. This is done online during a lecture or after the fact in the archive, and the ground truth obtained in this manner, provides further opportunities for machine learning to improve overall system performance over time. Of course, all these methods build on successful interaction with a human user and thus depend greatly on a well-designed multimodal user interface, designed to establishing context and obtain corrections naturally, seamlessly and unobtrusively.

- Code switching. Often, lectures and speeches contain quotations and phrases from other languages. Especially computer science lectures are peppered with English terms that are typically not translated into German. Germans talk about the "iPhone," "iPad," "cloud-basiertem Webcastzugriff" (cloud-based webcast access), or "Files," that are "downgeloaded" thus mixing English words with German declination rules and compounding!
- Pronouns. What do pronouns refer to? Here, problems occur rather frequently. The spoken word version of "Wir freuen uns, Sie heute hier begrüßen zu dürfen" may be translated as "We are happy to welcome her here" or "We are happy to welcome you here" (in writing, Germans use a capital and a small "s" to distinguish both versions).
- Readability. When people speak, they do not speak punctuation marks or the ends or starts of paragraphs contained in readable text. Hence, full stops, commas, question marks, paragraphs, and sometimes even titles have to be generated and inserted automatically [Cho et al. 2014].
- Spontaneous speech. Different speakers speak more or less syntactically. Hesitations, stuttering, repetitions, and discontinuations of speech aggravate readability and make translation difficult. A spoken sentence of a lecture transcribed by a perfect speech recognition system would contain all such

disfluencies and have no punctuation marks. We must therefore process the raw output from speech recognition first linguistically in order to make it readable in the source language. It can then be translated into readable text in the target language [Cho et al. 2014b].

- Microphones and noise. The Karlsruhe lecture translator presently is configured to accept input from a dynamic noise-canceling microphone that the lecturer wears. This is acceptable during lecturing, as lecturers carry microphones in auditoriums anyway, but for seminars and meetings this may be a distraction. Unfortunately, signals from distant microphones or table top microphones are distorted by reverberation, noise, and the potentially overlapping speech from several speakers, which leads to considerable losses in recognition performance.
- Linguistic scalability/portability. How can we implement the technologies developed not only in one or two languages, but extend it to cover communication among all languages and cultures on our planet? To achieve this, development costs of a translation system would have to be reduced considerably. Language-independent technologies, adaptation, inference, abstraction, better use of monolingual resources, and crowd sourcing (to better harvest the multilingual knowledge of mankind) are promising approaches.

The architecture of the Lecture Translator for practical use was introduced in 2005 as a research prototype between CMU, KIT, and Mobile Technologies, and went into service in 2012 at KIT. Infrastructure techniques and support were developed and merged with other sites under the EU-BRIDGE integrated project of the European Union (see Figure 14.12). Now, several lectures and multiple sites are supported in a cloud-based manner at the same time. In recent developments, Microsoft also launched a similar lecture interpreting service. Its features and capabilities are similar to the one described above. It provides integration with PowerPoint from the Microsoft Office suite to permit subtitling during presentations. Using the merged PowerPoint slides it also merges information from the slides' content in similar ways as described above (Waibel 2013–2018).

By means of this server architecture, translation services can be used in several auditoriums and other application scenarios (not only during lectures at university).

Apart from use at universities (where usually no translation support is provided), automatic systems can also be applied to support experts, for example human interpreters at parliaments. In 2012, 2013, and 2014, the lecture translation system

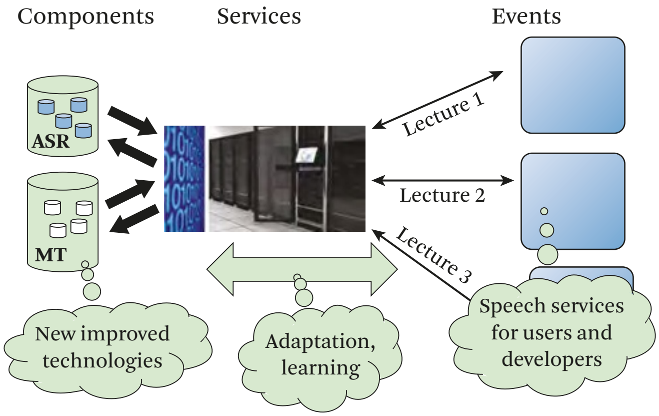

Figure 14.12 EU-BRIDGE: The automatic interpreter as a cloud-based service.

Figure 14.13 Automatically translated speech at the European Parliament.

was presented at the European Parliament, at several Rectors' Conferences (see Figure 14.13), and was featured in training courses for interpreters (see Figure 14.14). The goal of such discussion is to develop possible symbiotic human-machine arrangements that support human interpreters in their efforts to deliver high quality interpretation efficiently. A first test in interpreting booths of the European Parliament was carried out successfully in late 2014 under the EC Integrated Project "EU-BRIDGE" in Strasbourg. [Koehn et al. 2015]

Figure 14.14 First test of an automatic interpreter during voting at the European Parliament.

Because the European Parliament operates one of the world's broadest government interpretation efforts working continuously with more languages (23×23 language directions!) than any other organization and employs some of the most experienced and sophisticated human interpreters, interpretation services are already at their best in terms of quality and sophistication. Automated language processing and interpretation tools in these settings thus serve a different role from the settings discussed before: rather than performing fully automatic interpretation, they aim to support, enable and *amplify* human effort so as to achieve greater quality, speed, and scale in the face of overwhelming demand.

