Proceedings of the Workshop on Speech-to-Speech Translation:
Algorithms and Systems, Philadelphia, July 2002, pp. 129-135.
Association for Computational Linguistics.

The VI framework program in Europe:

some thoughts about Speech to Speech Translation research.

Gianni Lazzari
Centro per la ricerca scientifica e tecnologica ITC-irst
Via Sommarive 18 38050 Povo Trento
lazzari@itc.it

Abstract

Significant progress has been made in the field of human language technologies. Various tasks like continuous speech recognition for large vocabulary, speaker and language identification, spoken information inquiry, information extraction and cross-language retrieval in restricted domains are today feasible and different prototypes and systems are running. The spoken translation problem on the other hand is still a significant challenge: "Good text translation was hard enough to pull off. Speech to speech MT was beyond going to the Moon – it was Mars…" [Steve Silbermann, Wired Magazine].

Considering the major achievements of the last years obtained in the field and the related challenges, a question arise: what next ? Is it possible to foresee in the next decade real services and applications ? How can we reach this goal ? Shall we rethink the approach ? Shall we need much more critical mass ? How about data ? To answer to these questions a new preparatory action, TC_STAR_P, funded in the V framework, has been settled in Europe. Goals, objective and activities of this preparatory action will also be discussed in this paper

1 Introduction

In the last ten years, many projects addressed the speech to speech translation problem, S2ST, i.e. VERBMOBIL [1], C-STAR [2], NESPOLE! [3], EU-TRANS [4], BABYLON [5], .. Many results and advancements have been achieved in methodology, approaches and even performance. These projects have shown prototypes and demonstrations in different communicative situations: speech to speech translation over the telephone, machine mediated translation in a face to face communication (both in a real face to face or through videoconferencing). Some basic approaches have been explored: direct translation or data driven (both example based and statistical), indirect translation trough interlingua-interchange format (IF) and mixed approaches, i.e. multiengine. In terms of performance significant results have been obtained in the VERBMOBIL project using a statistical approach.

Real applications using ASR technology are used in many applications in every day life [6]. Dictation machines in limited domain, simple automatic services over telephone, command and control in car, spoken document retrieval from broadcast news. Despite the new economy bubble and some dramatic events, like the L&H case, speech companies are still on the market. However in terms of technology employed, we are far from providing a free communication functionality which is necessary when more complex automatic services are needed, even considering communicative situations where a small number of concepts are involved (very limited domain). Automatic time

table inquiry systems are working in a strictly menu driven approach. Automatic directory assistance services can also be classified in this class of applications. Here a further complexity is given by the high perplexity of the directory names, but in the end is still a complex communicative situation. In fact consider the difficulty in modelling the high number of sentences that can be used when trying to get the telephone number of an item of the Yellow Pages.

The microelectronic and telecommunication market offers new opportunity of communication by cells phones, pdas, laptops in a wired or wireless environment. The communication process in this case is helped or "complicated" by multimodal interfaces and multimedia information. A new framework could be offered by the Web, which "integrates" potentially multimedia data with multimodal communication. In this case the paradigm is shifted towards a multimedia, multimodal person to person communication, in which the meanings are conveyed by the language and enhanced with multimedia content and non verbal cues. The answer to a given question in a multilingual conversation could be more effective if given in text and/or visual form. In this case the problem to afford becomes a combination of language understanding, information extraction and multimedia generation in the target language. Document retrieval, summarization and translation could also be involved in this communication process. All these technologies should be thought as pieces of a whole: a new model for person to person, information mediated, communication that brings together all of the resources available: verbal and non verbal communication, multimedia, face to face. Approaching the multilingual communication as a whole means to implement each new technology as a brick within an entire edifice.

