Computer-Supported Human-Human Multilingual Communication

Alex Waibel[†‡]
with: *Keni Bernardin*[†] *and Matthias Wölfel*[†]

InterACT
International Center for Advanced Communication Technology
[†]Universität Karlsruhe (TH), Karlsruhe, Germany
[‡]Carnegie Mellon University, Pittsburgh, PA, USA
ahw@cs.cmu.edu

10.21437/Interspeech.2007-3

Abstract

Computers have become an essential part of modern life, providing services in a multiplicity of ways. Access to these services, however, comes at a price: human attention is bound and directed toward a technical artifact in a human-machine interaction setting at the expense of time and attention for other humans. This paper explores a new class of computer services that support human-*human* interaction and communication *implicitly* and *transparently*. Computers in the Human Interaction Loop (CHIL), require consideration of all communication modalities, multimodal integration and more robust performance. We review the technologies and several CHIL services providing human-human support. Among them, we specifically highlight advanced computer services for *cross-lingual* communication.

Index Terms: speech to speech, machine translation, simultaneous translation, domain-independence, multimodal interaction, perceptual user interfaces, language portability.

1. Introduction

It is a common experience in our modern world, for humans to be overwhelmed by the complexities of technological artifacts around us, and by the attention they demand. While technology provides wonderful support and helpful assistance, it also gives rise to an increased preoccupation with technology itself and with a related fragmentation of attention. But as humans, we would rather attend to a meaningful dialog and interaction with other humans, than to control the operations of machines that serve us. The cause for such complexity and distraction, however, is a natural consequence of the flexibility and choices of functions and features that the technology has to offer. Thus flexibility of choice and the availability of desirable functions are in conflict with ease of use and our very ability to enjoy their benefits. The artifact cannot yet perform autonomously and requires precise specification of the machine's behavior. Standardization, better graphical user interfaces, multimodal human-machine dialog systems, speech, pointing, mousing have all contributed to improve the interface, but still force the user to interact with a machine at the detriment of other human-human interaction.

To change the limitations of present day technology, machines must engage implicitly and indirectly in a world of humans, that is we must put Computers in the Human Interaction Loop (CHIL), rather than the other way round. Computers should assist humans engaged in human-human interaction, by providing implicit and proactive support. If technology could be "CHIL enabled" in this way, a host of new services could potentially be possible. Could two people be connected with each other at the best moment over the most convenient and best media, without phone tag, embarrassing ring tones and interruptions? Could an attendee in a meeting be reminded of participants' names and affiliations at the right moment without messing with a contact directory? Can meetings be supported, moderated and coached without technology getting in the way? And: Could computers enable speakers of different languages communicate and listen to each other gracefully across the language divide?

Human assistants often provide such services; they work out logistical support, reminders, helpful assistance, and language mediation; they can do it reliably, transparently, tactfully, sensitively and diplomatically. Why not machines? Clearly, a lack of recognition and understanding of human activities, needs and desires are to blame, and an absence of socially adept computing services that mediate rather than intrude. In the following we focus on these two elements: 1.) technologies to track and understand the human context, and 2.) computing services, that mediate and support human-human interaction.

2. Understanding the Human Context

In contrast to classical human-machine interfaces, implicit computer support for human-human interaction requires a perceptual user interface with much greater performance, flexibility and robustness, than is available today. This challenge has lead to research aimed at tracking all the communication modalities in realistic recording conditions, rather than individual modalities in idealized recording conditions. CHIL and AMI, both Integrated projects under the 6th Framework Program of the European Commission, as well as CALO, a DARPA program are among the more recent efforts aiming to take on this challenge.

In the following we will discuss computer services that support human-human interaction. To realize this goal, work concentrates on four key areas: The creation of robust perceptual technologies able to acquire rich and detailed knowledge about the human interaction context; the collection and annotation of realistic, audio-visual meeting and seminar data necessary for the development and systematic evaluation of such; the definition of a common software architecture to support reusability and exchangeability of services and technology modules; the implementation of a number of prototypical services offering proactive, implicit assistance based on the gained awareness about human interactions.

2.1. Audio-visual Perceptual Technologies

2.1.1. Introduction

Multimodal interface technologies "observe" humans and their environments by recruiting signals from multiple AV sensors to detect, track, and recognize human activity. The analysis of all AV signals in the environment (speech, signs, faces, bodies, gestures, attitudes, objects, events, and situations) provides the proper answers to the basic questions of "who", "what", "where", and "when", that can drive higher-level cognition concerning the "how" and "why", thus allowing computers to engage and interact with humans in a human-like manner using the appropriate communication medium (see Figure 1).

