Project Proposal TC-STAR

Make Speech to Speech Translation Real

Harald Höge

Siemens AG, Corporate Research
Otto Hahn Ring 6, D81739 München, Germany
harald.hoege@mchp.siemens.de

Abstract

The proposed project TC-STAR (technology and corpora for speech to speech translation), which is focused on technology, platform and service development for speech to speech translation components and systems. The components are speech recognition, speech centered translation and speech synthesis. The project is aimed to be launched as an integrated project in the 6th framework of the European Commission. To prepare TC-STAR a preparatory action TC-STAR_P has been launched to set up the infrastructure of TC-STAR. For further preparation the EU-funded project LC-STAR has been started to standardize and to create some corpora and lexica needed for all speech to speech translation components.

Keywords: speech to speech translation, language resources, evaluation

1. Introduction

Dictation and Speech-to-Speech-Translation have been and are one of the big challenges of speech processing technology. In the last decades researchers were quite optimistic, that these challenges could be mastered soon.

For dictation large application areas were predicted. But nowadays dictation is used only for very specific domains (e.g. medical diagnosis). The reason for this restriction lies in the fact, that the word error rate for dictation for an untrained user and an open domain is still in the range of 10% which is far too high for wide spread use. The word error rate must be brought down at least to 1% .

Research projects as C-STAR[1], Verbmobil[2], NESPOLE![3] have been set up to develop speech to speech translation technology. In order to achieve reasonable results the applications have been limited to narrow domains (e.g. time scheduling). But even for these limited domains the sentence error rate for speech centered translation was about 30% (Ney, 2000)[4]. For realistic applications the domains must be much broader and the error rate has to be substantial lower.

In order to build high performance speech to speech translation (SST) systems substantial improvement of the performance of the components

- speech recognition
- speech centered translation
- speech synthesis

composing SST systems have to be achieved. Observing the progress of these components within the last 20 years most progress was achieved whenever rule based approaches were substituted by data driven approaches. In the following we assume, that data driven approaches have enough potential to build the basis for high performance SST systems. What is needed is intensive basic research for all three SST components for several years.

Further we assume, that no deep semantic processing is necessary to achieve high performance for SST systems. Otherwise the break through of SST systems will not happen within this decade.

The architecture of data driven systems is based on a language independent processing kernel working on language dependent data. These data are created from language resources (LR) using a language independent training tool (see. Fig. 1).

This property makes the data driven approach very attractive because the processing kernels are nearly the same for all languages[5]. Thus research and development for the kernel can be focused on a few languages. For commercial localization of SST components the logistics to provide the necessary LR for many languages and application areas has to be build up.

The performance of a SST component can be judged how close the output of a SST component comes to an optimal output. In table 1 the input and output of the three SST components are denoted. To define the optimal output of a SST component for a given input is difficult due to human judgement involved. For speech recognition this definition is in general easy, because in most cases it is quite obvious to denote as a human from the uttered speech (input X) the corresponding word string (output Y). Within a translation system the recognizer has also to provide some prosodic markers, which can be used to detect phrase boundaries and to disambiguate possible syntactic constructions needed for translation. To transcribe these 'optimal' or 'correct' prosodic markers for a given utterance different language experts may come to different conclusions. So here the definition of the optimal output has a subjective component. These subjective judgements are even more evident for speech centered translation (what is the best translation for a

[1] www.C-STAR.org

[2] http://verbmobil.dfki.de/

[3] http://nespole.itc.it/

[4] in these experiments the word error rate of the recogniser was in the range of 25% (spontaneous speech)

[5] still some language specific processing steps are needed. Example are the handling of tonal languages, the handling of liaisons,...

given utterance?) and for speech synthesis (what synthesis sounds best for a given word string in a given context?).

