Speech Translation: Past, Present and Future

Alex Waibel

Interactive Systems Laboratories
Carnegie Mellon University, USA & University of Karlsruhe, Germany
ahw+@cs.cmu.edu

Abstract

A decade after its first beginnings, the grand challenge of building automatic systems that translate speech has grown into an active research area, a focus for speech and language researchers worldwide. Workshops, conferences, sessions, journal issues are devoted to it, and research is supported by governments in Asia, Europe and the US. The problem has attracted much attention, as practical needs in an increasingly globalized world converge with scientific advances that bring possible solutions within reach. While great progress has been made, the problem is certainly not solved and much remains to be done. At a midway point along the way, we present this paper as a review of the past, of the successes so far, and as an attempt to chart a course for the future.

1. Introduction

Over the years of our fascination with the problem, we were often asked why one should work on speech translation (ST), when speech recognition, translation and synthesis are all still research areas in their own right.

The answer is both pragmatic and scientific: Speech Translation, if realized, would have tremendous impact in a world, in which overcoming the linguistic divide emerges as the far greater challenge than the digital divide. It only takes watching the daily news, to convince us, that overcoming the separation between people by language and culture is the most dramatic challenge in society today. Yet, unifying languages is not the solution, as languages represent the cultural heritage and identity of people and hence must and will be maintained. Greater globalization will continue a trend of linguistic consolidation that will lead, regrettably, to the disappearance of many languages. However, cultural heritage, national identity and the size of sufficiently large language groups present opposing forces that will ensure that a large number of languages remain; too large a number for any one person to learn. Is our future therefore one of greater understanding and inclusion at a loss of cultural diversity? Or is it one of greater diversity at the expense of mutual understanding and social and economic inclusion. Speech translation provides a hope that we can address this dilemma.

Scientifically, the problem of speech translation is, of course, the ambiguity at all levels of the speech and language chain: each component needed for a speech translation system is imperfect and cannot produce an error-free output. et, we have learned in speech recognition (after all we can still not recognize phonemes!) and other fields that it is indeed possible to build complex large systems that achieve impressive performance, from poor and imperfect components. Key to the success is the use of multiple complementary knowledge sources and careful statistical modeling of a hypothesis space on which those knowledge sources operate. In speech recognition this means producing, searching, pruning and re-searching graphs of phones, words or concepts. This can be done in speech translation as well, e.g., by successive application of recognition, translation and generation, while managing ambiguity along the way. This observation encourages us to move toward even more ambitious speech translation problems.

2. The Past and Present

An impressive array of speech translation solutions has already been proposed. We will review a few example solutions in the following. Beyond detailed differences between individual system solutions, however, one can categorize speech translation systems by a few general system level capabilities that have or have not yet been addressed. These capability levels are shown in Table 1.

The distinguishing capabilities are generally related to the levels of uncertainty that each of the technology can tolerate. The more restrictions one imposes on the use and operation of a system, the greater constraints can be used to limit the search for translation hypotheses. In other words, better performance can be obtained, if one can impose greater restrictions on the incoming speech.

The very first speech translation systems [1], [2] and [3] were feasibility demonstrators that showed the concept of speech translation and that speech translation was possible at all. They attracted a great deal of attention, as they showed that bridging the language divide by spoken language might indeed be possible [4]. These early systems did not permit free dialog and required speakers to act out prescribed sentence patterns or allowable sentences in a read-speech speaking style according to a restricted syntax and/or a restricted vocabulary.

Table 1. Speech Translation System Capabilities

	Vocabulary	**Style**	**Domain**	**Platform**	**Example Systems**
Feasibility Exponents	Restricted	Constrained	Limited	Workstation	C-STAR I
Phrasebooks	Restricted	Constrained	Limited	Handheld	Phraselator, Ectaco
Two-way Dialog Systems	Unrestricted	Spontaneous	Limited	PC/Handheld	JANUS-III, ATR-MATRIX, C-STAR-III, Verbmobil, Nespole, Babylon, …
Simultaneous Translation	Unrestricted	Conversational	Unlimited	PC	---

Nevertheless they were systems that were proposed at a time, when speaker-independent, continuous speech was still a novelty and Machine Translation (MT) considered close to impossible, and thus, represented pioneering early work, the beginning of a new research endeavor.