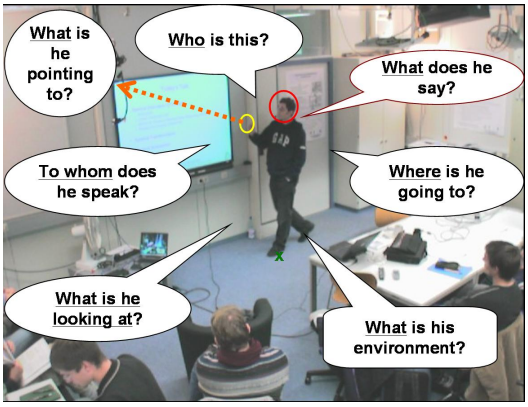

Figure 1: The "who", "what", "where", "when", "how" and "why" of human interaction

Research work performed and progress made on a number of such technologies is described next. Whereas technological advances for multimodal systems were hard to measure in the past for lack of common benchmarks, recent efforts in the community have led to the creation of international evaluations such as the CLEAR (Classification of Events, Activities and Relationships) [1] and RT (Rich Transcription) [2] evaluations, which offer a platform for large-scale, systematic and objective performance measurements on large audio-visual databases.

2.1.2. Person Tracking

Location and tracking of multiple persons behaving without constraints, unaware of audio/video sensors, in natural, evolving and unconstrained scenarios, still poses significant challenges.

Video-based approaches based on background subtraction are error prone due to varying illumination, shadows and occlusion, whereas those relying on the feature space (e.g. color histograms) are difficult to initialize reliably for every new acquired target. Many approaches that offer higher reliability are simply too computationally expensive to be used in online applications.

Audio-based localization and tracking requires the tracked person to be actively speaking, and have to deal with the variety of acoustic conditions (e.g., room acoustics and reverberation) and, in particular, the undefined number of simultaneous active noise sources and competing speakers found in natural scenarios.

Several strategies are being applied to face the challenges mentioned above. Distributed camera and microphone networks, including microphone arrays placed in different positions in space, provide a better "coverage" of each area of interest. Fusion of sensor data in multi-view approaches overcomes occlusion problems, as in the case of 3D background subtraction techniques combined with shape from silhouette [3]. Probabilistic approaches computing the product of single view likelihoods using generative models which explicitly model occlusion have proved efficient in managing the trade-off between reliable modeling and computational efficiency [4] (see also Figure 2). Fusion of multimodal data for speaker localization in e.g. particle filtering approaches increases robustness for speaker tracking [5]. Efficient tracking is a useful building block for all subsequent technologies. It has been shown, e.g. that multimodal fusion helps increase localization accuracy, and that this in turn has direct impact on the performance of far-field speech recognition [6,7] (see also Figure 3).

***Figure 2**: Audio-visual tracking of multiple persons. Targets are described by an appearance model comprising shape and color information, and tracked in 3D using probabilistic representations [4]. The system tracks 5 people in real-time through multiple persistent occlusions in cluttered environments.*

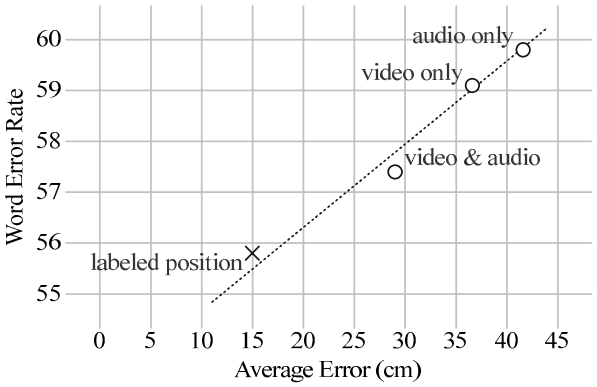

***Figure 3**: Acoustic, visual and multimodal 3D person tracking accuracies and resulting word error rate (after beamforming) on the CHIL 2005 dataset.*

2.1.3. Person Identification

The challenges for audio-visual person identification (ID) in unconstrained natural scenarios are due to far-field, wide-angle, low-resolution sensors, acoustic noise, speech overlap and visual occlusion, unpredictable subject motion, and the lack of position/orientation assumptions to facilitate well-posed signals. Clearly, employing tracking technologies and fusion techniques, either temporal, multi-sensor or multi-modal (speaker ID combined with face ID for example) is a viable approach in order to improve robustness.

Identification performance depends on the enabling technologies used for audio, video and their fusion, but also on the accuracy of the extraction of the useful portions from

the audio and video streams. The detection process for audio involves finding and extracting the speech segments in the audio stream. The corresponding process for video involves face detection. Developed mono- and multi-modal ID systems within CHIL have been successfully evaluated in the CLEAR'06 and '07 evaluations [1], reaching in many cases near 100% accuracies on databases of more than 25 subjects**.** Not only was steady progress made on the key technologies over the past years, showing the feasibility of person ID in unconstrained environments, it was also demonstrated that sensor and multimodality fusion help to improve recognition robustness (see Figure 4)