To date, three use cases were identified that instantiate such complementarity: (1) a generator of on-demand terminology lists and their translation, for example, if a session on "fishery" is scheduled, the system automatically serves up special terms pertaining to their domain and delivers it to the assigned interpreter along with appropriate dictionary lookups; (2) named entity and number tracking (to recall numbers and names more easily); and (3) the "Interpreter's Cruise Control," intended to handle repetitive (boring) segments of a session (such as, for example, voting sessions), or where human interpreters are not available. A sophisticated, multimodal interface is essential to deliver such human-machine symbiosis, seamlessly. The services access available resources (schedules, agendas, reports, dictionaries, and lexica) and deliver the desired support to EU interpreters on tablets or laptops. The services exist so far are in a prototype stage, but user studies and evaluations have shown the success of these tools, more than 60% of the interpreters were satisfied or very satisfied with the final tool [Stüker et al. 2015].

14.4 Multimodal Translingual Communication

In a multilingual and multicultural environment, language barriers are not only encountered in spoken dialogues, lectures, or text documents. They occur in many other communication situations, circumstances, and media: for example, important information can be found on road signs or in short text messages (SMS), TV news, lecture transparencies, gestures, and many more. To make the vision of a multilingual, language barrier-free world come true, our efforts have to go beyond the construction of better translation systems. The goal should rather be to build user interfaces that make language barriers transparent or move their existence into the background. Successful translingual communication is achieved, when people can interact with each other without being aware of the barriers between them! In the following, we discuss multimodal system designs, where this was attempted and achieved with varying degrees of effectiveness. The processing of multiple modalities is particularly important and beneficial in two situations: (1) recovering from miscommunications that may result from occasional human misunderstandings or from machine recognition or translation errors; and (2) when responding naturally to multimodal communicative clues in varying multicultural scenarios.

14.4.1 Multimodal Error Handling

Miscommunication in speech dialogue translation can result from errors during the speech recognition or the MT processes, and the causes are often not readily identifiable by the user. Worse, the translation of a misrecognized word rarely bears any resemblance to the translation of the correctly recognized word, so that the resulting output appears just confusing. Recognizing that an error has occurred, offering tools to recover from such errors, and algorithms to even learn from such correction, are subjects for considerable research on learning algorithms and effective multimodal interface design.

Several types of errors can occur in the process of cross-lingual interpretation.

- **Out-of-Vocabulary Words (OOVs).** Most commonly the problem arises when words are missing in the pronunciation dictionary of a recognizer, leading to one or more substitution errors. Named entities and specialty terms are particularly prone to this type of problem.
- **Speech confusions.** Words that are phonetically close (“forest” and “far East”) or homophones (“two” and “too”) can lead to substitution errors due to their acoustic similarity, even though they may differ semantically.

- **Translation Errors.** Words can have multiple translations. While translation and language models attempt to select the most appropriate translation, occasional inappropriate choices remain and need to be corrected.

14.4.1.1 Error Detection

Before attempting a correction, a problem has to be identified. A speaker may determine that a recognition error has occurred and intervene, if s/he pays careful attention to the transcript but this may not be possible in all situations. Translation errors may even be harder to detect for a speaker who does not know the target language. Two typical solutions are (1) Confidence measures to judge the reliability of the recognition and translation outputs and (2) back translation into a source language so that an input speaker may verify that a translation appears to be correct. A variety of confidence measures exist for ASR and MT engines; most typically compute an a posteriori probability of the word to be correct. Although the measures help identify errors in translation, there is unfortunately no guarantee that they will accurately flag problems or that flagged problems are actually errors. During consecutive translation, such methods are also far more likely to succeed, since both speakers have a joint interest in being understood, and have the time and interest to collaboratively resolve potential miscommunications.

14.4.1.2 Error Repair

How are miscommunications resolved, once they have been identified? Two methods have been proposed in the case of actual misrecognitions or mistranslations: (1) clarification dialogues; and (2) cross-modal repair. In the former, the system will initiate a disambiguation dialogue (triggered by a confidence measure) to get a user to resolve a potential error through a voice dialogue. In the latter, the user (or the system) may divert to another modality to clarify.

Clarification dialogue. In the former approach, once a putative error has been identified, the system attempts to initiate a clarification dialogue with the user. If the system misrecognizes "my name is Edwards," as "my name is *at words*," a clarification component might ask for clarification on the misrecognized word, if the error is correctly identified. In such a case, the human user is engaged to disambiguate the confusion through a clarification dialogue "is ATWORD a name?." Errors in recognition can be caused by OOVs, homophones, or substitutions with similar sounding words, but the detection of errors is a non-trivial classification task in itself. If an error is not recognized as an error, it is missed and cannot be repaired; if a correct word is flagged as an error, it may generate an unnecessary clarification dialogue and may be a nuisance to the user. Early versions of such clarification dialogues have already been explored in early studies [Block et al. 2000]. More rigorous evaluations using clarification dialogues were conducted under the DARPA program BOLT using simulated field data that investigated the efficacy of language-based error repair dialogues. Even though good performance in detecting and repairing errors was achieved through voice dialogues alone [Kumar et al. 2015], such dialogues take time and are generally much slower than and thus inferior to a multi- or cross-modal repair strategy. If an error is visible on the screen, and alternate input modalities are available, errors can be reliably detected by humans and corrected through typing, gesturing, handwriting or spelling, for example. Thus, unless the use-case is strictly a hands-eye-busy voice situation, multimodal repair appears to be more effective [Suhm et al. 1999, Kumar et al. 2015].