Already in 1992, however, it was recognized that these early concept demonstrators fall short of usable systems, as speakers had to speak in a well-behaved manner, and remember the words and sentences they would be allowed to say. The most unacceptable constraints were the vocabulary, syntax and speaking-style limitation. It was observed, that it is generally possible for humans to stick to a domain of discourse; indeed, many important applications are domain limited: hotel bookings, car rental, taxi and shopping negotiations, medical, emergency relief, Hotel/Hospital/ Conference registration, and many more, all require only dialogs in a limited domain. By contrast it is generally not possible for humans to speak in a limited speaking style (effectively reading sentences), or remember a limited set of words or syntactic patterns. Two solutions have been proposed:

The first is a pragmatic approach to addressing language needs in mobile situations (tourists, field workers, etc. Fig. 1), a mobile phrase book, such as the Phraselator [5] (Fig. 2). This approach does not address the problem of speaker style, but relaxes vocabulary restrictions and provide speakable phrases on a hand-held device. Sometimes called a "one-way" as it does not allow for free dialogs between two conversants (this requires spontaneous speech), but permits speech entry of a list of useful phrases for a given situation. Limitations in syntax are addressed by the user interface, which proposes alternatives that may still allow a speaker to make him/herself understood in a given situation.

Figure 1. Janus/Lingwear

Figure 2. Phraselator

Unfortunately, such systems are only practical in "one-way" mode, allowing one party to express a certain concept, but cannot process a free, spontaneous response. Users of one-way phrasebooks have to resort to gestures, pointing or other means of communication in this case.

More advanced technology was therefore necessary to address the serious limitation of speaking style: two-way dialogs almost always require handling free spontaneous speech input, both in recognition and translation. For spontaneous dialogs, in turn, we must relax syntactic constraints, as spontaneous speech is ill-formed and fragmentary. Without syntactic constraints in the translation process, we must turn to semantic constraints for guidance. Fortunately, this is possible in a limited domain application, where the typical concepts and arguments one wishes to convey, can be enumerated and represented. Many two-way speech translators have therefore used a semantic representation (or interlingua) as intermediate step, based on which a sentence could be produced in the target language. Semantic parsers or classifiers extract key concepts and arguments from an ill-formed input string and map them onto the interlingua [6], [7] and [8].

3. Where do we go from here?

From the discussion in the previous sections, several research directions emerge that chart our course for the future. Clearly, further improvements can be achieved by continued development of the present technology. Improvements in grammar coverage, use of tree banks, modularization and integration of multiple domains, faster implementations and smaller footprint, etc. will all lead to greater usefulness of these advanced speech translation systems to different uses, additional domains and languages.

However, two problems appear to be particularly daunting:

- The problem of portability: How much effort and cost is necessary to develop speech translation technology for a new domain and new language. For most of the 4,000-6,000 languages (many of which from poor and small communities), the present cost of development would be prohibitive.
- The problem of domain limitation: What exactly would it take to go beyond current domain limited spoken language systems and to realize domain unlimited translation of conversational speech. Many important applications that require such capability exist and include: translating telephones, translating TV's, Radio's, and simultaneous lecture and meeting translators. The challenge is formidable: for such tasks we can neither rely on the syntactic structure of the input, nor fully describe the semantics of the domain.

In response to these challenges, we believe that greater reliance on machine learning is key to making more adaptable, portable and ultimately more extendable speech translation systems [9] [10]. In the US and Europe, we have begun work under three new initiatives to address these problems: Projects STR-DUST (NSF) and TC-STAR (EC-FP-6) [11] are aiming at the latter and Babylon/Cast (DARPA) at the former. In the following sections we describe four ideas and techniques that we are developing as possible solutions in our research. We will illustrate these concepts by recent pilot experiments at our own laboratories.