SST Component	Input X	Output Y
Speech recognition	Uttered speech	Prosodic marked word strings
Speech centered translation	Prosodic marked word strings	Translated word strings
Speech synthesis	Translated word strings	Fitting speech segments[6]

Table 1: Input and output of SST-modules

In general we observe a variability in the input X and in the output Y. The variability in speech recognition concerns the variability of the input X due to the fact that the same prosodic marked word string can stem from different speech signals (e.g. from different speakers, different acoustic environments). The variability in speech centered translation and speech synthesis is more observed in the output due to the variable subjective judgements. Variability can be expressed by a probability function P(Y|X), which denotes the probability that a specific output Y is optimal for a given input X. Based on pattern recognition theory (Duda, 1973) we can define the "optimal" output Y_{opt} of SST component for a given input X by

$$Y_{opt} = \arg\max_{Y} P(Y \mid X) \qquad (1)$$

i.e. for a given input X that output Y is defined to be optimal, where P(Y|X) is maximal (maximum likelyhood approach).

The probability function P(Y|X) is closely related to the language dependent data structure shown in Fig. 1. Generally P(Y|X) is very complex and is not known. In order to achieve tractable results P(Y|X) has to be approximated by a much simpler function $P_M(Y|X)$. Such functions are called models. To find good approximations for the optimal model is called the modeling problem. In general the models have free parameters, which have to be estimated (trained) from pairs (Y, X). Those pairs have to be provided by language resources. The more free parameter are needed by the function $P_M(Y|X)$ the more pairs (Y,X) must be provided for sufficient accurate parameter estimation.

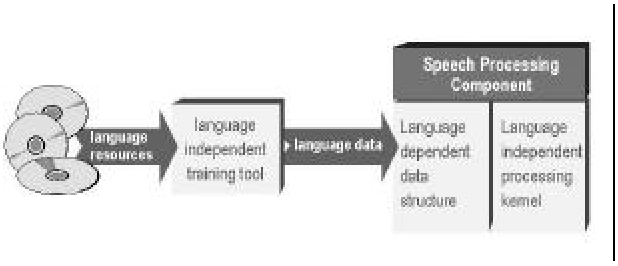
language resources
language independent training tool
language data
Speech Processing Component
Language dependent data structure
Language independent processing kernel

Fig. 1 Language resources allow to transfer efficiently SST-components to other languages

To find the optimal value Y_{opt} as defined in (1) the argmax operator has to be implemented. This problem is called the search or parsing problem. In general a direct implementation of this operator would lead to the task to try out all possible values for Y to determine the maximum of $P_M(Y|X)$. The complexity of this approach increases exponentially with the number of items to be searched for. To make implementation feasible an approximation for the theoretic optimal parsing algorithm have to be found. Such parsing algorithms are closely related to the data structure of $P_M(Y|X)$.

Given the framework of data driven approaches for speech to speech translation two major problems have to be solved

1. The modeling (and derived parsing) problem i.e. find the best models and derived parsers
2. The language resource problem i.e. provide adequate data to train the free parameters of the models

Both problems are related to the questions

Q1: What are the best language resources to improve the models

Q2: What are the best models given available language resources?

Q3: What are the best language resources for given state of the art models?

Within many research institution working on SST components the question Q2 is studied within 'small scale projects'. Often in such projects not sufficient large LRs are used making the presented results questionable. In larger project as the DARPA (Price,1988) and Verbmobil projects question Q1 and Q2 were handled simultaneously. At the beginning of those projects focus is more on Q1 to provide the initial LR to start research on models with related parsers. In the following those LR are called 'experimental LR'. Concerning speech to speech translation no project has been launched which treated Q1 for all SST-components in equal depth. Most focus was on LR needed for speech recognition (Höge, 1998). The reason can be seen in the fact that data driven approaches for speech centered translation and speech synthesis are rather new.

Q3 is a typical commercial question, where LR are needed for many languages and application areas. They are specified in a way to lead to optimal performance for state of the art models. Examples of such LR are those specified and created within the SpeechDat[7] family. In the following such LR are called 'applied LR'.

