Research Opportunities in Automatic Speech-to-Speech Translation

Sebastian Stüker, Teresa Herrmann, Muntsin Kolss, Jan Niehues, and Matthias Wölfel

I. Introduction

Over the last decades the process of globalization has brought dramatic changes to the way the world interacts. Fast and affordable long-range transportation has brought an unprecedented mobility to people - professionals and tourists alike. The removement of trade barriers has lead to the establishment of an international market in which huge quantities of goods flow between the different countries. The technological revolution on the communication and information sector has enabled instant interaction between every conceivable part of the world. By extending the network that once started as a telecommunication network to an omnipresent data network in the form of the Internet, people were given the capability to instantly access huge amounts of information in all countries in the world.

While technology has provided the means for people to interact, either remotely or face-to-face, it has not given yet the full solution to communication. Though trade and travel barriers have been removed, the diversity of languages in the world remains as a barrier inhibiting intra-cultural interaction.

Ethnologue, a catalogue for languages, lists 7,299 living languages. This richness in diversity of languages is by many considered to be the distinctive and crowning achievement in human evolution. The language of a people or ethnic group is usually of paramount importance to their cultural heritage and an integral part of it.

However, in today's globalized world, languages are frequently disappearing. Linguists, e.g. Janson, estimate that, if the current trend holds, in as little as one hundred years half of today's languages will be extinct. With this loss of languages comes a loss in cultural diversity which needs to be prevented. One reason why many languages die out is the tendency of speakers of languages with a comparatively small population to adopt more widely used languages, such as English. If technology were able to provide translation technology that would enable users to access information provided in any language or to freely communicate with a person speaking a different language, it could counteract this trend.

Today, translations are mostly performed by human translators, e.g. simultaneous interpreters. These manual translations are in most situations, in which they would be useful, either too expensive in terms of money or even impossible due to a lack of suitable translators. Here, modern technology in the form of automatic *speech-to-speech translation* (STST) systems can play a critical role in empowering people to communicate with speakers of a different language and to access or present information in a cross-lingual way—ubiquitously, with acceptable accuracy, and at affordable costs in terms of price of the translation.

Due to the fact that politics and society have recognized the importance of developing this field further it currently receives significant funding all over the world in the form of multiple, highly visible research projects. The European Union (EU), for example, has recognized the importance of its linguistic diversity as an inalienable component of the cultural heritage of the people on its territory—currently 27 countries in which 23 languages are spoken. This led to the adoption of the principle of equality for all languages in the European Union. Since under this principle all languages within the EU are equally important, all communication within the EU, such as speeches in the European Parliament or documents produced by the institutions of the EU, must be provided in all 23 languages. Thus, the speeches in the European Parliament must be simultaneously translated from and into all the languages; just as all written documents must be translated likewise. This forces the EU to spend € 1.1 billion per year on human translations. In order to implement this principle outside its administration, the EU sponsors a multitude of actions in education, training, e.g. language learning programs, and research in linguistic diversity—but also programs seeking technical solutions to the problem.

Other funding agencies across the world support programs with similar goals. Also, commercial products are starting to be available that enable meaningful interactions in, as of now, limited scenarios, using automatic translation devices. For example, *National Telephone and Telegraph* (NTT) DoCoMo of Japan in cooperation with ATR-Trek provides a speech translation system for the domain of frequent travellers' expressions. The system works on a common cell phone, providing voice-activated Japanese-English and Japanese-Chinese translation by utilizing a server operated by ATR-Trek.

Despite a long tradition of research, the problem of automatic STST is far from being solved and many research opportunities present themselves to the interested engineer and computer scientist. This, in combination with the existing funding and appreciation by the public, as well as the existence of commercial companies developing and marketing the technology, makes it an attractive field for advanced students or recent graduates seeking to pursue a career in either research or product development.

II. Projects and their Use Cases

The importance and livelihood of the research field of automatic speech-to-speech translation is mirrored by a multitude

Sebastian Stüker, Teresa Herrmann, Muntsin Kolss, Jan Niehues, and Matthias Wölfel are with the Fakultät für Informatik, Universität Karlsruhe (TH), Karlsruhe, Germany e-mail: {stueker,therrman,kolss,jniehues,wolfel}@ira.uka.de

of recent or ongoing, large scale research projects—a selection of which we would like to briefly present here.