Starting from the state of the art in speech to speech translation research, considering the experience carried on in setting real applications in ASR and having in mind the opportunities offered by new devices in a wired and wireless environment, a question arise in order to develop real multilingual communication in the next decade: what next? Which are the main breakthroughs needed? Many issues need to be addressed. First of all how can we reach the necessary performance required by the three basic technologies needed, i.e. speech recognition, synthesis and machine translation. Shall we need a shift in the paradigm of research ? Is it mainly a matter of amount and quality of data needed? How important are issues as devices, multimedia information involved in a human to human dialog, environmental-contextual information provided by intelligent networks? How to integrate all these contextual information in a consistent way ? Many steps and advancements are needed in order to answer these questions. These are some of the questions addressed in a project whose acronym is TC-SPAR_P, technology and corpora for speech to speech translation, recently funded by European Union in the last call of the V framework. In what follows, first of all a state of the art of the basic technologies involved in a S2ST systems is summarized, then the most important challenges are listed and finally the TC-STAR_P project is presented.

2 State of the art

2.1 Speech recognition

In the last 15 years a number of speech recognition tasks have been studied and evaluated. Each task presented different challenges. The features characterizing these tasks are: type of speech (well formed vs spontaneous), target of communication (computer, audience, person), bandwidth (FWB, full bandwidth TWB, telephone bandwidth, FF, far field). Some of these tasks are dictation (WSJ), broadcast news, switchboard, voicemail and meetings. In what follows, they are ordered in terms of the word error rate (wer)

Dictation: 7%, well formed, computer, FBW
Broadcast news: 12%, various, audience, FBW
Switchboard : 20-30% spontaneous, person, TBW
Voicemail: 30% spontaneous, person, TWB
Meetings: 50-60% spontaneous, person FF

At present the spontaneous speech is the feature with the largest effect on word error rate, followed by environment effect and domain dependence.

The main challenge for the next years will be to develop speech recognition systems that mimics human performance. This means in general independent of environment, domain and working as well for spontaneous as for read speech. The focus areas will mainly concentrate first of all *on improving the spontaneous speech* models (i.e prosodic features and articolatory models, multispeaker speech, collect adequate amount of conversational speech,…), modeling and training techniques for multi-environment and multi-domain. Then another key issue will be *language modeling*. It is well known that different static language models work best on specific domain. To implement a language model that works well on many domains will be an important achievement towards the goal of mimicking the human performance. A very quick dynamic adaptation at the level of word/sentence is an important target of the research. Finally other factors driving progress will be the continuous improving of computer speed over time, the independence from vocabulary and the involvement of all the potential researchers in the field, not only a few institutions. Improving the performance of conversational speech and introducing highly dynamic language models are the two fundamental requirement for improving S2ST performances. This is maybe the most critical point because performing under 10%, in conversational speech, seems today an hard problem.

2.2 Speech synthesis

Speech synthesis is an important component in a speech to speech translation system. To mimics human voice is still one of the most challenging goal for speech synthesis. The multilingual human to human communication framework introduce new challenges, gender, age and cultural adaptation. Emotion and prosody are also very important issues [7] [8].

Today the most effective way to generate synthetic speech is based on the concatenation of different acoustic units. This approach is in contrast to traditional rule-based synthesis where the design of the deterministic units required explicit knowledge and expertise. In a corpus based approach the unit selection process involves a combinatorial search over the entire speech corpus, and consequently, fast search algorithms have been developed for this purpose as an integral part of current synthesis systems.

Three are the main factors of the corpus-based methods for a specification of the speech segments required for concatenative synthesis: first of all a unit selection algorithm, then some objective measures used in the selection criteria and finally the design of the required speech corpus. From the application point of view the huge amount of memory necessary for exploiting the concatenation of speech units, strongly limits the class of application.

Prosody and speaker characteristics are, together with speech segments design, the other two important issues in speech synthesis. In order to control prosody, it is necessary to ensure adequate intonation and stress, rhythm, tempo and accent. Segmental duration control and fundamental frequency control are needed. Speech waveforms contain not only linguistic information but also speaker voice characteristics, as manifested in the glottal waveform of voice excitation and in the global spectral features representing vocal tract characteristics. Moreover paralinguistic factors cause changes in speaking styles reflected in a change of both voice quality and prosody.