One of the goals of TC-STAR is to find answers to the questions Q1 till Q3 for all three SST components. In this context not only the 'technical' questions Q1,Q2 and Q3 have to be answered but also the 'organisatorial' question:

[6] Although the output is the synthesised speech, the speech segments selected from the segment database used in concatenative speech synthesis can be seen as the primary output. The transformation of the selected segments to the speech signal is a language independent processing step.

[7] www.speechdat.org

Q4 what is the best organizational model to work on the three questions given the European framework of research and development. In order to answer the technical questions three major activities can be seen:

A1: provide 'experimental LR' to drive the development of models

A2: improve models and relating parsing techniques based on 'experimental LR'

A3: provide 'applied LR' to transfer state of the art SST-components to requested languages and application fields.

To answer question Q4 the action

A4: set up an organizational framework given the 6th framework of the EC and given the European institutions active in research and development of SST components and systems.

has to be performed.

This framework should be optimal to achieve fast progress in actions A1, A2, A3

Currently a European consortium has been set up to launch a project TC-STAR, which is focused on the actions A1, A2, A3, A4. To start this project a preparatory action TC-STAR_P focused on action A4 is planed to start in summer 2002.

Another preparatory project is the EU-funded project LC-STAR started to work on certain aspects of A1, A2, A3.

In the following the project LC-STAR and the basic ideas for the proposed project TC-STAR together with the preparatory action TC-STAR_P is presented.

2. The project LC-STAR[8]

The project LC-STAR (Lexica and Corpora for speech to speech translation) has been launched on 1. Feb. 2002 with the goal to provide lexica and annotated text corpora needed for SST components (see table 3).

LC-STAR is the first industrial guided project with the goal to deliver large lexica.

SST Component	Lexica	Corpora
Speech recognition	Pronunciation Lexica	-
Speech centered translation	Bilingual word Lexica; POS Lexica	Aligned (tagged) bilingual text corpora
Speech synthesis	Pronunciation and POS Lexica	-

Table 3: LR created in LC-STAR

The lexica needed for speech recognition and speech synthesis are in line with the action line A3 as defined in the introduction. Based on state of the art models of speech recognition and synthesis for modeling the relations between graphemes and phonemes the content of the pronunciation lexica is specified. The main activity is to provide these lexica for many languages and to provide a quasi industrial standard for their structure.

The provision of lexica and corpora needed for speech centered translation is in line with action A1, i.e. within the project it is explored what lexica and corpora is needed to provide optimal models for speech centered translation.

2.1. Lexica for speech recognition and speech synthesis

For providing lexica for speech recognition and speech synthesis following activities are performed:

- For 12 languages (see table 4) create word lists in the range of 50 000 general words and evaluate word coverage based on large corpora of more than 10 million words for each language.
- For 12 languages create word lists of 50 000 proper names per language.
- For 12 languages specify lexica suited for speech recognition and speech synthesis containing 50 000 proper names per language and containing sufficient general words yielding a high coverage within each language.[9]
- For 12 languages create and validate the specified lexica.

Russian
Turkish
Italian
Greek
Spanish
Catalan
German
Classical Arabic
Hebrew
US-English
Finnish
Mandarin

Table 4: List of 12 languages

Currently word lists are specified within the project.

2.2. Experimental LR for speech centered translation

For providing experimental LR the following work has to be performed:

- For selected language pairs build up experimental language resources suited for evaluating the kind of LR needed. As far as possible existing LR will be used.
- Build up an experimental platform for speech centered translation based on existing SST components.
- For selected language pairs (see table 5) test the impact on translation quality with respect to the amount of text corpora used and the kind of lexicon information used.

[8] www.LC-STAR.com

[9] To achieve a high coverage for each language about 50 000 words are envisaged

Language pairs
Catalan/ US-English
Spanish/Catalan
Spanish/US-English

Table. 5: language pairs for aligned bilingual text corpora and related lexica

Currently available experimental LR from Verbmobil (German/US-English) are extended to Spanish and Catalan.

3. TC-STAR_P

TC-STAR_P is scheduled to start in summer 2002 with a duration of 1 year and has as goal to prepare the main project TC-STAR.