The individual projects address different domains and application scenarios, often motivated by the needs of the agencies funding the projects. Funding agencies are usually national or international, government backed entities, military or civil in nature. Ongoing projects such as Global Autonomous Language Exploitation (GALE) or the recently started program Quaero offer excellent opportunities and environments for young engineers to become involved in cutting-edge research and advancing the state-of-the-art in a lively and technologically challenging area.

A. TC-STAR (http://www.tcstar.org)

From April 2004 until March 2007 the European Commission sponsored the project *Technology and Corpora for Speech to Speech Translation* (TC-STAR), an effort to advance research in all core technologies for STST—automatic speech recognition, spoken language translation, and speech synthesis. TC-STAR aimed at a breakthrough that significantly reduces the gap between human and machine translation performance. Participants to the project came from seven European countries. One of the participants also had departments in the United States which contributed to the project. The project targeted translation of unrestricted conversational speech on large and unconstrained domains of discourse. The main task chosen was the translation of speeches delivered in the European Parliament. Given the high costs of manual translations at the European Union, support for researching and developing automatic translation systems for this specific domain is a promising investment. In order to foster progress, periodic, competitive evaluations were conducted. TC-STAR was the first large scale program to target an unconstrained domain for translation.

B. GALE (http://www.darpa.mil/ipto/programs/gale/gale.asp)

On the military side, the United States' (US) *Defense Advanced Research Projects Agency* (DARPA) sponsored program *Global Autonomous Language Exploitation* (GALE). GALE develops technologies to absorb, analyze, and interpret huge volumes of speech and text in multiple languages. In the end it should be possible to deliver pertinent, consolidated information in easy-to-understand forms to military personnel and monolingual English-speaking analysts in response to direct or implicit requests. In the program, speech technology to recognize huge amounts of foreign speech (e.g. Chinese and Arabic) is developed as well as technologies to translate this information into English. At the end, a distillation engine is responsible for integrating information of interest to its user from multiple sources and documents. Military personnel will interact with the distillation engine via interfaces that will include various forms of human-machine dialog (not necessarily in natural language). The laboratories and companies participating in GALE are from the US, France, Germany, Great Britain, Hong Kong, and Switzerland.

C. Lecture Translator (http://interact.ira.uka.de)

Lectures, seminars, and oral presentations form a core part of knowledge dissemination and communication. For members of the audience that are native speakers of languages other than that of the presentation, the opportunities for collaboration and exchange of information are severely reduced. In some cases human translation services can be an option to overcome the language barrier, but high costs limit their utilization to some exceptional circumstances. In all other scenarios automatic translation systems can offer considerable practical benefits. Therefore, researchers at the *International Center for Advanced Communication Technologies* (InterACT)—a collaboration between Universität Karlsruhe (TH), Carnegie Mellon University, Hong Kong University of Science and Technology, and Waseda University—are developing systems for the simultaneous translation of lectures and speeches. An initial system translating from English to Spanish was presented to the public in 2005. Since then, work has been expanded to improve performance and include additional languages, such as German and Mandarin Chinese. Figure 1 shows the laptop based set-up of the translation system. In addition to displaying the translations as subtitles on a screen, it offers the possibility of projecting them into the field of vision of individual members of the audience by means of *goggles* with headmounted displays. Synthesized output can be directed at a sub-set of the audience using a *targeted audio* device that produces a narrow corridor of sound without disturbing the remaining listeners.

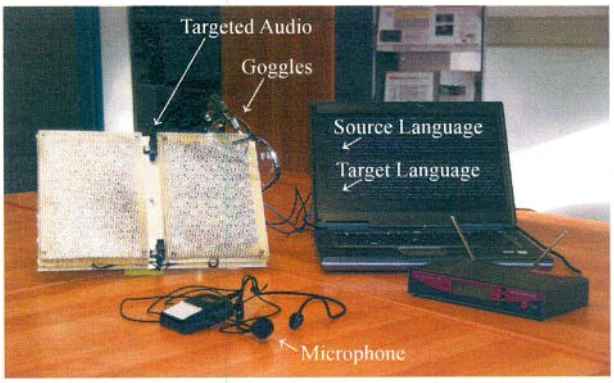

Fig. 1. Lecture translation system.