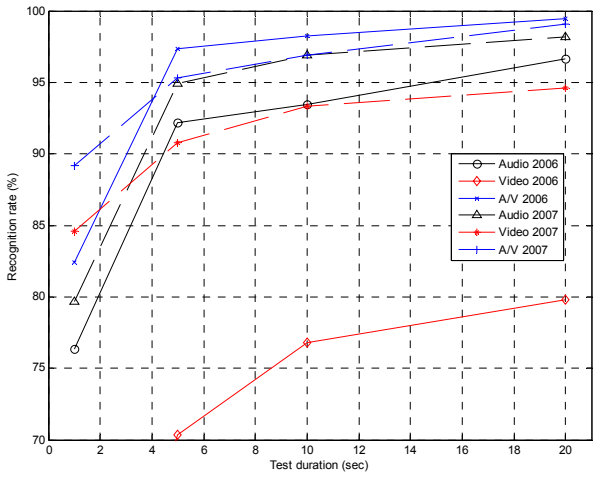

Figure 4*: Acoustic, visual and multimodal identification results for the CLEAR 2006 and 2007 evaluations (only best results shown). Systems were trained on 15 second sequences and tested on 1, 5, 10 and 20 second test sequences. Shown are accuracies for 25 users from 5 sites.*

2.1.4. Head Pose, Focus of Attention

Understanding human interaction requires not only to perceive the state of individuals, but also to determine their person or object of interest, the addressees of speech, and so forth. Since people's head orientations are known to be reliable indicators for their direction of attention [8], systems were developed to estimate the head orientations of people in a smart room using multiple fixed cameras (see also Figure 5). In the CLEAR 2006 head pose dry run evaluation, the first formal evaluation for a task of this kind, classification of pan angles into 45° classes was attempted and accuracies of 44.8% were reached [1]. The challenging CHIL database drove the development of more accurate systems and already in 2007, estimation of exact angles was performed and error rates as low as 7° pan, 9° tilt and 4° roll could be achieved.
Once head orientations are estimated, they can be used to automatically determine the foci of attention of people [9].

2.1.5. Activity Analysis, Situation Modeling

Another useful type of information for unobtrusive, context-aware services is the classification of a user's or a group's current activities. In experiments performed in one of the CHIL sites, typical office activities such as "paperwork", "meeting" or "phone call" were distinguished in a multiple-office setup using only one camera and one microphone per room [10]. A hierarchical classification ranging from low level isolated events such as desk activity, to complex activities, such as leaving a room and entering another, could be achieved. The event classes were learned by clustering audio-visual data recorded during normal office hours over extended periods of time. Figure 6 depicts an example of data-driven clustering of activity regions within an office.

Figure 5*: Estimating Head Pose and Focus of Attention [9]. Head orientations are estimated from four camera views. These are then mapped to likely focus of attention targets, such as room occupants*

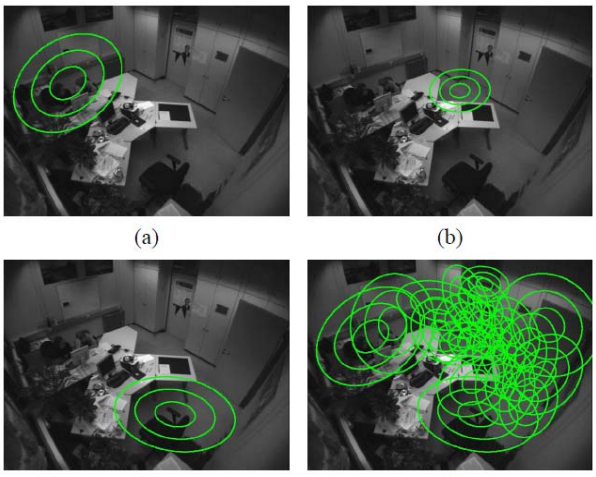

Figure 6*: Data-driven training of activity regions in an office room[10]. The regions labeled as a), b) and c) represent the learned areas of activities by office workers and their visitors, whereas d) depicts all resulting clusters. Evaluation of an unconstrained one week recording session revealed accuracies of 98% for "nobody in office", 86% for "paperwork", 70% for "phone call" and 60% for "meeting"*

2.1.6. Speech Activity Detection, Speaker Diarization

These two related technologies are relevant not only for Automatic Speech Recognition (ASR), but also for speech detection and localization and for speaker identification. Speech activity detection (SAD), addresses the "when'' of the speech interaction, and speaker diarization, addresses both "who'' and "when". Both have been evaluated on the CHIL interactive seminar database in the latest CLEAR and RT evaluations, using primarily far-field microphones.

2.1.7. *Recognition of Speech and Acoustic Events*

Speech is the most critical human communication modality in seminar and meeting scenarios, and its automatic transcription is of paramount importance to real-time support and off-line indexing of the observed interaction. Although automatic speech recognition (ASR) technology has matured over time, natural unconstrained scenarios present significant challenges to state-of-the-art systems. For example, spontaneous and realistic interaction, with often accented speech and specialized topics of discussion (e.g., technical seminars), as well as overlapping speech, interfering acoustic events, and room reverberation degrade significantly the ASR performance. These factors are further exacerbated by the use of far-field acoustic sensors, which is unavoidable in order to free humans from tethered and obtrusive close-talking microphones.

Various research sites have been developing ASR systems to address these challenges, and have benchmarked their performance, e.g. in the recent RT'06 and '07 evaluations. There, the best far-field ASR system achieved a word error rate (WER) of 44% (52% in 2006), by combining signals from multiple (up to four) table-top microphones. It is interesting to note that this is considerably higher than the 31% (also 31% in 2006) WER achieved on close-talking microphone input – with manual segmentation employed to remove unwanted cross-talk. These results demonstrate the extremely challenging nature of the task at hand.