Cross-Modal Repair. In cross-modal repair, the error is corrected by diverting to an alternate, hopefully orthogonal modality, such as typing, handwriting, spelling, paraphrasing, etc. The advantage of this approach is that it can proceed in parallel to speaking. Generally, it is thus considerably faster to correct an error by pointing, clicking, and editing, rather than through a disambiguating dialogue [Suhm et al. 1996, 1999, Waibel et al. 1991, 1998a, 1998b, Oviatt and VanGent 1996]. The simplest form of such repair is to simply observe the error and correcting it through typing. Alternatively, however, it is possible to point to the error and spell, handwrite, paraphrase, or respeak a correction. Such correction is fast, potentially more natural, and exploits the orthogonal sources of errors in each modality to obtain a jointly optimal result [Suhm et al. 1999, Waibel et al. 1998a, Oviatt and VanGent 1996].

Learning Words. OOVs are particularly troublesome for speech translators, since no matter how the user may correct the input, recognition and translation will fail every time, if the missed word is not included in the processing dictionaries. A typical way to handle OOVs in MT is to simply pass the unknown word through to the other side.[5] If the two languages in question use the same script (say, English and Spanish), this may lead to acceptable results: an unknown name, for example, may appear in the same script as the name in the other language. However, this is not acceptable if the scripts (e.g., Chinese and English) differ. On the recognition side, OOVs are particularly problematic, since their absence from the

5. It is worth noting that this approach is no longer so simple in current implementations of NMT, due to the absence of phrase tables.

recognition lexicon will force another match and thus lead to substitutions errors.[6] In a speech translator then, such substituted words will be translated in curious, irrelevant ways that have no resemblance (neither phonetically nor semantically) from the original intended message [Kaiser 2005]. In research speech translation systems, the problem of OOVs is mostly handled by adding the missing words to the various dictionaries and language components manually. This involves a total of eight modifications: the pronunciation dictionary in language L1 has to be provided ("Paul"—[P AO L]), the language model has to be modified to include the "Paul" in a word sequence (e.g., "my name is Paul"), "Paul" has to be translated to "Pablo," and we may need a pronunciation dictionary to properly pronounce "Pablo" in Spanish. If the system we are building is a bidirectional dialogue system, the appropriate modifications have to be made in the reserve direction as well.

The modifications involve knowledge of phonetics and statistical language models that are easily done in research labs, but they cannot be performed by non-expert users in the field. OOVs (for example, the occurrence of names) found in the field are also not predictable a priori and modifications really must be done by the user. Interactive multimodal interface solutions have thus been proposed [Waibel and Lane 2015, Kaiser 2006] that shield the required technical detail and allow a non-expert to make vocabulary additions in the field, interactively and intuitively. The interface accepts orthography of a word/name to be added, it then generates the appropriate model entries automatically in the background and modifies all system components dynamically. It provides intuitive, interactive sound checks to make sure the pronunciation is correctly represented. When the same name is then uttered again (in either language), it is recognized and translated appropriately. This functionality was extensively tested during humanitarian deployments and in a commercial deployment on a smartphone App (Jibbigo).

14.4.2 Multimodal, *Flexi*-modal Communication

In cross-lingual communication, multimodal interfaces are not only useful to recover from errors generated by speech recognition or machine translation, they also open up a broad array of cross-lingual communication channels. We recall that our goal is not just speech or text translation, but to provide humanity with a human-human communication experience in which linguistic and cultural barriers become transparent. As such, we must be concerned with the full breadth of

6. Here, recent character based neural approaches may offer potential solutions in the future by generating character strings directly without the use of dictionaries [Zenkel et al. 2017, Miao et al. 2015, Zweig et al. 2016].

Figure 14.15 Road sign translator (2001) and Google Translate (Wordlense) 2015.

human expression and assist in their mediation, in whatever modality, context, or situation humans may find themselves. Human language is given by speech, but also images, text, handwriting, even gestures, and facial expressions. Input and output of language may be suitable using one modality in one situation but awkward in another. A successful cross-lingual communication system design must carefully match input and output modalities with devices in each situation.

Over the last 10–20 years considerable progress has been made in achieving this goal.

- **Road sign translators**. As early as 2001 [Yang et al. 2001, Waibel 2002], first systems were developed and commercialized to read and translate road signs with the help of a mobile device and camera. Translations were inserted into the image of the scene and the system was tested first on a (then applicable) PDA platform. Translations of text found in road sign images were then displayed as subtitles under the signage. Meanwhile, similar applications have been developed and issued as iTunesTM and AndroidTM apps for smartphones. A more recent development offered by Wordlense, a startup company (now incorporated in Google Translate), combined a simple recognition engine for Western script and translation dictionaries with graphical rendering that inserts the translated word back into the original image. Both applications (shown in the left and right of Figure 14.15) demonstrate how an integrated multimodal design is equally essential to achieving our goal of language transparency as the language technology itself.