D. EuroMatrix and EuroMatrixPlus (http://euromatrix.net)

The principle of equality of languages within the EU leads to a growing need for translation and interpreting services, since EU documents and communication need to be available in all official languages. The EuroMatrix project was created to address this need for translating between the growing number of European languages. The project focused on statistical and hybrid machine translation to advance research in that field. One specific goal was to make translation systems available that cover all European language pairs. EuroMatrix was funded by the European Commission and included participants from

the Czech Republic, Germany, Hungary, Italy, and Scotland. Recently, EuroMatrixPlus was launched as a follow-up project continuing and building on the research efforts of EuroMatrix.

E. Quaero (http://www.quaero.org)

Quaero is a recently started French research and development program with German participation. It targets to develop multimedia and multilingual indexing and management tools for professional and general public applications such as the automatic analysis, classification, extraction, and exploitation of information. The projects within Quaero address five main application areas:

- Multimedia internet search
- Enhanced access services to audiovisual content on portals
- Personalized video selection and distribution
- Professional audiovisual asset management
- Digitalization and enrichment of library content, audiovisual cultural heritage, and scientific information.

Also included in Quaero is basic research in the technologies underlying these application areas, including automatic speech recognition, machine translation, and speech-to-speech translation. The vision of Quaero is to give the general public as well as professional user the technical means to access various information types and sources in digital form, that are available to everyone via personal computers, television, and handheld terminals, across languages.

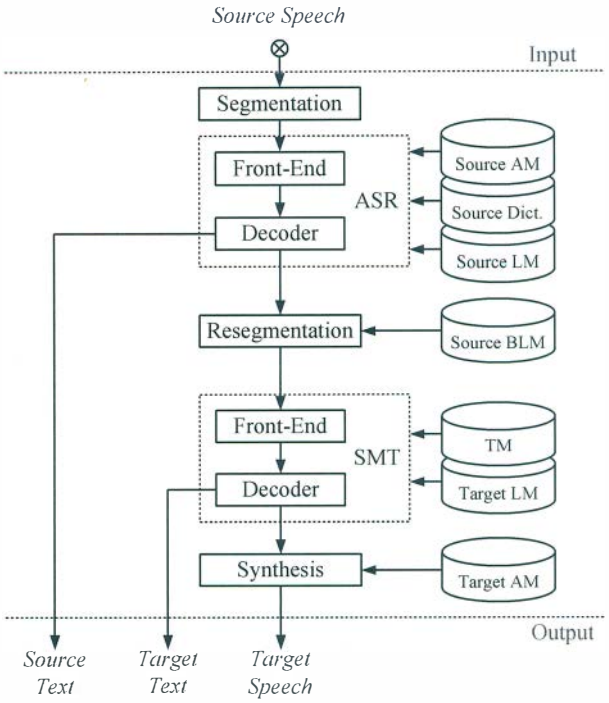

Fig. 2. The basic speech-to-speech translation chain with an automatic speech recognition (ASR) module and a, in this case statistical, machine translation module (SMT) at its core

III. System Overview

The basic layout of a current automatic speech-to-speech translation system is a chain of modules which pass their output to the next module in the chain. The first module performs a *segmentation* of the audio input. The resulting segments are then transcribed by the *automatic speech recognition* component. The resulting transcriptions are further processed by a *resegmentation* which combines and cuts the speech recognition result into segments that are then translated by the *machine translation* component. Finally, a *speech synthesis* module speaks the resulting translation. Figure 2 depicts a set-up using *statistical machine translation* and shows the different components that the individual modules use for their task. These modules and components will be described in more detail in the following sections. While segmentation, resegmentation, and speech synthesis can be seen as supporting techniques, automatic speech recognition and machine translation are fundamental to the operation.