Various research approaches are being currently investigated to improve far-field ASR. Some employ multi-sensory acoustic input, for example beamforming that aims to efficiently combine acoustic signals from microphone arrays [6], and speech source separation techniques that attempt to improve performance during speech overlap segments. A different multimodal approach considered is to recruit visual speech information from the speaker lips, captured from properly managed pan-tilt-zoom cameras, in order to improve recognition through AV-ASR.

Finally, one should note that speech is only one of the acoustic events occurring during human interaction scenarios. Technology is being developed to detect and classify acoustic events that are informative of human activity, i.e., clapping, keyboard typing, door closing, etc. [1].

2.2. Technology Evaluations, Data Collection & Software Architecture

To drive rapid progress of the presented audio-visual perceptual technologies, their systematic evaluation using large realistic databases and common task definitions and metrics is essential.

Technology evaluations, undertaken on a regular basis, are necessary so that improvements can be measured objectively and different approaches compared. An important aspect is to use real-life data covering the envisioned application scenarios. In CHIL, large numbers of seminars and meetings were collected in five different smart rooms, equipped with a range of cameras and microphones. The recordings were manually enriched with acoustic event and speech transcriptions as well as several visual annotations that allowed to train and evaluate various technology components (see for example [1] for further details). In contrast to many of the evaluation benchmarks that exist for individual technologies such as face recognition, for example, the data from such realistic scenarios is extremely challenging, containing a combination of many difficulties for perceptual technologies, such as varying illumination, viewing angles, head orientations, low resolution images, occlusion, moving people, varying speaking accents, behaviours, room layouts and technical sensor setups.

Starting in 2006, a large effort was undertaken to create an international forum for evaluation of multimodal technologies for the analysis of human activities and interactions. The CLEAR workshop was created in a joint effort between CHIL [11], the US National Institute of Standards and Technology (NIST) and the US Video Analysis Content Extraction (VACE) [12] program. The goal was to provide the needed discussion forums, databases, standards, and benchmarks necessary to drive the development of multimodal perceptual technologies, much like the NIST Rich Transcription Meeting Recognition (RT) workshop for diarization, speech detection and recognition, or the TRECVID [13], PETS [14] and ETISEO [15] programs for visual analysis and surveillance. More than a dozen evaluation tasks were conducted, including face and head tracking, multimodal 3D person tracking, multimodal identification, head pose estimation, acoustic scene analysis, acoustic event detection, etc.

To offer support for the integration of developed technological components, to realize higher level fusion of information and modeling of interaction situations, and to provide well-defined interfaces for the design of useful user services, a proper architectural framework is of great importance. An example of such an infrastructure is the CHIL Architecture [16]

2.3. Human-Human Computer Support Services

Building on the perceptual technologies and compliant to the software architecture, several prototypical services are being developed that instantiate the vision of context-awareness and proactiveness for supporting human-human interaction.

The target domains are lectures and small office meetings. In the following, some example services, relying on the robust perception of human activities and interaction contexts are presented:

2.3.1. *The Meeting Browser*

The Meeting Browser provides functionality for offline reviewing of recorded meetings, automatic analysis, intelligent summarization or data reduction, generation of minutes, topic segmentation, information querying and retrieval, etc. Although it has been a topic of research for quite some time [17,18], advances in perceptual technologies (such as face detection, speaker separation and far-field speech recognition) have increased its user-friendliness by reducing the constraints on the interaction participants or the need for controlled or scripted scenarios.

2.3.2. *The Collaborative Workspace*

The Collaborative Workspace (CW) [19] is an infrastructure for fostering cooperation among participants. The system provides a multimodal interface for entering and manipulating contributions from different participants, e.g., by supporting joint discussion of minutes or joint accomplishment of a common task, with people proposing their ideas, and making them available on the shared workspace, where they are discussed by the whole group.

2.3.3. *The Connector*

The Connector is an adaptive and context-aware service designed for both efficient and socially appropriate communication [20]. It maintains an awareness of users'

activities, preoccupations, and social relationships to mediate a proper moment and medium of connection between them.

2.3.4. The Memory Jog

The Memory Jog (MJ) provides background information and memory assistance to its users. It offers "now and here" information by exploiting either external databases: (Who is this person? Where is he/she from?) or own ones (Who was there that day? What did he say?), the latter including information gained from the observation of the interaction context [21]. The MJ can exploit its context-awareness to proactively provide information at the proper time and in the most convenient way given the current situation.

2.3.5. Cross-Lingual Communication Services

Another exciting class of services concern cross-lingual human-human communication. Is it possible to communicate with a fellow human speaking a different language as naturally as if he/she spoke your own? Clearly this would be a worthwhile vision in a globalizing world, when international integration demand limitless communication, while national identity and pride demand recognition and respect for the cultural and linguistic diversity on this planet. How could technology be devised to make this possible? We devote the following section to a discussion of this potentially revolutionary class of human communication support and an area of growing speech, language and interface research.