Prosodic modeling is probably the domain from which most of the improvements will come. Investigation in this direction, try to master linguistic and extra-linguistic phenomena, will address probably multicultural issues, which are very important in a multilingual human to human communication framework.

2.3 Machine Translation

Beside speech recognition and synthesis the translation component is the core of a speech to speech translation system. The classical machine translation (MT) problem, to translate a text in a given language, i.e. Italian, in a target language, i.e. Chinese, is a completely different problem from the S2PT problem First of all in the classical MT problem no human is involved. The process is a one way process. The text is supposed to be linguistically ‘correct’. In the S2ST process two humans are involved, the process is bi-directional, the language is conversational, spontaneous, ungrammatical and mixed with non verbal cues. Moreover the environment, in terms of acoustic noise and modality of interaction is a critical issue.

A near real time translation is mandatory in S2ST. Then, because humans are involved directly in the process, the understanding phase is carried on by humans in a collaborative way. Finally given that anyhow a machine is involved in the translation an important issue related to human machine communication has also to be considered. In order to afford the S2ST problem all these factors have to be taken into account.
Different architectures have been exploited: some using an intermediate language (interlingua, interchange format), some exploiting a direct translation method. A typical example of the first case is represented by JANUS [9] and NESPOLE! architectures. The Italian implementation of NESPOLE! S2ST system architecture] consists of two main processing chains: the analysis chain and the synthesis chain. The analysis chain converts a Italian acoustic signal into a (sequence of), IF representation(s) by going through: the *recognizer*, which produces a sequence of word hypotheses for the input signal; the *understanding module*, which exploits a multi-layer argument extractor and a statistical based classifier to deliver IF representations. The synthesis chain starts from an IF expression and produces a target language synthesized audio message expressing that content. It consists of two modules. The *generator* first converts the IF representation into a more language oriented representation and then integrates it with domain knowledge to produce sentences in Italian. Such sentences feed a speech *synthesizer*.

An example of the direct translation approach is represented by the ATR-MATRIX [10] architecture, which exploit a cascade of a speech recognizer with a direct translation algorithm, TDMT, whose produced text is then synthesized. The direct translation approach is implemented using example based algorithms. A second example of direct translation, based on statistical modeling , has been pioneered by IBM[11] [12], starting from text translation. Statistical translation has also been developed in the European project EU-TRANS and in the framework of German project VERBMOBIL.

At the moment research is going on in order to develop unified or integrated approaches. To unify speech recognition, understanding, and translation as an entire statistical processing is the ultimate goal of this approach as well stated in [13] " We consider this integrated approach and its suitable implementation to be an open question for future research on spoken language translation"
From the performance point of view the most important experience obtained in the VERBMOBIL project, in particular a large-scale end-to-end evaluation, showed that the statistical approach resulted in significantly lower error rates than three competing translation approaches: the sentence error rate was 29% in comparison with 52% to 62% for the other translation approaches.

Finally a key issue for S2ST systems is the end to end evaluation methodology. The goal is to develop a methodology based on objective measurement. Evaluation methodologies have been proposed and developed in VERBMOBIL, C-STAR, and by many other groups.

3 Major Challenges

3.1 Improve significantly the end-to-end performance

This is the first challenge to be addressed in the near future. It seems that unified methodologies based on statistical modeling are very promising, provided that some key issues will be afforded and suitable solutions worked out. This methodology allows to include acoustics, phonetic context, speaking rate, speaker variations, language features such as syntax or semantics, etc. into one unified way. Then this approach jointly optimizes acoustics, language and speaker effects. From the modeling point of you it represents quite a shift from the source model. Much more work is needed in proposing new computational tools and building up. This approach is also consistent with the speech synthesis perspective: corpus based and data driven
A challenge will also be the exploitation of real applications in a limited domain, i.e. tourism, of systems based on interlingua approaches. Key issues in this case are portability and robustness.