A. Automatic Speech Recognition

The goal of *automatic speech recognition* (ASR) is to produce a written transcript of human speech. Due to the highly variable nature of human speech—two recordings of the same word sequence uttered by the same speaker under the same circumstances always will look different—current recognition systems apply statistical methods. For this, a speaker is recorded using a microphone. Under the assumption that a speech signal is stationary in a time frame of 10–20ms, the input from the microphone is preprocessed by applying short time frequency analysis. In this analysis the frequency components are extracted from 10 to 20ms long overlapping windows cut from the speech signal. Further processing discards unnecessary information and enhances the distinctive information in the signal, leading to features suitable for performing speech recognition. This chain of processing steps is called *front-end* or *pre-processing*.

ASR is a pattern classification problem. The features from the pre-processing are the pattern that is classified while the correct word sequence that corresponds to the audio recording is the class that the pattern needs to be assigned to.

The search for the class with the highest probability, given the extracted pattern, is the actual recognition process. For calculating the probability that a recording belongs to a certain word sequence, ASR makes use of two models. The *acoustic model* (AM) captures the relation between words and the recorded sound pattern, while the *language model* (LM) is concerned with how likely it is that certain word sequences are being uttered in the first place.

The acoustic models of current state-of-the art ASR systems for recognizing continuous, large vocabulary speech utilize *Hidden Markov Models* (HMMs). Words are being decomposed into phonemes using a pronunciation dictionary and phonemes are further divided into smaller units that serve as states in the HMMs. The states in the HMMs emit patterns, such as they are the result of the pre-processing. The emission probabilities are calculated by *Gaussian Mixture Models* (GMMs). The parameters of the models are trained on large

amounts—hundreds to thousands of hours—of transcribed recordings of speech from many speakers.

The language models used today are normally *N-Gram language models*. They calculate the probability of a longer word sequence by considering sequences of four to five words at a time. Their parameters are learned on large amounts of text data for the task that the ASR system is supposed to work on.

Using the acoustic model and the language model, graph algorithms in the *decoder* search for the most likely word sequence for a given recording. Using parameters, the size of the space of possible word sequences that is being searched can be influenced. The larger the search space is chosen the longer the run-time of the search, but the better the recognition result.

B. Machine Translation

The first systems for automating the human translation process, introduced in the late 40s of the last century, were *rule-based machine translation* (RBMT) systems that focus on language properties. With the increase in computational power in the early 1990s, an additional, data-driven approach towards machine translation emerged that is based on statistics—similar to ASR systems. This approach is referred to as *statistical machine translation* (SMT).

RBMT systems are built on manually created rules implementing linguistic knowledge. These rules capture grammatical structures and lexical information about source and target language words as well as crosslingual correspondences. The prototypical RBMT system is composed of a lexicon and a grammar. Those two components interact in the translation process which can be divided into three consecutive steps: Analysis, transfer, and generation. A grammar module performs the grammatical analysis of the source sentence and constructs an abstract representation of it. The transfer from the source language representation to target language representation is conducted by the lexicon according to bilingual correspondences. Finally, in the generation step, grammar rules determine how the abstract representation of the target sentence should be converted into a well-formed target language sentence.

The linguistic knowledge underlying the rules within RBMT systems ensure grammatically motivated translations. However, the input text must be well-formed in order for the analysis to succeed. Furthermore, to build such a system or merely to extend it with additional language pairs requires a language scientist who develops lexicon and grammar rules manually.

Therefore, SMT was introduced as a method that can automatically learn how to translate and is currently the most promising approach for machine translation in large vocabulary tasks. In this approach two probabilistic models, the *translation model* (TM) and the *target language model* (target LM), are used to find the most probable translation for the source sentence. The translation model is automatically learned from a large collection of parallel sentences that are translations of each other. In the beginnings of statistical MT, word-to-word translation probabilities were used that describe how probable the source word is translated into a target word. Nowadays, nearly all systems are phrase-based, modeling the probabilities of sequences of words instead of only single words. The target language model is used to generate fluent sentences in the target language by describing the probability of the target word sequence. One weak point of the statistical approach is the translation of grammatical constructions that do not have a direct correspondence in the target language. Furthermore, due to the lack of linguistic control mechanisms, the resulting target language sentence may be grammatically incorrect.