3. Cross-Lingual Human-Human Communication Services

In the past decade, Speech Translation has grown from an oddity at the fringe of speech and language processing conferences, to one of the main pillars of current research activity. The explosion in interest is driven in part, by considerable market pull from an increasingly globalizing world, where distance is no longer measured in miles but in communication ease and cost. Indeed, effective solutions that overcome the linguistic divide may potentially offer considerable practical and economic benefits. For the research community, the linguistic divide may ultimately prove to be a more formidable challenge than the digital divide as it presents researchers with a number of fascinating new problems. The goal is, of course, good human-to-human communication without interference from technical artifacts, and effective solutions must combine efficient and reliable speech & language processing with effective human factors and interface design.

Early developments provided first prototypes demonstrating the concept and feasibility [22,23]. In the mid '90's a number of projects aiming at spontaneous speech two-way speech translators for limited domains (e.g. JANUS-III, Verbmobil, Nespole) followed suit. The Consortium for Speech Translation Advanced Research (C-STAR) was founded in '91 to promote international cooperation in speech translation research. With the turn of the millennium, activity has proceeded in two directions: The first continues to improve domain-limited two-way translation toward *fieldable*, *robust* deployment where domain limitation is acceptable (humanitarian, health-care, tourism, government, etc.). The second has begun to tackle the open challenge of domain limitation for applications such as broadcast news, speeches and lectures. Large new initiatives (NSF-STR-DUST, EC-IP TC-STAR and DARPA GALE were launched in the US and Europe in '03, '04, and '06, respectively, in response. In the following we review these advances.

3.1. Domain-Limited Portable Speech Translators

Fieldable speech-to-speech translation systems are currently developed around portable platforms (laptops, PDA's) which impose constraints on the ASR, SMT, and TTS components. For PDA's memory limitations and the lack of a floating point unit require substantial redesign of algorithms and data structures. Thus, a PDA implementation may impose WER increases from 8.8% to 14.6% [24] over laptops. In addition to continued attention to speed, recognition, translation and synthesis performance, however, usability issues such as the user interface, microphone type, place and number, as well as user training and field maintenance must be considered. One of the resulting speech-to-speech graphical user interfaces (GUI) of a PDA pocket translators is shown in Figure 7.

***Figure** 7: A PDA pocket translator [English-Thai]* [1].

The GUI window is divided into two regions, showing the language pairs. These regions can be populated by recognized speech output (ASR), translation output (SMT), or by a virtual PDA keyboard for backup. A back-translation is provided for verification; a push-to-talk button activates the device and aborts processing for false starts and errors. Projects (e.g DARPA Transtac) and workshops (e.g. IWSLT, sponsored by C-STAR) provide for collaboration, data exchange and benchmarking that improve performance and coverage in this space.

3.2. Translation of Parliamentary Speeches and Broadcast News

For speech-translation without domain limitation, component technologies first had to be developed that deliver acceptable ASR, SLT (and TTS) performance in face of spontaneous speech, unlimited vocabularies, broad topics, and speaking style characteristic of spoken records. In TC-STAR, speeches from the European Parliament (and their manual transcriptions and translations) were used as data to train and evaluate. Figure 8 shows the improvements over the years in speech recognition and automatic translation within the project. In these experiments it has been seen that there is almost a linear correlation between WER and machine translation quality. We also found that a WER of around 30% is influencing the machine translation quality significantly while a WER of 10% provides for reasonable translation

[1] Courtesy of Mobile Technologies, LLC, Pittsburgh

compared to reference transcriptions. The goal of a different ambitious speech translation project, GALE (Global Autonomous Language Exploitation) [25], is to provide relevant information in English, where the input comes from huge amounts of speech in multiple languages (a particular focus is on broadcast news in Arabic and Chinese). However, progress is not measured by WER and BLEU, but how fast a particular goal can be reached.

Figure 9 compares human and computer speech-to-speech translations on five different aspects by human judgment: was the message understandable (understanding), was the output text fluent (fluent speech), how much effort does it take to listen to the translation (effort) and what is the overall quality, where the scale ranges from 1 (very bad) to 5 (very good). The fifth result shows the percent accuracy by which questions of content could be answered by human subjects based on the output from human and machine translators. It can be seen that automatic translation quality still lags behind human translation, but reaches usable and understandable levels already close to human translations. It is interesting to note, that the human translations also fall short of perfection due to the fact that humans translators occasionally omit information.