3.1. ST by Direct Statistical Translation

Statistical Machine Translation (SMT) has been used in limited domain translation tasks in parallel to the development of Interlingua-based translation systems. For the Nespole project, a dialog system for the tourists [12], only a very small domain specific bilingual corpus of about 3,500 sentence pairs German-English was available at CMU to train a statistical lexicon and also to extract phrase-to-phrase translation. However, one of the advantages of using a statistical translation system is that electronic dictionaries or other available data can easily be added. The English part of the Verbmobil corpus (collected at CMU), which is also conversational, was added to the data used for training the language model. When compared with the Interlingua-based system our statistical MT system achieved higher translation quality, both in terms of an automatic evaluation metric and in terms of human evaluations [13]. Results are given in the

following table, where we show both the automatically computed Bleu scores and human judgments (good, acceptable, bad) as measures of quality.

Table 2. SMT vs. Interlingua-based system on ST Tasks

	System	Bleu	Good	Okay	Bad
Text	IF	0.068	77	104	227
	SMT	0.333	124	80	205
Speech	IF	0.059	64	101	243
	SMT	0.262	95	83	227

The results demonstrate that statistical machine translation is comparable or even better in performance to grammar-based translation systems and that this is already the case when only very small corpora are available for training. This is explainable in part by the greater speed and ease of training statistical translation system compared to manually programming grammars. Coupling between the recognizer and the translation engine can also be achieved by performing the translation search on output word lattices, thereby exploiting the availability of alternative recognition hypotheses for translation.

3.2. ST by Statistical Interlingua Based Translation

A direct statistical translation approach clearly has the advantage, that it eliminates the complexities and cost of building ontologies, programming large grammars, or tagging large tree banks. On the other hand it does deprive us of three benefits that an interlingua has meant for us: 1) an elegant solution for connecting N language systems to each other without the cost of building N^2 translators, 2) providing an attractive way for feedback by paraphrasing the input in one's own language, and 3) the possibility of using the extracted semantics of an utterance to achieve other additional side-effects (updating a database, filling in forms, generating minutes and summaries, etc.) besides the translation itself. For all these reasons, it is interesting to consider applying statistical modeling in the development of statistical interlingua based machine translation systems. Here, a spontaneous input sentence is 'translated' into the interlingua as output language, and an interlingua is 'translated' into an output sentence.

A first system in this spirit was proposed by Kauers et al. [14]. The statistical translation framework was reformulated to allow for an interlingua as output languages. To do so, the translator must map a sentence into a tree instead of an output string, and must represent constituents in an order-invariant manner, since an interlingua is a tree-structured representation of concepts and arguments.

An alternate design philosophy is proposed by Reichert et al. [15]. Here, we use English as the "interlingua". The advantage of doing so is that it is simple and straightforward to "design" an interlingua, since it is a plain English translation of the input. A potential drawback is that English is a natural language, i.e., inherently ambiguous. Translating from one language into English and from there to another language will add errors as a result. Also English does not allow for direct links to databases, forms, etc., as a side effect, without another translation step. Nevertheless, English text can be enriched by various linguistic tokens and tags that improve specificity and translation quality. While such linguistic tags cannot be obtained for any language, they can certainly be obtained for English. Performing a variety of processing steps (tagging, compressing, simplifying, disambiguating, etc.) can be performed for English where a rich body of tools, language resources and know-how exist. The results on translation suggest that MT systems can be successfully constructed for any language pair by cascading multiple MT systems via English. Moreover, end-to-end performance can be improved, if the interlingua language is enriched with additional linguistic information that can be derived automatically and monolingually in a data-driven fashion.