Since both MT paradigms have well-known weaknesses, hybrid MT approaches combining rule-based and statistical MT are currently under investigation.

C. Performance of ASR and SMT

ASR and SMT are not perfect technologies, but rather commit errors in their operations. For both technologies, performance measures exist that can automatically measure the quality of systems on a test set.

The quality of ASR systems is described by measuring its *word error rate* (WER) on a set of speech recordings whose correct transcription is known. WER describes how many percent of the words compared to the correct transcriptions have been wrongly recognized. It counts three types of errors: the substitution of a correct word with a wrong one, the omission of a word in the correct transcript, and the insertion of a word which is not in the correct transcript.

For SMT systems, the *bilingual evaluation understudy* (BLEU) score is one of the most popular performance measures. It is calculated by comparing reference translations against the translations found by the SMT on a given test set. BLEU calculates the precision of the SMT result in the reference translation. In doing so it considers word sequences of different lengths. So, it calculates, how often word sequences of one, two, three, etc. words in length that are present in the reference translation have been found by the SMT system. BLEU can also use multiple reference translations to take into account the fact that multiple correct translations for one source sentence can exist. Besides BLEU other metrics exist. For example, the *Translation Edit Rate* (TER) counts the number of modifications needed to transform the MT output into the reference translation.

Figure 3 plots the performance of the state-of-the-art in automatic speech recognition, evaluated in word error rate, and statistical machine translation, evaluated in BLEU score, over time. The WER and BLEU scores have been determined by annual evaluations hosted by the Multimodal Information Group (http://www.itl.nist.gov/iad/mig) at the *National Institute of Standards and Technology* (NIST) in order to measure and evaluate technologies that provide more effective access to multimedia and multilingual information.

Comparing the WERs and BLEU scores in the graphs, we can see that in general the performance of the systems on the same tasks improve over time. The decreases in performance are either due to a switch to a more difficult task (e.g. it is

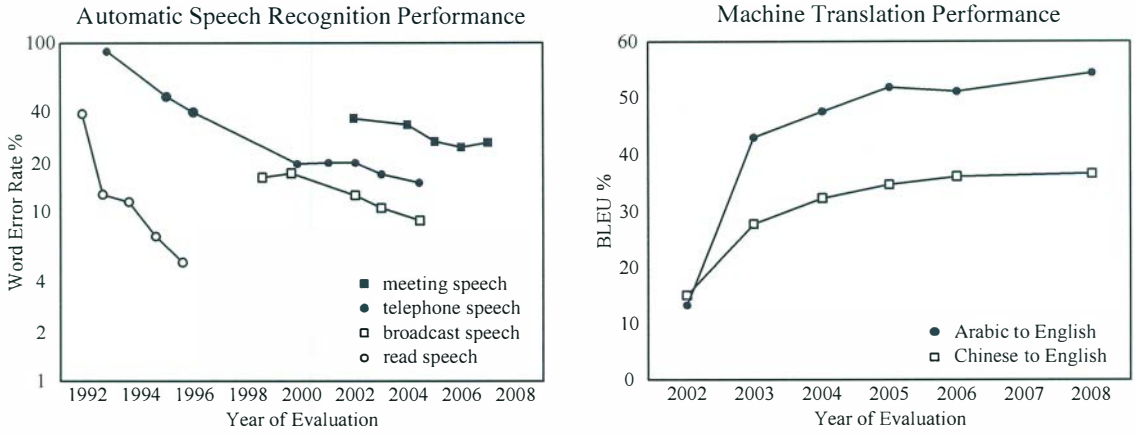

Fig. 3. Performance of the best available speech recognition systems in terms of word error rate and the best machine translation systems in terms of BLEU score over time as determined by evaluations hosted by NIST

more difficult to recognize meetings than read speech) or to a more difficult test set within the same task.