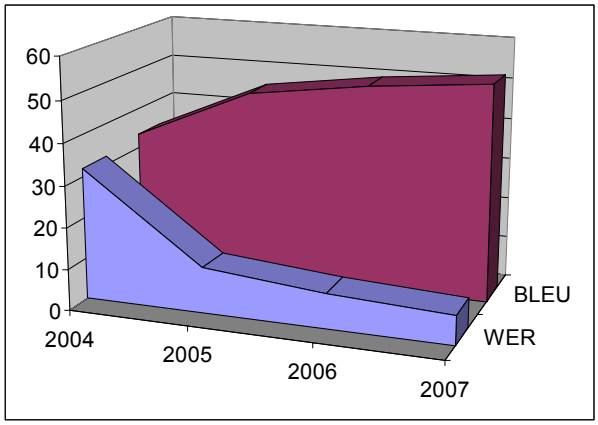

***Figure 8**: Improvements in Speech Translation and Automatic speech recognition over the years on English EPPS and translation into Spanish. (source [26,27])*

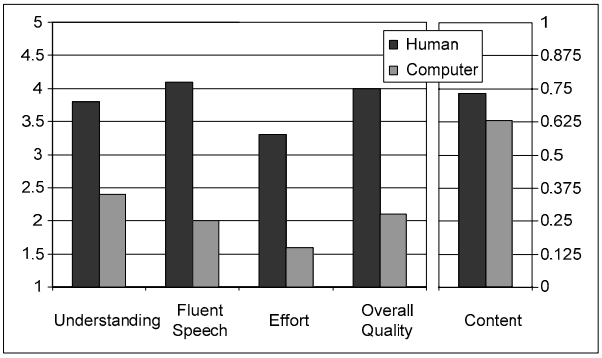

***Figure 9**: Human vs. automatic translation performance. (source [28])*

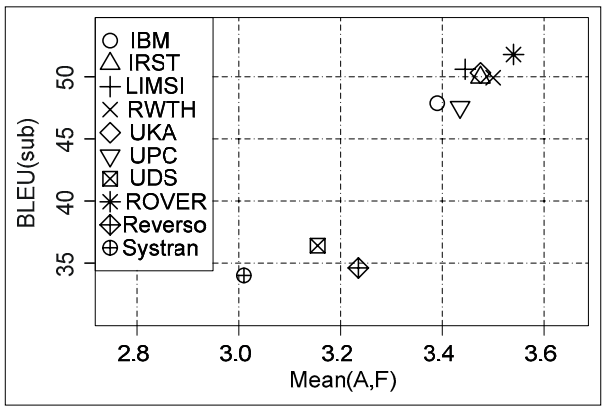

***Figure 10**: BLEU scores show good correlation with human judgements (fluency & accuracy) for English to Spanish translations. (source [27])*

An important aspect in all automatic evaluations are good metrics that can be evaluated automatically and repetitively. While WER is an established method to measure accuracy of automatic speech transcriptions, automatic MT metrics have only recently been proposed. Figure 10 shows the BLEU score (one of several popular MT scoring metrics) and its good correlation with human judgements (adequacy, fluency) on the European Parliament data.

3.3. Unlimited Domain Simultaneous Translation

The ultimate cross-lingual communication tool would be a simultaneous translator that produces simultaneous real-time translation of spontaneous lectures and presentations. Compared to parliamentary speeches and broadcast news, lectures, seminars, presentations of any kind, present further problems for domain-unlimited speech translation by

- Spontaneity of free speech, the disfluencies, the ill-formed nature of spontaneous natural discourse
- Specialized vocabularies, topics, acronyms, named entities and expressions in typical lectures and presentations (by definition specialized content)
- Real-time and low-latency requirements and on line adaptation to achieve *simultaneous* translation and
- Selection of translatable chunks or segments

3.3.1. The Lecture Translator

To address these problems in ASR and MT engines, changes to an off-line system are introduced as follows:

- To speed up recognition, acoustic models can be adapted to a particular speaker. The size of the acoustic model is restricted (for additional speed up when evaluating the Gaussian mixture model one can use techniques such as Gaussian selection) and the search space is more rigorously pruned.
- To adapt to particular speaker style and domain, the language model is tuned offline on slides and publications provided by the speaker, either by reweighting available text corpora or by retrieving relevant training material through the internet or on previous lectures given by the same speaker.
- As almost all MT systems are trained on data split at sentence boundaries and therefore ideally expect sentence like segments as input, particular care has to be taken for suitable online segmentation. We have observed that extreme deviations from

sentence based segmentation can lead to significant decreases in performance. In view of minimizing overall system latency, however, shorter speech segments are preferred. In addition to providing efficient phrase translation on-the-fly, word-to-word alignment is optimally constrained for entire sentence pairs[29].

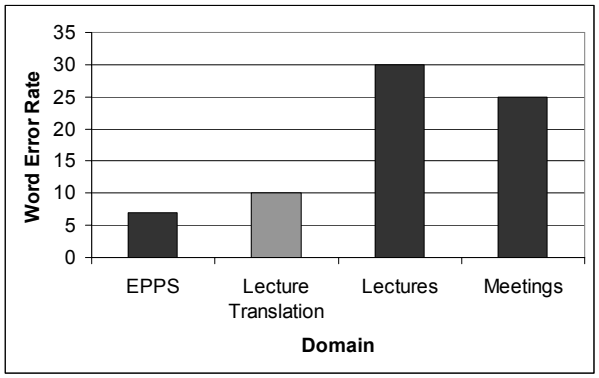

***Figure 11**: Current performance of speech recognition systems on different domains (source [28,30,31], black = speaker independent off line system, gray = speaker dependent online system)*

Figure 11 compares WERs on different domains for English. With a tweaked speaker dependent lecture recognition system we reach a sufficient good performance of 10% WER. On an end-to-end evaluation of the system from English into Spanish we got a BLEU score of 19 while on reference transcripts we got a score of 24 (source [30]).