3.2 Produce aligned multilingual corpora and lexica

In order to afford the challenge of developing new models with the hope to improve significantly per-

formance a key issue is given by corpora and lexica. In order to afford the problem of spontaneous speech recognition, there are proposals [14] of collecting and transcribing 5000 hours of spontaneous speech. This issue is controversial; anyhow this is what we have learn from the past experience in speech recognition. The test data could be a mixture of current and new sources. For translation aligned multilingual text corpora are also crucial. An effort is going on in a joint cooperation with ATR and IRST and with the other member of C-STAR III consortium in order to set up an aligned text corpora composed by the transcription and translation of phrase book in the tourism domain. This phrase book cover a broad range of situations: emergency, time table, transport, sightseeing, directions, attractions, hotels, shopping…Aligned multilingual lexical are also important language resources for future S2ST systems development. A current activity is under development in LC-STAR [15] a new funded project in the Vth framework by EU.

3.3 Integrate speech to speech translation components in a real applications

Real services and application involving speech communication need to manage the "interface problem", i.e. the physical impact of the user with a device which involves multimodal, multimedia in a ubiquitous environment. A wearable device, a PDA or 3G cellular cannot be operated by keyboard, and requires sophisticated natural multimodal human interfaces. Speech, vision and handwriting seem natural candidates for human-machine interaction. But how can a system provide seamless integration between human-machine services and human-human services? How can the system blend the two, provide assistance and guidance for a user to access and understand databases and information resources, but also to serve as a go-between to facilitate the interaction with other humans or with a user's direct environment?

4 A new action in Europe

Given the challenges previously discussed and the experience carried on in the previous and ongoing projects a new and innovative initiative is needed to tackle to problem. This initiative in order to be successful need first of all a critical mass of researchers. Within Europe few research groups have the capability to build up complete SST systems. Most research groups are small and work only on some research themes, i.e prosody, acoustic modeling, language modeling, speech synthesis. Although these small groups may have excellent researchers, their work has less impact on the development of SST-components. This new initiative should provide an appropriate infrastructure to use in a effective way the intellectual potential of European researchers. Given the big shift needed in order to set up this new action, a group of European major players in the spoken language technology, both research institutions, industrial entities, and ELDA proposed a preparatory action, which acronym is TC-STAR_P (Technology and Corpora for speech translation).

4.1 Goals and activities.

The preparatory action, under negotiation, fits with the action line IST2002-III.5.2 c) "preparing for future research activities". It is scheduled to begin in July 2002. The duration will be one year with the purpose of preparing and getting ready an integrated project for the VI Framework. An integrated project as is a large scale action with the purpose to create the European Research Area, ERA. The activity of the TC-STAR_P will be carried on by the cooperation of the four groups: an industrial group, with proven experience in SST technology development, a research group, with proven experience in research in SST-technologies, an infrastructure group, with proven experience in producing language resources for SST components and with proven experience of evaluation of SST components and systems. Then a dissemination group will be in charge of using and spreading the project's results

Three are the main goals of this action:

- developing research roadmaps and associated implementation models
- identifying and bringing together all relevant actors in the Speech to Speech Translation (SST) area
- investigating effective mechanisms for managing future activities

4.1.1 Preparing RTD roadmaps and associated implementation models

The consortium is composed of different RTD communities: industrial, academics, and infrastructure entities. All these organizations will contribute to develop common visions and analyze research requirements for SST systems. As a result of these tasks, industrial partners will prepare roadmaps for technical implementations and services; the scientific and academic groups will prepare roadmaps for technology improvements; and the infrastructure group will provide roadmaps for LR-production and evaluation campaigns.
The work will include a case study where industrial partners and research partners will provide application-oriented and research input respectively. The infrastructure group will focus on preparatory tasks for setting up production, evaluation and validation centers for the needed LR.