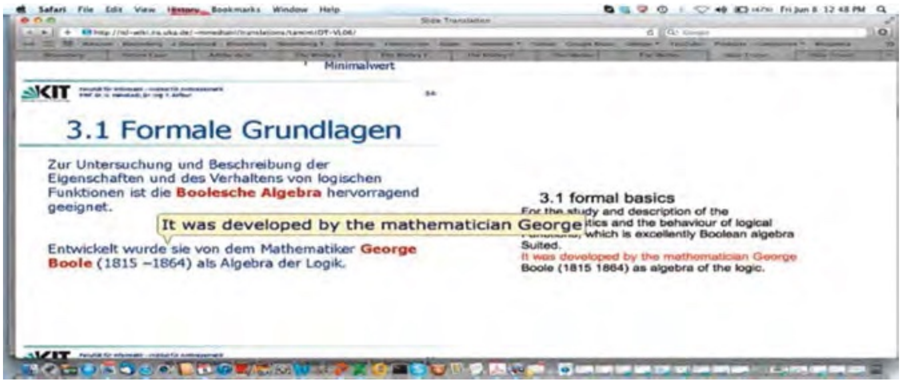

Figure 14.16 Translation of lecture slides.

Handwriting recognizers recognize handwritten text and provide text in translation. The problem of text translation in real images has been solved partly solved by the road sign translator or OCR scanners as discussed above [Zhang et al. 2002], but using handwriting still offers additional real-time low latency opportunities. The difficulties of recognizing handwriting also require more sophisticated recognition akin to speech recognition. Early neural network-based and HMM-based systems have been proposed since [Jaeger et al. 2000, 2001, Manke et al. 1995, Starner et al. 1994], and recent solutions have matured in performance and usability,[7] so as to permit integration into sophisticated commercial grade multimodal communication interfaces.

- **Translation of lecture presentation material**. If foreign students have difficulties understanding lectures in a foreign language, then the lecturer's presentation slides or handouts might generate communication issues, too. For this reason, translation can also be applied to material across these different media, as well. Figure 14.16 shows a prototype translation system for PowerPointTM slides (explored at KIT) that translates the text on a slide, when hovering with the mouse over the appropriate text. The translated text is then displayed in a speech bubble.
- **Distant speech input**. A continuing issue problem with speech translation devices is the placement of the microphone. When wearable or hand-held

7. http://www.myscript.com.

microphones are acceptable (e.g., lectures or mobile smartphones), this is of little concern, since the speaker's speech is well discernable and signal quality generally good. However, in meetings, noisy public places, and many other situations, signal separation, reverberation, and noise become significant factors in delivering successful translation services. One method to mitigate these factors are microphone arrays that are placed in a strategic location (e.g., on a table in a meeting room), worn (necklaces), or moved on a robot. Fujitsu and NICT are testing directional microphone technology under a national project aimed to deliver communication for the 2020 Olympics in Japan. Other mobile microphone arrays were proposed as attachments on smartphones or individual devices and (of course) for non-translation purposes conversational speech dialogue pods for the home such as Alexa and Google Home.

- **Silent speech input**. Speech is audible and thus perceived as noise for those for whom it was not intended. Is noise-free speech conceivable? Alternate non-vocal speech recognition systems have indeed been proposed, where articulated mouth movements are captured by electromyography, even though the language is not spoken out loud. Such "silent speech" can be recognized (although recognition is not as good as for spoken language), translated, and made audible by synthesis [Maier-Hein et al. 2005]. Subsequently, articulation of silent speech can be translated into audible speech in another language. The underlying technology is not yet mature, but the proposed prototypes (see Figure 14.17) show that input devices could be devised that can accept *silent* spoken language as an alternative modality where speaking aloud would create disturbances (or privacy concerns). Using silent speech input technology might thus be conceivable so that *anyone* can produce loud speech in *any* language, by (silent) articulatory motion in another.
- **Targeted Audio**. Synthetic speech output in a speech translation system can also be delivered selectively by directional loudspeakers. Such speakers were proposed experimentally (CHIL-project, [Waibel and Stiefelhagen 2009]), and commercially distributed (Sennheiser AudioBeam Ultrasonic Directional Loudspeaker).[8] By directing such speakers to different points in a room different listeners can then listen to simulatenous translation in different languages without a headset. It is as if each listener has a personal interpreter whispering the interpreted result into his or her ear. Early steer-

8. Sennheiser Electronic GmbH & Co. KG, product currently discontinued.

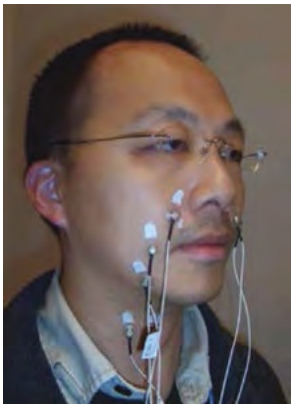

Figure 14.17 Silent speech as input to translator.