3.3.2. Delivering Translation Services (Output Modalities)

Aside from speech and language challenges, lecture translation also presents human factor challenges, as the service should be provided unobtrusively, i.e., with minimal interference or disruption to the human-human communication. Several options are being explored:

- *Subtitles*: Simultaneous translations can be projected to the wall as subtitles. This is suitable if the number of output languages is small.
- *Translation goggles*: Heads-up display goggles that display translation text as captions in a pair of personalized goggles. Such goggles provide unobtrusive translation and exploit the parallelism between the acoustic and visual channel. This is particularly useful, if listeners have partial knowledge of a speaker's language and wish to obtain complementary language assistance.
- *Targeted Audio Speakers*: Under the project CHIL, a set of ultra-sound speakers with high directional characteristics has been explored, that can provide a narrow audio beam to an individual listener or a small area in the audience, where simultaneous translation is required. Since such speakers are only audible in a narrow area, it does not disturb other listeners, or could be complemented by similar translation services into other languages to several other listener areas. [32]
- *PDA's, Display Screens or Head-Phones*: Naturally, output translation can also be delivered through traditional display technology, i.e., displayed on a common screen, a personalized PDA screen or acoustically via head-phones.

3.4. The Long Tail of Language

With promising solutions to the language divide under way, language portability remains the unsolved issue. At current estimates, there are more than 6,000 languages in the world, but language technology is only being developed for the most populous or wealthy languages of the world. Most languages along the long tail of language (Figure 12) remain unaddressed. Overcoming the language divide thus requires workable solutions to providing solutions to the long tail of language, at reasonable cost. Most current research is focused on improving cross-lingual technology by employing ever larger data, personnel or computational resources. To address the long tail of language, an orthogonal direction should be concerned with making do with less at lower cost.

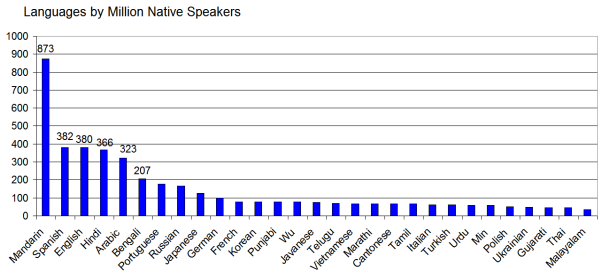

***Figure 12**: The long tail of languages*

At our center, we are therefore exploring several intriguing possibilities that lower cost that could some day bring this problem within reach as well:

- Language independent or adaptive components (this was demonstrated already for acoustic modeling[33]
- More selective parsimonious use of data and data collection [34]
- Interactive and implicit training by the user [35]
- Training on simultaneously *spoken* translation thereby eliminating the need for parallel text corpora [36]

4. Acknowledgements

The work presented here was supported in part by the *European Union* (EU) (projects CHIL (Grant number IST-506909) and TC-STAR (Grant number IST-506738)), by NSF (ITR STR-DUST), by DARPA (projects TRANSTAC and GALE). I would also like to thank the CHIL, TC-STAR, GALE, TRANSTAC partners and the InterACT research team at Karlsruhe and Pittsburgh for their collaboration and for data and images reported in this paper. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the author and do not necessarily reflect the views of the funding agencies or the partners.

5. References

[1] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garafolo, D. Mostefa, P. Soundararajan, "The CLEAR 2006 Evaluation", Proceedings of the First International CLEAR Evaluation, Springer LNCS 4122.

[2] J. Fiscus, J. Ajot, M. Michel, and J. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation,'' Proc. MLMI, Washington DC, 2006.