IV. Research Topics

From the use cases in the research projects above, it becomes clear that a large number of interesting research questions in the fields of ASR, MT, and their fusion into STST systems exists. In the following we have collected a selection of research topics that we believe to be relevant. For researchers or young engineers interested in pursuing a career in the area of STST, these listed research topics present an opportunity for entering this area and for making an impact on the development of future STST systems.

A. ASR

1) Language Diversity: Roughly 7,000 languages exist in the world today. For only a tiny fraction of those languages, automatic speech recognition systems have been created so far. Languages addressed are mainly those with either a large number of speakers, high economic power, or high political impact. As already described before, it is desirable to address all languages in the world in order to keep up a high cultural diversity, enable access to information for and from all people in the world, and avoid technological discrimination of certain languages. The traditional methods for training speech recognition systems are costly in terms of time and money, and it is clear that with them it will not be possible to address all languages. Therefore, one current area of research deals with developing new techniques that allow to rapidly and cost efficiently create speech recognition systems for new languages. Techniques being deployed often leverage from applying the knowledge and models gained from already studied languages in an automatic or semi-automatic way to the new languages, and refining them with as little effort and training data as possible. However, systems generated that way still lack in performance compared to systems trained with a high effort in the traditional way. Here, opportunities exist for refining existing and developing new methods, so that well performing speech recognition systems can be created in a way in which it will eventually be possible to address a by magnitudes larger number of languages than today.

2) Robustness: In a realistic speech recognition scenario one will always be confronted with interfering signals or significant background noises. While body mounted microphones, if mounted correctly, clearly lead to recordings with the lowest distortions and thus lowest word error rate, they are inconvenient and may generate too much distraction to the speaker. Thus, ASR which works reasonably well on recordings captured with mid- or far-field microphones is essential to providing convenient and robust automatic transcriptions. With the rule of thumb that *each doubling of the distance brings down the sound pressure level by approximately* 6 *dB* it is clear that a significant loss in signal quality arises by moving the microphone away from the speaker's mouth which is directly transferred into a lower transcription quality. To reduce the quality gap of automatic transcriptions between close and distant speech capture, technologies need to be developed to be able to estimate, track, and compensate for distortions introduced by *environment noise* and *reverberation*. Besides acoustic model adaptation, hot research topics to improve robustness of automatic transcriptions include speech feature extraction methods which are based on spectral envelopes and/or neural networks and speech feature enhancement methods. Enhancement techniques can, for example, compensate for:

- non-stationary additive distortions by Bayesian filters,
- convolutive distortions by techniques based on correlation such as multi-step linear prediction and
- both kinds of distortions by particle filters which integrate reverberation estimates or by microphone array processing techniques such as blind source separation and beamforming.

3) Higher Level Knowledge: When humans perform speech recognition they often rely on higher order knowledge such as world knowledge, common sense, or knowledge about the semantic context. Especially in adverse acoustic conditions, this kind of knowledge helps us humans to correct misrecog-

nized words or completing only partially recognized sentences, by either guessing or performing plausibility checks on the recognized speech. ASR systems, as they exist today, do not have the capability to apply this kind of world knowledge. On the linguistic side, the only source of knowledge for them is the language model which usually keeps statistics of which word sequences of four or less words are likely to occur. If ASR technology wants to perform equally well as us humans or even surpass us, it has to find a way to include this kind of knowledge in its recognition process. Currently no models for representing what one might call common sense exist in science. Also, it is not clear how this kind of knowledge has to be integrated into the current recognition framework. In order to achieve significant advances in this difficulty, a scientific revolution in ASR research has to abolish the current statistical framework and to establish a paradigm to enable such sanity checks at ASR algorithms level.

B. MT

1) Word Reordering: One important and still not satisfactorily solved problem in SMT is the word reordering problem. Even in related languages like German and English the word order is very different. A difficult example is the position of the verb: in a German subordinate clause the verb is at the last position, whereas it is at the second position in the English sentence. Furthermore, German verbs might consist of two parts where the auxiliary verb is at the second position of the sentence while the main part is at the end. Consequently, reordering over a rather limited distance does not lead to a reasonable translation quality, as the readability of the translation is strongly affected by the awkward word order. One promising way to address this problem is the use of additional linguistic knowledge.