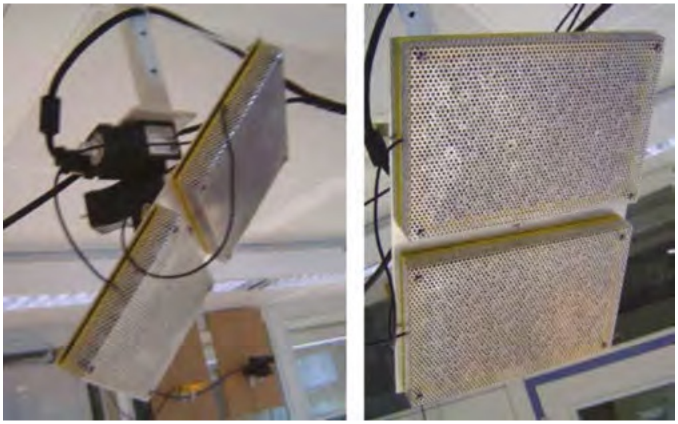

Figure 14.18 Translated speech delivered through targeted audio speakers: Personal, audible interpretation without headsets.

able prototypes have been developed and proposed [Waibel and Stiefelhagen 2009] for meeting rooms where the appropriate output interpretation can be positioned toward specific individuals (or guided by face recognition) (see Figure 14.18).

- **Speech translation goggles**. As alternative to personalized acoustic delivery of translation output, such output can also be delivered in visual form. In 2005, such a cross-modal speech translated was first proposed at a press

Figure 14.19 First demonstration of heads-up display "translation goggles" at Carnegie Mellon (2005).

conference at CMU/KA where translated output from a simultaneous (lecture) translation system was delivered textually via heads-up display goggles (Figure 14.19). In this configuration, the user faces a conversation partner or a lecturer and the translation of spoken words is displayed as text as subtitles in the glasses. While this still seemed like science fiction in 2005, such configurations are now becoming reality as mobile computing platforms (smartphones, smart watches) can be connected with wearable augmented and virtual reality eye-glasses (Google "Glass", Snapchat's "Spectacles", Facebook's Oculus) that are also becoming pervasive and commonplace. Google already proposed a speech translator just like it as a feature for Google Glass (Figure 14.20), Google Glass Demo[9] and others are sure to follow.

- **Earplugs and Pixel-Buds**. Another form factor that has recently attracted attention are earphone style devices (perhaps inspired by the Babel Fish from the science fiction series Hitchhiker's Guide to the Galaxy). Here, a set of earplugs provides input and output for a speaker attempting to dialogue with others. The underlying technology is similar to the systems described above, but speech is delivered through earbuds instead of to a phone's microphone. A young start-up, "Waverly Labs," announced to bring a product (the "Pilot") to market, and Google recently launched a similar product called "Google Pixel Buds". Google's system combines Google Translate with speech I/O from and to the earbuds.

9. https://www.youtube.com/watch?v=MqZuscmCYi4.

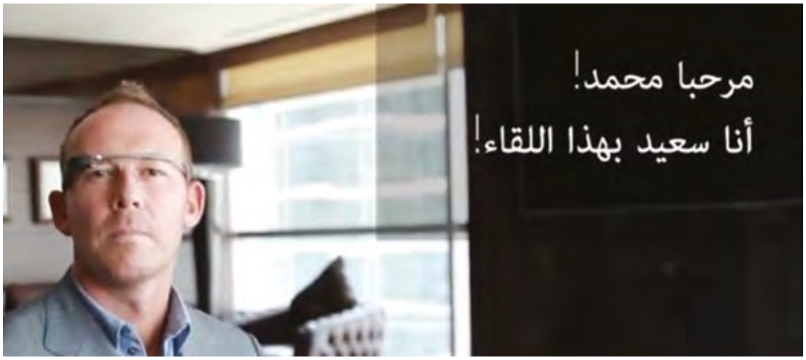

Figure 14.20 MITE: translation via Google glass (2014).

Given all the advances with systems that can provide a translation function in a variety of situations, speaking styles and across multiple modalities, true integrated multilingual and multimodal environments become possible. With input accepted from personal, directional microphones, by electromyography speech or handwriting, and output translations delivered via directional targeted audio speakers, heads-up display goggles, personal displays on smartphones and tablet, language transparent conversations, and meetings become possible. So far, such fully integrated systems have been demonstrated only as prototypes or concept demonstrations (see Figure 14.21), but with continued progress they will likely transform the way we communicate in the global village of the future.

14.5 Conclusion

Multimodal interfaces represent a critical dimension to building effective systems that support cross-lingual communication. Depending on the situation (lectures, meetings, one-on-one conversations, mobile dialogues, telephone conversations, blackboard notes, handwriting, texting, and many more), language is communicated in different modalities and at different speeds. Input must be accepted in different forms and output translation delivered in different modes and presentation styles, depending on use-case and personal preference. Perceptual input processing and translation technology has to be carefully adapted optimized to deliver the best language translation accuracy, at the appropriate speed, and latency, as required by the application. All processing steps also have to sensitive to the conversational context. With multimodal ("fleximodal") interfaces, dialogues across language barriers can be supported effectively. Multimodal interfaces can better

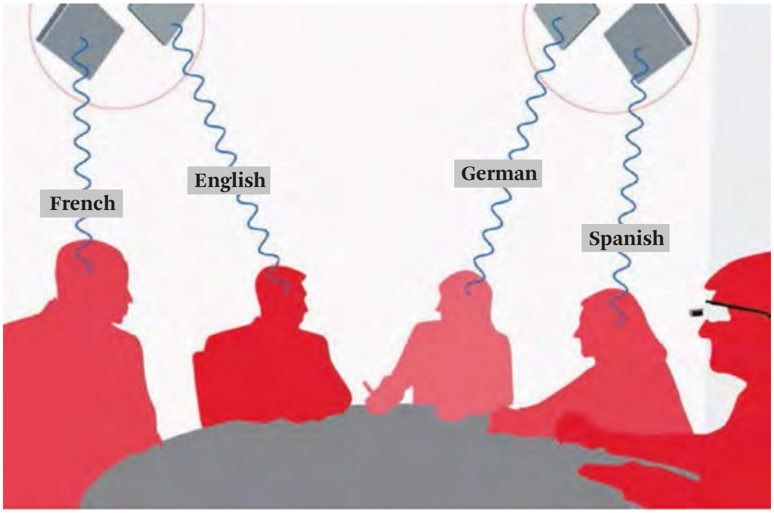

Figure 14.21 Individually adapted simultaneous translation in meetings.

compensate for errors, improve the speed of communication, adapt and scale, and respond to user communication needs and environments. Suitable interface design is as important to the success of cross-lingual communication tools in practice, as the performance of the underlying technology components.