[3] C. Canton-Ferrer, J. R. Casas, M. Pardàs, “Human Model and Motion Based 3D Action Recognition in Multiple View Scenarios”. EUSIPCO, Firenze, September 2006
[4] O. Lanz, “Approximate Bayesian Multibody Tracking”. IEEE Trans. PAMI, vol. 28, no. 9, September 2006
[5] R. Stiefelhagen, K. Bernardin, H.K. Ekenel, J. McDonough, K. Nickel, M. Voit, M. Wölfel, “Audio-Visual Perception of a Lecturer in a Smart Seminar Room”. Signal Processing, Vol. 86 (12), December 2006, Elsevier.
[6] M. Wölfel, K. Nickel, and J. McDonough. “Microphone array driven speech recognition: Influence of localization on the word error rate”, Proc. of MLMI, Edinburgh, UK, 2005.
[7] H. K. Maganti and D. Gatica-Perez “Speaker Localization for Microphone Array-Based ASR: The Effects of Accuracy on Overlapping Speech”, ICMI, Banff, Canada, Nov. 2006.
[8] C. Wojek, K. Nickel, R. Stiefelhagen, “Activity Recognition and Room-Level Tracking in an Office Environment”. Proc. of the IEEE Intl. Conference on Multisensor Fusion and Integration for Intelligent Systems, Heidelberg, Germany, 2006.
[9] R. Stiefelhagen, J. Yang, A. Waibel, “Modeling Focus of Attention for Meeting Indexing”. ACM Multimedia, Orlando, Florida, Oct. 1999
[10] M. Voit, R. Stiefelhagen, “Tracking Head Pose and Focus of Attention with Multiple Far-field Cameras”. ICMI, Banff, Canada, Nov. 2006.
[11] CHIL – Computers in the Human Interaction Loop, http://chil.server.de
[12] VACE – Video Analysis and Content Extraction, http://www.ic-arda.org
[13] TRECVID – TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/t01v/
[14] PETS – Performance Evaluation of Tracking and Surveillance, http://www.pets2006.net/
[15] ETISEO – Video Understanding Evaluation, http://www.silogic.fr/etiseo
[16] “D2.2 Functional Requirements & CHIL Cooperative Information System Software Design, Part 2, Cooperative Information System Software Design”, http://chil.server.de
[17] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen, “Meeting browser: Tracking and summarizing meetings”. In Proceedings of the Broadcast News Transcription and Understanding Workshop, pp. 281-286, Lansdowne, Virginia, 1998.
[18] M-M. Bouamrane and S. Luz, “Meeting browsing”, Multimedia Systems, Springer-Verlag, 12 (4-5):439-457, 2006.
[19] Q. Y. Wang, A. Battocchi, I. Graziola, F. Pianesi, D. Tomasini, M. Zancanaro, C. Nass. “The Role of Psychological Ownership and Ownership Markers in Collaborative Working Environment”. ICMI. Banff, Canada, 2006
[20] M. Danninger, T. Kluge, R. Stiefelhagen, “MyConnector – Analysis of Context Cues to Predict Human Availability for Communication”. ICMI, Banff, Canada, 2006.
[21] J. Neumann, J. R. Casas, D. Macho, J. Ruiz, “Multimodal Integration of Sensor Networks”. Proc. of AIAI, pp. 312-323, Athens, Greece, 2006.
[22] A. Waibel, A. N. Jain, A. E. McNair, H. Saito, A. G. Hauptmann, J. Tebelskis. “JANUS: A Speech-to-speech Translation Using Connectionist and Symbolic Processing Strategies.” In Proc. of ICASSP'91, pages 793-796, May 1991.
[23] 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.
[24] R. Hsiao, A. Venugopal, T. Köhler, Y. Zhang, P. Charoenpornsawat, A. Zollmann, S. Vogel, A. W. Black, T. Schultz, A. Waibel, "Optimizing Components for Handheld Two-way Speech Translation for English-Iraqi Arabic System", Proceedings of Interspeech, 2006
[25] GALE – Global Autonomous Language Exploitation http://www.darpa.mil/ipto/programs/gale
[26] J. L. Gauvain "Speech transcription: general presentation of existing technologies within TC-Star”. TC-Star Review Workshop, Luxembourg, May 28-30, 2007
[27] H. Ney "TC-Star: Statistical MT of Text and Speech". TC-Star Review Workshop, Luxembourg, May 28-30, 2007
[28] K. Choukri “Importance of the Evaluation of Human-Language Technologies”. TC-Star Review Workshop, Luxembourg, May 28-30, 2007
[29] M. Kolss, B. Zhao, S. Vogel, A. Hildebrand, J. Niehues, A. Venugopal, and Y. Zhang. “The ISL Statistical Machine Translation System for the TC-STAR Spring 2006 Evaluation” In Proc. of the TC-STAR Workshop on Speech-to-Speech Translation, Barcelona, Spain, June 2006.
[30] C. Fügen, M. Kolss, M. Paulik, A. Waibel: “Open Domain Speech Translation: From Seminars and Speeches to Lectures”, In Proc. of the TC-STAR Workshop on Speech-to-Speech Translation, Barcelona, Spain, 2006.
[31] J. Fiscus and J. Ajot “The Rich Transcription 2007 Speech-To-Text (STT) and Speaker Attributed STT (SASTT) Results”, The Rich Transcription 2007 Meeting Recognition
[32] D. Olszewski, F. Prasetyo, and K. Linhard, “Steerable Highly Directional Audio Beam Louspeaker”. In Proc. of the Interspeech, Lisboa, Portugal, September 2006
[33] Tanja Schultz, “Multilinguale Spracherkennung - Kombination akustischer Modelle zur Portierung auf neue Sprachen”. PhD thesis, Universität Karlsruhe, June 2000
[34] M. Eck, S. Vogel, A. Waibel, “Low Cost Portability for Statistical Machine Translation based on N-gram Frequency and TF-IDF”. Proc. of IWSLT, Pittsburgh, PA, Oct 2005
[35] M. Gavalda, A. Waibel, “Growing semantic grammars”. In Proceedings of the COLING/ACL, Montreal, Canada, 1998.
[36] M. Paulik, S. Stüker, C. Fügen, T. Schultz, T. Schaaf, A. Waibel, “Speech Translation Enhanced Automatic Speech Recognition”. ASRU, Cancun, Mexico, December 2005