4.1.2 Identifying and bringing together all relevant actors

The consortium includes some of the most relevant actors in the SST field. One of the objectives during the lifetime of the project is to attract further key actors from the industrial, research and infrastructure groups, as well as SMEs working with SST applications and related fields.
Within the infrastructure group, a key action is to attract and prepare contacts with national agencies for funding language specific LR-production in the future FP6, and with entities working on evaluation and validation of language resources. The development of language resources is a very expensive activity, which must be best tackled by coordinated funding actions at national and European levels.

4.1.3 Investigating a new management model

According to the IST 2002 Work programme, Action Line 3.5.2 should focus on building and strengthening RTD *communities* by encouraging research, business and user organisations to develop together common visions and analyse research requirements in order to identify common challenges and objectives; and on investigating effective mechanisms for managing future activities.

Moreover, a cornerstone of the future work to be developed under the Integrated Project is the management structure. In accordance with Action Line 3.5.2., the work to be performed under TC-STAR_P includes exploring a new organizational model in order to allow partners to smoothly collaborate in pursuing the final goal. This important task will be investigated during the project. Issues such as distribution of work and resources, admission and withdrawal of participants, engagement of additional parties, scientific guidance and monitoring, etc. will be examined. The model has to be effective to reach the envisaged goal, to react to external new trends, needs and demands coming from the market, society and scientific community Section 2

References

[1] W. Wahlster (Ed.): *Verbmobil: Foundations of speech-to-speech translations.* Springer-Verlag, Berlin, Germany, 2000

[2] C-STAR Website: http:// www.c-star.org/

[3] NESPOLE! Website: http:// nespole!.itc.it

[4] EU-TRANS Project; Instituto Tecnologico de Informatica (ITI, Spain), Fondazione Ugo Bordoni (FUB, Italy), RWTH Aachen, Lehrstuhl f. Informatik VI (Germany), Zeres GmbH Bochum (Germany): Example-Based Language Translation Systems. *Final report of the EuTrans project* (EU project number 30268), July 2000.

[5] BABYLON Web site www.darpa.mil/ipto/research/babylon/approach.html

[6] R.V.Cox, Candace A. Kamm, Lawrence R. Rabiner, J. Schroeter, J. G. Wilpon Speech and Language Processing for Next-Millennium Communications Services, Proceedings of the IEEE, Vol. 88, No. 8, August 2000

[7] Sammy Lemmety Review of speech Synthesis Technology Master Thesis Helsinky University of Technology, 1999

[8] Ron Cole (Ed): Survey of the State of the Art in Human Language Technology, Spoken Output Technology, chapt. 5 Cambridge University Press 1996

[9] A. Lavie, L. Levin, A. Waibel, D. Gates, M. Gavalda, L. Mayfield: JANUS: Multi-lingual translation

of spontaneous speech in a limited domain. *2nd Conf. Of the Association for Machine Translation in the Americas* pp. 252-255, Montreal, Quebec, Oct. 1995.

[10] Takezawa et al. A Japanese-to-English speech translation system: ATR-MATRIX. In Proceeding of ICSLP 1998 pp. 2779--2782

[11] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer: The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistic*s, Vol. 19, No. 2, pp. 263–311, 1993

[12] Yuqing Gao et alias. Speech to speech translation. *In proceeding of C-STAR III Workshop* Guillin China march 13-14 2002

[13] H. Ney: Speech Translation: Coupling of Recognition and Translation. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* pp. I-517-520, Phoenix, AR, March 1999.

[14] M.Padmanabham, M. Pichney, Large Vocabulary Speech Recognition Algorithms, *IEEE Computer*, pag 42-50 april 2002

[15] LC_STAR Website www.lc-star.com