3.3. ST by Learning of Semantic Grammars

As an alternative to direct and interlingua-based statistical translation methods, it is also possible to scale grammar based approaches and obtain greater coverage at lower cost, as long as we can learn these grammars automatically, as opposed to programming them. This can be done in two ways:

- Training from Tree-Banks: a training corpus of example sentences in the target domain is tagged according to semantic labels, and grammars are trained using the resulting tree-bank. Various statistical and connectionist approaches have been applied successfully to achieve this for spoken language understanding as well as translation tasks [16], [17], but they require semantic tagging of a large corpus to be effective.
- Interactive Training of Grammars: the underlying assumption for this approach is that it is acceptable to build a grammar as long as it occurs implicitly during use and does not require special expertise. Work by Gavalda [18] has shown, that grammars can expand interactively from basic seed grammars, but assuming the most likely meaning of a sentence and asking confirmation questions when in doubt. Experiments with naïve users show, that grammars can be built in this manner interactively without requiring expertise. The resulting grammars generalized well toward new and unseen data.

3.4. ST in Domain Unlimited Spontaneous Tasks

Unlike a formal written language, conversational spoken language is much less well-formed and contains various kinds of disfluencies, such as filled pauses, repetitions, corrections, false starts, etc. Statistical analysis of the Chinese Callhome corpus (120 dialogues) showed that 15~20% of words are speech disfluencies. These disfluencies cause more difficulties for speech translation because they not only introduce unwanted uncertainties in the speech recognition and translation hypothesis, but also decrease the readability of the translation. Some initial experiments show that after manually removing the disfluencies from manual speech transcripts, the translation Bleu score of a translation hypothesis (which compares n-gram matches between automatic translation and reference translation, here n=1,2,3,4) is increased by 0.02 when using a statistical word-based translation system.

The following example shows the automatic translation of original manual transcriptions ("Orig") and disfluency-removed ("Clean") transcriptions (the disfluencies are marked with italic font and underscored), compared with the reference translation ("Ref"):

Orig: *这个* 美国, *这个* 教学 进度 比较 快 ,

Translation: *this* usa, *this* education progress relatively quick,

Clean: 美国 教学 进度 比较 快,

Translation: the american education progress relatively quick,

Ref: the education progress in america is faster

Honal et al. [19] proposed an approach for disfluency cleaning based on statistical machine translation framework. In this framework, the original disfluent speech was regarded as the "source language" while the cleaned speech was regarded as the "target language". The cleaning process was treated as a monotone translation. Several translation models were trained on the VerbMobil English corpus, where disfluencies have been manually annotated. Evaluation on the held-out data showed 77.2% recall and 90.2% precision. This approach was also applied to the Chinese Callhome corpus, and achieved 49.4% recall and 78.8% precision.

In conclusion, a combined approach of recognition, conversion into a more textual form, and translation, appears to have merit for further pursuit.

4. Conclusion

We have reviewed the state of affairs in speech translation research. After a decade of tremendous advances, the field is more active than ever before. A variety of successful research systems as well as commercial systems have emerged, that begin to provide practical language assistance. Considerable challenges remain, however, as we must improve language portability and remove domain limitations in our systems. Domain unlimited translation of conversational speech (e.g., telephone conversations, meetings, lectures) represents the grand challenge for the decade ahead.

Acknowledgements

I would like to thank Fei Huang, Stephan Vogel and Ying Zhang for their help during the writing of this paper.

References

[1] Alex Waibel, Ajay N. Jain, Arthur E. McNair, Hiroaki Saito, Alexander G. Hauptmann, and Joe Tebelskis. “JANUS: A Speech-to-speech Translation Using Connectionist and Symbolic Processing Strategies.” In Proc. of ICASSP'91, pages 793--796, May 1991.

[2] Louise Osterholtz, Charles Augustine, Arthur McNair, Ivica Rogina, Hiroaki Saito, Tilo Sloboda, Joe Tebelskis, Alex Waibel, “Testing Generality in JANUS: A Multi-lingual Speech Translation System,” ICASSP, 1992, volume 1, pages 1-1.