2) Hybrid Approach: The strength of SMT is the ability to automatically learn translation with sentence context being considered. On the other hand, SMT is not guaranteed to produce grammatically correct sentences. Contrary to that, the linguistic knowledge implemented in rule-based MT systems ensures well-formed target language output.

As a result of this contrariety of the two paradigms, hybrid approaches combining SMT and RBMT receive growing interest. They aim at the combination and mutual compensation of the aforementioned strengths and weaknesses. Two main types of hybrid MT systems can be distinguished: integrated systems and multi-engine systems. Integrated systems interleave statistical and rule-based methods. In order to do so, either a hybrid system has to be built from scratch applying linguistic rules and statistics as required for the different subtasks. Or the internals of existing rule-based and statistical systems have to be broken up and original system parts are rearranged. Contrary to that, multi-engine systems use existing MT systems as black boxes to obtain multiple translation hypotheses for the same input. Then, selection mechanisms identify the best hypothesis of the set. Alternatively, the sentence parts of the different system outputs are recombined to obtain a new, possibly better hypothesis. First translation results from hybrid systems of both types have shown an improvement of translation quality over RBMT and SMT.

3) Language Coverage and Bridge Language: In recent years, the number of languages used in machine translation systems increased. While parallel corpora are generally available for large common languages, it is rare to find large parallel corpora for more unusual language pairs (e.g. Paschtu-Catalan) and domains. Furthermore, the number of translation systems needed grows quadratically with the number of language pairs.

An alternative strategy, in contrast to building translation systems for every language pair, is the use of a bridge language. This could be English since for many language pairs there exists a parallel corpus that includes English. On the downside, information gets lost by first translating to English. Future research will have to focus on compensating for that, e.g. by building on past approaches in enriching the English used as pivot language.

4) Comparable Corpora: In statistical machine translation systems, parallel corpora at sentence level are the most important knowledge source. While there are plenty of sources for monolingual corpora, only very few parallel corpora are available. For example, for many European language pairs the only big parallel corpus is the collection of plenary speeches from the European Parliament. Since the topic of this corpus is restricted to political issues, a system trained on this corpus will not be able to deliver good translations for other domains. Thus, additional knowledge sources need to be found in order to broaden the domain covered by such translation systems.

One solution to this problem is the use of comparable data. These corpora do not consist of direct translations of each other, but they are about the same topic. Although only parts of the text can be used to train a statistical translation system, they can be found more easily and far more comparable data exists, for example on the Internet, than parallel data.

5) Evaluation in MT: Unlike as in ASR, the automatic evaluation of an MT system is a research topic on its own. The main problem is that there is not one single correct translation for most of the sentences. Instead, many different variants of a correct translation are conceivable. For example, individual words can have more than one correct translation and the target sentence may be realized by using alternative correct word orders. To account for that problem several reference translation and a bigger amount of test sentences are used.

BLEU and TER perform quite well when evaluating statistical machine translation approaches. However, the scores obtained by these evaluation measures do not correlate well with human judgements on translation quality when evaluating rule-based MT approaches or hybrid ones. Therefore, more advanced methods to compare the translation output with the reference translation have been investigated in recent years. These try, for example, to take also synonyms and the morphology of the target sentence into account. Within the annual *Workshop on Machine Translation* (WMT), one of the tasks is dedicated to automatic evaluation metrics and the assessment of their relation to human judgements.

C. STST

1) Segmentation and Tight Coupling: Machine translation systems need a certain minimum context and perform best

when given more or less well-formed sentence or utterance units to translate. The standard approach to coupling speech recognition and machine translation components is to segment the ASR output prior to passing it to the translation component. The segmentation operates on the unstructured first-best ASR hypothesis and divides it into shorter, sentence-like units. Choosing good segment boundaries, ideally corresponding to semantic or strong syntactic boundaries, can have a big impact on the final translation performance. Some of the most useful features for segmentation are a source language model, pause information from the original audio signal, and observing reordering and phrase boundaries within the translation system. By combining them, meaningful segment boundaries can be identified.