TC-STAR_P will be carried out 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
- a dissemination group, which will be in charge of using and spreading the project's results

During the lifetime of the TC-STAR_P project, all groups will be responsible for completing their group with relevant actors in their field. These groups will form part of the future project TC-STAR and will work out an organisational model for the implementation of the envisaged action lines as defined in the next chapter.

The projects TC-STAR_P and TC-STAR are industry driven in order to find an organizational model, which assures a fast transfer from results of research to deployed services based on SST components and SST systems. The industrial partner in TC-STAR_P agree that the following basic requirements should be fulfilled.

The basic requirements to the research group are:

- substantial improvement of performance of speech recognition (decrease of error rate)
- substantial improvement of performance of speech centered translation (decrease of error rate)
- substantial reduction of memory consumption for high quality concatenated speech synthesis (learn to manipulate speech segments without loss on speech quality)

The requirements to the infrastructure group will be:

- capability to evaluate the performance requirements for systems and SST components for potential SST-services
- capability to evaluate the performance of developed SST systems and SST components
- capability to create cost effective language resources with given specifications and quality criteria

Main expected results of TC-STAR_P will be:

- the industrial group will provide roadmaps for technology development and service creation for SST components and systems in coordination with roadmaps delivered by the research group and the infrastructure group
- the research group will provide roadmaps for potential improvements of the SST Technologies over time
- the infrastructure group will provide roadmaps for LR-production and SST-technology evaluation
- the dissemination group will create several fora for spreading the project's results
- all groups will be enlarged with the collaboration of other key SST players
- a working model suitable for the management of the integrated project TC-STAR. 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

4. The TC-STAR project

The integrated project TC-STAR (technology and corpora for speech to speech translation) is intended to start in 2003 as a large scale, integrated project embedded in the 6th framework of the EC. The duration of the project should be about five years. TC-STAR has as goal to overcome the language barrier and make speech to speech translation real. This goal will be achieved via a concerted action of SST key actors making SST-technology more and more mature and deploying SST-Services compatible with the maturity achieved.

The action lines of TC-STAR will be:

KA1: basic research to improve substantially the basic performance of SST components and systems

KA2: applied research to adapt SST components to the requirements of application areas and to develop SST components and systems on specific platforms

KA3: development of solutions on application specific platforms using SST components

KA4: Creation of LR and evaluation of SST components and systems

The goal of TC-STAR – i.e. to make speech to speech translation real – is achieved, as soon as many SST-solutions are deployed successfully. So the success of TC-STAR has to be measured on the results on KA3. Consequently the key actions have to be organized in such a way, that the flow of results is running fast from basic to applied research and from applied research to solutions.

KA4 is a service, which has to be fitted optimally on the requirements of the other three key actions.

KA1 is typically performed by public funded institutions. But also companies[10] can be involved.

KA2 lies in the core business of companies deploying SST-components implemented on specific platforms for specific application areas.

KA3 is the field of integrators, which build solutions on given platforms.

KA4 is in line with the activities of ELRA. In this framework many LR has been created and distributed. Currently ELRA extends it's activities to evaluation of LE-systems.

[10] e.g. IBM pioneered dictation and data driven translation

4.1. Overall Organization of TC-STAR

Each action line can be performed by a single or several projects , but the link between the projects working on the same and different action lines must be strong in order to achieve a fast flow of the results from KA1 via KA2 to KA3 and to have efficient support from KA4.

Each action line should be handled by a technical board, which determines and supervises all work done within each action line.

Headed to these technical boards is a steering committee and an administrative unit. The steering committee and the administrative unit have close links to the EC. The administrative unit handles all financial and legal issues within TC-STAR. In order to gain flexibility it is foreseen that 'common money' located at the administrative unit is made available at the beginning of TC-STAR. This money can be used to launch new projects or new activities within projects.

The duty of the steering committee is to coordinate the work between the four action lines and to steer TC-STAR to success i.e. assure that many services based on SST components and systems will be deployed . Due to this goal the steering committee has to be industry driven.