Focus Questions

14.1. What are the three partial tasks that need to be solved for speech-to-speech translation?

- ASR
- MT
- Speech Synthesis

14.2. What are the three main components of a modern speech recognition system?

- Acoustic Model
- Dictionary
- Language Model

14.3. What NN architecture is typically used in neural machine translation?

- Recurrent NN

14.4. What are current research questions in speech translation?

- Compound words
- Agreement
- Technical terms, jargon and unknown words
- Code switching
- Pronouns
- Readability
- Spontaneous Speech
- Microphones and noise
- Linguistic scalability/portability

14.5. Which types of errors can occur during cross-lingual interpretation?

- Out-of-Vocabulary Words
- Speech confusion
- Translation errors

References

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 584, 586

H. U. Block, St. Schachtl, and M. Gehrke. 2000. Adapting a large scale MT system for Spoken Language. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pp. 394–410. Springer Published, Berlin/Heidelberg, Germany. 606

H. Bourlard and N. Morgan. 1994. *Connectionist Speech Recognition, A Hybrid Approach*. Kluwer Academic Publishers.

H. Bourlard and Ch. Wellekens. 1989. Speech pattern discrimination and multilayer perceptrons. In *Computer Speech and Language*, (3), pp. 1–19. 587

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2): S. 263–311. 590

E. Cho, J. Niehues, T. L. Ha, M. Sperber, M. Mediani, and A. Waibel. 2016. Adaptation and Combination of NMT Systems: The KIT Translation Systems for IWSLT 2016. In *Proceedings of the 13th International Workshop on Spoken Language Translation, IWSLT*. Seattle. 588

E. Cho, J. Niehues, and A. Waibel. 2014a. Tight integration of speech disfluency removal into SMT. *EACL, 2014*, 43. Gothenburg, Sweden. DOI: 10.3115/v1/E14-4009. 585

E. Cho, J. Niehues, and A. Waibel. 2014b. Machine Translation of Multi-party Meetings: Segmentation and Disfluency Removal Strategies. *IWSLT*. Lake Tahoe, US. 601

E. Cho, C. Fügen T. Herrmann, K. Kilgour, M. Mediani, C. Mohr, J. Niehues, K. Rottmann K. Saam, S. Stüker, and A. Waibel. 2013. A Real-World System for Simultaneous Translation of German Lectures. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*. Lyon, France. 596, 598

K. Cho, B. van Merrienboer, Ç, Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pp. 1724–1734. Doha, Qatar. DOI: 10.3115/v1/D14-1179. 600

M. Eck, I. Lane, Y. Zhang, and A. Waibel. 2010 Jibbigo: Speech-to-Speech translation on mobile devices. In *2010 IEEE Spoken Language Technology Workshop*, 165–166. DOI: 10.1109/SLT.2010.5700843. 592

The Economist. June 12, 2006. How to build a Bablefish. In *The Economist*.

Ectaco. 1989. Ectaco eBook Readers and Translators. http://www.ectaco.com. 592

C. Fügen, A. Waibel, and M. Kolss. 2007. Simultaneous Translation of Lectures and Speeches. In *Journal Machine Translation*, 21(4), 209–252. DOI: 10.1007/s10590-008-9047-0. 597

A. von Greve-Dierfeld. 2012. Uni-Übersetzungs-Automat: Don't worry about make. Spiegel-Online. http://www.spiegel.de/unispiegel/studium/dolmetscher-fuer-die-vorlesung-kit-entwickelt-uebersetzungsprogramm-a-838409.html 598

T.-L. Ha, J. Niehues, and A. Waibel. December 2016. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*, 8–9. Seattle, WA. 586

Handelsblatt: Übersetzung. July 1991. Handelsblatt. http://isl.anthropomatik.kit.edu/cmu-kit/downloads/1991.07.30_Handelsblatt.pdf. 588

K. Hatazaki, J. Noguohi, A. Okumura, K. Yoshida, and T. WatanabeT. 1992. INTERTALKER: an experimental automatic interpretation system using conceptual representation. In *Second International Conference on Spoken Language Processing*. 588

S. Hourin, J. Binder, D. Yeager, P. Gamerdinger, K. Wilson, and K. Torres-Smith. 2013. Speech-to-Speech Translation Tool Limited Utility Assessment Report. Report OMB No.0704-0188. 593

S. Jaeger, S. Manke, J. Reichert, and A. Waibel. 2001. Online handwriting recognition: the NPen++ recognizer. *International Journal on Document Analysis and Recognition*, 3(3): 169–180. DOI: 10.1007/PL00013559. 609