[3] T.Morimoto, T.Takezawa, F.Yato, S.Sagayama, T.Tashiro, M.Nagata, and A.Kurematsu, "ATR's speech translation system: ASURA," Proc. 3rd European Conf. on Speech Communication and Technology, pp. 1291--1294, Sep. 1993.

[4] Andrew Pollack, "Computer Translator Phones Try to Compensate for Babel", New York Times, Jan 28, 1993, page A1.

[5] Sarich, A., Phraselator, one-way speech translation system, http://www.sarich.com/translator/, 2001.

[6] M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P.Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki-Waibel, A. Waibel, W. Ward, “Recent Advances in JANUS: a Speech Translation System,” In Proceedings of Eurospeech, 1993.

[7] Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavaldà, Torsten Zeppenfeld, Puming Zhan. "JANUS-III: Speech-to-Speech Translation in Multiple Languages," In Proceedings of ICASSP-97, Munich, Germany, April 1997.

[8] Rayner, M., Alshawi, H., Bretan, I., Carter, D.M., Digalakis, V., Gamback, B., Kaja, J., Karlgren, J., Lyberg, B., Price, P., Pulman, S. and Samuelsson, “A Speech to Speech Translation System Built From Standard Components,” Proceedings of the Workshop on Human Language Technology, 1993.

[9] S. Yamamoto, "Toward speech communications beyond language barrier - research of spoken language translation technologies at ATR", ICSLP2000, 2000-10

[10] Y Yuqing Gao, B. Zhou, Z. Diao, J. Sorensen, H. Erdogan, R. Sarikaya, F. Liu, M. Picheny, "A Trainable Approach for Multi-Lingual Speech-To-Speech Translation System ", Human Language Technology Conference 2002, San Diego, CA., March 2002

[11] Technology and Corpora for Speech to Speech Translation (TC-STAR), http://www.tc-star.org/

[12] F. Metze, C. Langley, A. Lavie, J. McDonough, H. Soltau, A. Waibel, S. Burger, K. Laskowski, L. Levin, T. Schultz, F. Pianesi, R. Cattoni, G. Lazzari, N. Mana, E. Pianta, L. Besacier, H. Blanchon, D. Vaufreydaz, and L. Taddei. "The NESPOLE! Speech-to-Speech Translation System," HLT 2002, San Diego, California U.S, March 2002.

[13] Stephan Vogel, Alicia Tribble. "Improving Statistical Machine Translation for a Speech-to-Speech Translation Task." In Proceedings of ICSLP-2002 Workshop on Speech-to-Speech Translation. Denver, CO. September 2002.

[14] Manuel Kauers, Stephan Vogel, Christian Fügen, and Alex Waibel, “Interlingua Based Statistical Machine Translation,” In Proceedings of ICSLP 2002, San Diego, USA.

[15] Alex Waibel, Tanja Schultz, Stephan Vogel, Christian Fügen, Matthias Honal, Muntsin Kolss, Jürgen Reichert, Sebastian Stüker, “Towards Language Portability in Statistical Speech Translation,” to appear in the Proceedings of International Conference on Acoustics, Speech, and Signal Processing. Montreal, Quebec, Canada. May, 2004.

[16] W. Minker, "Stochastically-Based Natural Language Understanding Across Tasks and Languages," Proc. Eurospeech, pp. 1423—1426, September 1997.

[17] Finn Dag Buoe, Alex Waibe, “Learning to parse spontaneous speech,” In Proceedings of the International Conference on Spoken Language Processing, ICSLP 96, Philadelphia, PA.

[18] Marsal Gavalda, "Interactive Grammar Repair," Proceedings of the ESSLLI-98 Workshop on AUTOMATED ACQUISITION OF SYNTAX AND PARSING, August 24 - 28, 1998

[19] Matthias Honal and Tanja Schultz, “Correction of Disfluencies in Spontaneous Speech using a Noisy-Channel Approach,” Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech-2003), Geneva, September 2003.