It is foreseen that the steering committee has strong influence on the projects because it determines

- what new projects or actions within a project as proposed by the technical boards should be accepted, modified or rejected
- How the common money is distributed
- Which new partners as proposed by the technical boards should join TC-STAR
- what projects should be stopped, if they do not work successful

This organization has the advantage that the administrative work is centralized and need not to be done for each project. Further the supervision of the work within the different project is decentralized as far as possible by the technical boards. Each technical board build the link between the projects in it's action line. They have to track the deliverables and have to present progress of basic deliverables and milestones to the steering committee and the EC.

4.2. Organization of Basic Research (KA1)

The main task of basic research is

- To improve the performance of the SST-components substantially using the European research potential
- To make the transfer of the research results to applied research as fast and easy as possible.

These two issues must be reflected in the organization of basic research.

To improve substantially the performance of SST-systems 'evolutionary' and ' revolutionary' ideas are needed.

Evolutionary models are those using basic models $P_M(Y|X)$ and performing gradual improvements on these.

Revolutionary ideas are those which provide new basic models $P_M(Y|X)$.

The last revolutionary ideas in speech recognition were born in the late 70th where the HMM's (Baker,1975) and the language models (Jelinek, 1985) were invented. Within several DARPA projects these models have been improved successfully by an evolutionary process. Improvement was achieved by large research groups, which were driven by strong evaluation campaigns. It is important to note that the DARPA projects started with non data driven approaches as ANGEL (Adams, 1986). The revolutionary process during the first DARPA project was when Kai Fu Lee at CMU (Lee, 1988) showed, that speaker independent continuous speech recognition based data driven approaches achieves far better results than the other approaches (The research group around Kai Fu Lee was at the beginning of the DARPA project very small. Later this group crew very fast and developed the open software recognition system SPHINX).

Nowadays it seems that speech recognition needs again revolutionary ideas because only small improvements can be observed.

Speech centered translation was explored to great extent within the large German project Verbmobil. The project started with non data driven approaches. In Verbmobil data driven approaches were explored rather late, but finally proved to achieve the best performance (Ney, 2000). The 'revolutionary' ideas of this approach was explored by the research group within IBM (Brown,1993).

For speech synthesis recently data driven approaches using models for prosody and pronunciation generation were developed. Further fitting segments selected from recorded speech of a speaker (concatenative synthesis) are used.

Summarizing the development of SST-components it is evident that to organize the research a 'revolutionary track' and a 'evolutionary track' is needed. The evolutionary track is driven by competitive strong evaluation campaigns. Within the revolutionary track small groups proposing new ideas should be funded. Due to the current status of the different SST-components it can be stated

- For speech recognition initial focus should be on the revolutionary track The basic models and corresponding search algorithms nowadays used seem to have no great potential for further improvement (Recently progress was mainly in the field of robust feature extraction methods).
- For speech centered translation both tracks should be followed. Speech centered translation based on data driven approaches is a new field and some basic models with corresponding parsers have been developed. These models should be explored via an evolutionary process. Nevertheless the search for new basic models has to be done in parallel.
- For speech synthesis focus should be on the evolutionary track because recently many new ideas around concatenative synthesis have been worked on, which must be settled and further refined

4.2.1. Research groups and Infrastructure

To be successful as a research group a critical mass is needed. This issue is also a matter of infrastructure and focus. Research groups should be supported by an appropriate infrastructure, e.g. the creation of experimental LR and support of evaluation should be 'outsourced' (Key action KA4). Further the focus should not to be on applied research and demonstrator development but on fundamental performance improvement verified by evaluation.

Within Europe few research groups are large enough to be able to build up complete SST systems. Most research groups are small and work on certain aspects of

SST-components. Although these small groups may have excellent researchers their work contribute insufficiently to the improvement of SST-components. The aim of TC-STAR should not be to install many large research groups (this would take too much time and resources) but should provide an appropriate infrastructure to use the intellectual potential of European researchers .