2) ASR with automatically found word units: In STST systems it is often not necessary to have the exact written representation of the spoken words that need to be translated as an intermediate result. It is sufficient when the result of the ASR in the processing chain is suitable for the translation component to produce the desired output. Recent research has started to explore the possibility of automatically finding word-like units in new, previously untreated languages. This can be very helpful when trying to develop a translation system for a new language, about which is little known by the system developers and for which no expert might be readily available. This scenario addresses specifically the multitude of languages, often under-resourced and less prevalent, for which no writing systems exist. The exploration of words in those languages can be aided by learning in the field, e.g. from human simultaneous translators which can be observed by the system.

3) Run-on Decoding - Low Latency: Systems for simultaneous translation of speeches or lectures must run in real-time in order to keep up with the lecturer. In addition, the system should have a low overall *latency*. That is, the time difference between the speech in the original language and the translation should be as short as possible. Long delays between the original speech and the translation might otherwise severely impact the understanding of the presented topic, for example when the speaker is referring to text or visual information presented on slides. The main bottleneck to achieving shorter latencies in current STST systems is the machine translation component. This is due to MT's need for a long word context and the need for word reordering. The traditional approach of segmenting the ASR output into fixed segments prior to translation, introduces additional errors and delays.

An alternative approach is directly processing the continuous input word stream from the speech recognizer and decoupling decisions of when to generate partial translation output from a fixed input segmentation. Using a flexible sliding window within which full word reordering is possible, the same translation performance is reached at greatly reduced latency values. This stream decoding approach works well when translating between languages with similar word order, but the long-distance word reordering requirements of language pairs such as German-English interfere with the desire for short latency. The challenge here is to find appropriate models that selectively extend the search horizon only when necessary, in a manner that mimics the strategies of professionally trained human interpreters.

V. Conclusion

The field of speech-to-speech translation is a well established area of research that has a noteworthy tradition. It addresses a problem—the communication between people and the access to information across languages—that is of high relevance in today's globalized world. By contributing to the field, recent graduates in Electrical Engineering and Computer Science can make significant contributions to a challenging, technical field. At the same time their work has a highly valuable impact on human interaction, and is advancing society on its way to becoming a multilingual world-wide community.

The plentitude and diversity of current research projects depicted in this article demonstrates that this field of research is a very active one. The ongoing projects receive significant funding and attention by the public, and are backed and initiated by governments world-wide. As shown, the activities in research communities happen at an international scale and projects can be found in many countries around the globe. The combination of these projects and the current and future research topics that we have outlined give graduates an ideal environment for performing challenging and interesting research. By pursuing a career in this field of research, young engineers and scientists are given the opportunity to make an impact on the future development of this area by advancing the state-of-the-art of automatic speech-to-speech translation systems.

References

[1] V. Steinbiss, "Human language technologies for europe," Work comissioned by ITC-irst, Trento, Italy to Accipio Consulting, Aachen, Germany, April 2006.
[2] "A new framework strategy for multilingualism," Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, November 2005.
[3] R. G. Gordon, Junior, Ed., *Ethnologue, Languages of the World*, fifteenth ed. SIL International, 2005.
[4] T. Janson, *Speak – A Short History of Languages*. Oxford University Press, 2002.
[5] A. Waibel and C. Fügen, "Spoken language translation—enabling cross-lingual human-human communication," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 70–79, May 2008.
[6] "How to build a bable fish," *The Economist Technology Quaterly*, p. 20, June 10th 2006.
[7] P. Koehn, *Statistical Machine Transation*. Cambridge University Press, 2009.
[8] J. Hutchins, "Iamt compendium of translation software," http://www.hutchinsweb.me.uk/Compendium.htm.
[9] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Prentice Hall, 2001.
[10] "Statistical machine translation," http:\\www.statmt.org\.
[11] D. Nettle and S. Romaine, *Vanishing Voices*. New York, NY, USA: Oxford University Press Inc., 2000.