S. Jaeger, S. Manke, and A. Waibel. September 2000. NPen++: An On-line Handwriting Recognition System. In *Proceedings of the 7th International Workshop on Frontiers in*

Handwriting Recognition, IWFHR 2000, Amsterdam, The Netherlands, 11–13. DOI: 10.1.1.30.158. 609

A. Jain and A. Waibel. August 1989. A Connectionist Parser Aimed at Spoken Language. In *Proceedings of the 1st International Workshop on Parsing Technologies, IWPT 1989*, Pittsburgh, PA. 28–31. DOI: 10.1109/ICASSP.1990.115782. 585

M. Johnson, M. Schuster, Q.V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. 2017. Google's multilingual neural machine translation system: enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351. 586

E. C. Kaiser. 2006. Using redundant speech and handwriting for learning new vocabulary and understanding abbreviations. In *Proceedings of the 8th ACM International Conference on Multimodal Interfaces*, pp. 347–356. ACM Press. DOI: 10.1145/1180995.1181060. 607

E. C. Kaiser. 2005. Multimodal new vocabulary recognition through speech and handwriting in a white-board scheduling application. In *ACM Intelligent User Interfaces Conference, IUI '05*, pp. 51–58. ACM Press, New York. DOI: 10.1145/1040830.1040851. 607

N. Kalchbrenner and P. Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1700–1709. 584

P. Koehn, Y. Zhang, C. Dugast, J. Gauthier, S. Grimsey, S. Fünfer, M. Mueller, S. Stueker, and V. Steinbiss. EU-BRIDGE D6.3 Final Evaluation Report, (www.eu-bridge.eu). 602

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL, Prague, Czech Republic. 590

P. Koehn, and K. Knight. 2003. Empirical Methods for Compound Splitting. *EACL*, pp. 187–193. Budapest, Hungary. DOI: 10.3115/1067807.1067833. 599

R. Kumar, S. Hewavitharana, N. Zinovieva, M. E. Roy, and E. Pattison-Gordon. 2015. Error-Tolerant Speech- to-Speech Translation. In *Proceedings of MT Summit XV, Volume 1: MT Researchers' Track*, pp. 229–239. Miami, FL. 606

A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zhan. 1997. JANUS III: Speech-to-Speech Translation in Multiple Languages. *International Conferences on Acoustics, Speech, and Signal Processing, ICASSP*. Munich, Germany. DOI: 10.1109/ICASSP.1997.599557. 590

W. Lewis. November 2015. Skype Translator: Breaking Down Language and Hearing Barriers. *AsLing's 37th Translating and the Computer Conference*, 27–28. London. 595

L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel. November 2005. Session Independent non-audible speech recognition using surface electromyography. In *Proceedings of ASRU*, Cancun, Mexico. DOI: 10.1109/ASRU.2005.1566521. 610

S. Manke, M. Finke, and A. Waibel. 1995. The use of dynamic writing information in a connectionist on-line cursive handwriting recognition system. *Advances in Neural Information Processing Systems*, 1093–1100. 609

Y. Miao, M. Gowayyed, and F. Metze. 2015. EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-Based Decoding. *Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ. 588, 607

R. Miikkulainen and M. D. Dyer. 1989. A modular neural network architecture for sequential paraphrasing of script-based stories. In *Proceedings of the International Joint Conference on Neural Networks*. IEEE. DOI: 10.1109/IJCNN.1989.118677. 585

T. Morimoto, Takezawa, F. Yato, S. Sagayama, T. Tashiro, M. Nagata, and A. Kurematsu. 1993. ATR's speech translation system: ASURA. In *Proceedings Eurospeech '93*, pp. 1291–1294. Geneva, Italy. 588, 590

G. Neubig. 2016. Lexicons and Minimum Risk Training for Neural Machine Translation: NAIST-CMU at WAT2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT)*. 588

J. Niehues and A. Waibel. 2011. Using Wikipedia to Translate Domain-specific Terms in SMT. In *Proceedings of the Eight International Workshop on Spoken Language Translation (IWSLT)*. 600

F. J. Och, and H. Ney. 2004. The alignment template approach to statistical machine translation. In *Journal Computational Linguistics*, 30(4): pp. 417–449. DOI: 10.1162/0891201042544884. 590

S. Oviatt and R. VanGent. 1996. Error resolution during multimodal human-computer interaction. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, Vol. 1, pp. 204–207. DOI: 10.1109/ICSLP.1996.607077.

S. Oviatt and P. R. Cohen. 1992. Spoken Language in interpreted telephone dialogues. In *Computer Speech and Language*, 6(3): pp. 277–302. DOI: 10.1016/0885-2308(92)90021-U. 580

D. B. Roe. 1992. A spoken language translator for restricted-domain context-free languages. In *Speech Communication*, 11(2–3: pp. 311–319. DOI: 10.1016/0167-6393(92)90025-3. 588

S. Stüker, M. Federico, Ph., Koehn, H. Ney, M. Rödder, M. Simpson, V. Steinbiss, and A. Tescari. 2015. EU-BRIDGE Final Report. www.eu-bridge.eu 603

S. Stüker, C. Zong, J. Reichert W. Cao, M. Kolss, G. Xie, K. Peterson, P. Ding, V. Arranz, J. Yu and A. Waibel. 2006. *Speech-to-Speech Translation Services for the Olympic Games 2008, 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, MLMI 2006*, Washington D.C. 592