Such an infrastructure could be:

- Within TC-STAR the few large European research groups build up open software for each SST-component, which allow easy to integrate new modules in these components and are suited to improve performance. Further these modules must be easy to handle for performance evaluation. Examples of such open software are HTK or Festival.
- One large research group is responsible for building up a complete SST system for evaluation
- Experimental LR is provided free off charge for those institutions making research on SST components.

The advantages of such an infrastructure is

- Large research groups can push SST technology in the evolutionary track
- The intellectual potential of small groups can be explored in both tracks
- The aspects of special features of the European languages can be considered
- The funding agencies of the European countries can be integrated to support national groups and ‘national’ LR creation.

4.2.2. Transfer of Results

In order to achieve a fast transfer of results from basic research to applied research the open platform infrastructure is an adequate mean. Recently this has been demonstrated in the AURORA project, where in the framework of the open platform HTK noise robust feature extraction methods needed for distributed processing applications have been developed. The research was done industry driven and has finally let to a standard.

5. Organization of applied research (KA2)

Currently for speech processing technology, where the speech processing components speech recognition, dialog, speech synthesis are deployed, the following 4 application areas are identified:

- Network based server (e.g. automated call server)
- Mobile devices (e.g. voice control of mobile phones)
- Automotive (e.g. voice supported navigation)
- Office (e.g. dictation)

It is assumed that in future also speech centered translation will be used as a new innovative component in all these application areas.

For all these application areas specific platforms are used, which differ mainly in available computing power provided by specific processors, available memory and APIs . Further each application area has additionally specific requirements on the SST components (e.g. noise robust recognition for mobile and automotive applications; different vocabulary size for small or large domain applications;...).

The main goal of applied research is to deliver standardized SST components, which can be implemented on platforms covering all 4 application areas. To achieve this goal several tasks have to be performed:

- Adapt algorithms developed in KA1 to the needs of application areas and their platforms
- Specify and create applied LR as far they are not covered by the SpeechDat family and the LC—STAR project (this task is strongly supported by KA4)
- Specify APIs for all SST components and all application areas
- Deliver standardized SST components, which can be implemented on platforms covering all 4 application areas

A specific issue are the IPRs. Companies deploying SST components as their core business can not deliver their SST components royalty free. Special arrangements for the partners within TC-STAR have to be made (e.g. provider of SST-components can give a restricted number of free royalties to their partners in KA3).

6. Conclusion

The paper provides first thoughts to launch an integrated project within the 6th framework of the European commission for speech to speech translation.

7. References

Adams, D., Bisiani, R., 1986. The Carnegie_Mellon University Distributed Speech Recognition system. *Speech Technology March/April 1986*: 14 - 23

Baker, J. K. , 1975. Stochastic modelling as a means of automatic speech recognition. *PHD dissertation, Carnegie Institute of Technology, Carnegie-Mellon University*

Brown, P., F., Della Pietra, S., A., Della Pietra, V., J.,Mercer, R., L.,1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics,* V19, 2:263-311

Duda, H., Hart, P., 1973. Pattern Classification and Scene Analysis. New York: John Wiley & Sons

Höge, H., 1998. Spoken Language Resources for Voice Driven Man Machine Interfaces. *LREC 1998 Proceedings*

Jelinek, J., 1985. The development of an Experimental Discrete Dictation Recognizer. *Proc. IEEE* Vol73,11:1616-1624

Lee, K. F., Hon, H., 1988. Large-vocabulary speaker-independent continuous speech recognition using HMM. *Proc. IEEE ICASSP88:* 123-126

Ney, H., F. J. Och, F., J., Vogel, S.,2000. Statistical Translation of Spoken Dialogues in the Verbmobil System. *In Workshop Multi-Lingual Speech Communication 2000,* Kyoto, Japan, October 2000: 69-74.

Price, P., Fisher, W.M., Bernstein, J., Pallet, D. S. 1988. The DARPA 1000-Word Resource Management Database for continuous speech recognition. In *Proc. IEEE ICCASP88* :651-654