M. Seligman, A. Waibel, and A. Joscelyne. 2017. TAUS Speech-to-Speech Translation Technology Report. *TAUS Report*.

R. Sennrich, B. Haddow, and A. Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT16)*. Berlin, Germany. 588

T. Starner, J. Makhoul, R. Schwartz, and G. Chou. 1994. On-line cursive handwriting recognition using speech recognition methods. *Acoustics, Speech, and Signal Processing, ICASSP*. DOI: 10.1109/ICASSP.1994.389432. 609

B. Suhm, B. Myers, and A. Waibel. May 1999. Model-based And Empirical Evaluation Of Multimodal Interactive Error Correction. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, CHI 1999*, Pittsburgh, PA. DOI: 10.1145/302979.303165. 606

B. Suhm, B. Myers, and A. Waibel. 1996. Interactive recovery from speech recognition errors in speech user interfaces. In *Proceedings of the International Conference on Spoken Language Processing*, pp. 861–864. Philadelphia, PA. DOI: 10.1109/ICSLP.1996.607738. 606

I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014* pp. 3104–3112. Quebec, Canada. 584, 585

T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto. 1998. A Japanese-to-English Speech Translation System: ATR-MATRIX. In *Proceedings ICSLP'98*, pp. 779–782. Sydney, Australia. 590

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, l. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. *CoRR*, abs/1706.03762. 586

Voxtec, Advancing Voice Technology™. http://www.voxtec.com. 592

W. Wahlster. 1993. Verbmobil—Translation of Face-To-Face Dialogues. In *Grundlagen und Anwendungen der Künstlichen Intelligenz*. Springer-Verlag, Berlin Heidelberg. DOI: 10.1.1.109.6407. 588

A. Waibel. 2002. Portable object identification and translation system. US Patent App. 10/090,559. 608

A. Waibel. 2015a. Translation Training with Cross-Lingual Multimedia Support, 2018, US Patent 9,892,115 B2; CIP Application #14/589,658, filed 2015.

A. Waibel. 2015b. Translation and Integration of Presentation Materials with Cross-Lingual Multimedia Support, 2017 US Patent 9,678,953 B2, 2017; CIP Application #14/589,653, filed 2015

A. Waibel. 2014. Translation and Integration of Presentation Materials with Cross-Lingual Multimedia Support, 2014; Patent Pub. #US 2014/0365202 A1; Appl. # 14/302,146, filed 2014

A. Waibel and I. R. Lane. 2015. Enhanced speech-to-speech translation system and methods for adding a new word, US Patent. 607

A. Waibel and R. Stiefelhagen, editors. 2009. *Computers in the Human Interaction Loop*. Springer, London. 610, 611

A. Waibel and A. McNair. 1998b. Locating and correcting erroneously recognized portions of utterances by rescoring based on two n-best lists, US Patent 5,712,957b. 606

A. Waibel, B. Suhm, and A. McNair. 1998a. Method and apparatus for correcting and repairing machine-transcribed input using independent or cross-modal secondary input. US Patent 5,855,000. 606

A. Waibel, A. Jain, A. McNair, H. Saito, A. Hauptmann, and J. Tebelskis. May 1991. JANUS: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toronto. DOI: 10.1109/ICASSP.1991.150456. 588, 606

A. Waibel, A. Hanazawa, G. Hinton, K. Shikano, and K. Lang. March 1989. Phoneme Recognition Using Time-Delay Neural Networks. In *IEEE Transactions of the Acoustics, Speech and Signals Processing Society*, 347(3). DOI: 10.1109/29.21701. 587

A. Waibel. March 1989. Modular Construction of Time-Delay Neural Networks for Speech Recognition. In *Journal for Neural Computation*, MIT Press Journals, 1. DOI: 10.1162/neco.1989.1.1.39. 587

Y.-Y. Wang and A. Waibel. July 1997. Decoding Algorithm In Statistical Machine Translation. In *Proceedings of the 35th Annual Meeting of the ACL joint with the 8th Meeting of the European Chapter of the ACL 1997*, ACL/EACL 1997, Madrid, Spain. DOI: 10.3115/979617.979664. 590

Y.-Y. Wang and A. WaibelA. May 1991. A Connectionist Model for Dialog Processing. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, ICASSP 1991, Toronto, Canada. DOI: 10.1109/ICASSP.1991.150090. 585

J. Yang, J. Gao, Y. Zhang, X. Chen, and A. Waibel. 2001. An Automatic Sign Recognition and Translation System. *Workshop on Perceptual User Interfaces 2001*, PUI 2001. DOI: 10.1145/971478.971490. 608

T. Zenkel, R. Sanabria, F. Metze, J. Niehues, M. Sperber, S. Stüker, and A. Waibel. 2017. Comparison of Decoding Strategies for CTC Acoustic Models. In *Proceedings of Interspeech*, pp. 513–517. Stockholm, Sweden. DOI: 10.21437/Interspeech.2017-1683. 588, 607

Y. Zhang, B. Zhao, J. Yang, and A. Waibel. September 2002. Automatic SIGN Translation, 7th International Conference on Spoken Language Processing, 2nd Interspeech Event, ICSLP 2002 - Interspeech 2002, Denver, CO. 609

G. Zweig, C. Yu, K. Droppo, and A. Stolcke. March 2016. Advances in All-Neural Speech Recognition. *The 41st IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China. 588, 607