[From: *Tralogy*, 3-4 March 2011, Paris]

Dimitra Anastasiou

Survey on Speech, Machine Translation and Gestures in Ambient Assisted Living

Abstract

In this paper we provide the state-of-the-art of existing proprietary and free and open source software (FOSS) automatic speech recognition (ASR), speech synthesizers, and Machine Translation (MT) tools. We also focus on the need for multimodal communication including gestures, furnishing some examples of 3D gesture recognition software. Our current experiment is based on interoperability between FOSS ASR, MT, and text-to-speech applications, while future experiments will include gesture recognition tools. Our application environment is an ambient assisted living lab at the University of Bremen, suitable for the elderly and/or people with impairments. In a nutshell, our goal is to provide a single uniform multimodal interface combining FOSS speech processing, MT, and gesture recognition tools for people in need.

Introduction

Today we are experiencing the aging population phenomenon. In 2002 United Nations stated that the number of older persons has tripled over the last 50 years and will more than triple again over the next 50 years. In 2010 the German Federal Statistical Office emphasized that in Germany in 2060 there will be as many 80 years old and older people as younger than 20 years old. Ambient assisted living (AAL)' s initiative is to make the lifestyle of the elderly or people in need more autonomous in a domestic environment by means of advanced technology. Technology used by people in need should be easy-to-use, user-friendly, accessible, flexible, and interaction efficient. All these are advantages of multimodality (see D' Andrea et al., 2009), which is the seamless combination between different modes of interaction, such as haptics, speech, gesture, and vision. Besides multimodality, multilinguality facilitates not only human-human, but also human-computer interaction.

Multilinguality is enabled through translation. Translation, in general, facilitates communication between people in different locales. Using intelligent devices of AAL in crosslingual environments means not only that user interfaces are multilingual, but also that speech-to-speech translation systems should be implemented.

This paper is concerned with speech processing, specifically speech recognition and synthesis, machine translation, and gesture recognition. We provide definitions, history, and a survey on proprietary as well as free/open-source software (FOSS). We cover initiatives and tools both from Academia and Industry. We also refer to speech-to-speech translation systems which combine speech recognition, machine translation, and text-to-speech. Dialog systems and their components will be also pointed out. Gesture will be defined, gesture recognition software will be furnished, and culture-specific gestures and their localization will be discussed.

Then we present related work about how speech and gestures are supported in assisted living environments. Our experiment with the goal of having a symbiotic relationship between multimodality and multilinguality, is the development of a unique platform combining FOSS speech-to-text, machine translation, and text-to-speech tools; gesture recognition software will be a future component.

Definitions and Systems Survey

In this chapter we present definitions, the history, and state-of-the-art tools[1] of speech recognition and synthesis, and Machine Translation (MT); speech-to-speech translation, dialog systems, and gesture recognition and localization will also be discussed.

Automatic Speech Recognition

Speech processing is the study of speech signals and the processing methods. It is tied to natural language processing (NLP), as its input can come from and/or the output can go to NLP applications. Speech recognition, speaker recognition, and speech synthesis are included *inter alia* under speech processing. A distinction between speaker recognition and speech recognition should be made here; the former is about recognizing *who* is speaking and the latter *what* is being said. As for speaker recognition, there are three types of speaker models (put in an ascending complexity order): speaker dependent, speaker independent, and speaker adaptive. The first recognizes the speech patterns of only one person, the second of a large group of people, while the third starts with a speaker independent model and continues with training by the speaker.

Jurafsky and Martin (2009: 285) define the goal of Automatic Speech Recognition (henceforth ASR) as to build systems that map from an acoustic signal to a string of words. Automatic Speech Understanding (ASU) produces some sort of understanding of sentences, rather than just words. Some application areas of ASR, which appeal to those who need or want hands-free approach, are automated commercial phone systems, call-routing, and dictation. Specifically to translation, the ALPAC report (Pierce et al., 1966) mentioned that ‘productivity of human translators might be as much as four times higher when dictating’ (see ‘ASR-MT Combination’).

Speech models

Speech models in ASR are categorized into language and acoustic models. On the one side, language models help a speech recognizer figure out what constitutes a possible word, what words are likely to co-occur, and how likely a word sequence is (see Jurafsky & Martin, 2010: chapter 4, Manning & Schütze, 2009: chapter 6). Microsoft Research[2] characteristically describes a language model as the model that tries to make the right guess when two different sentences, such as *It's fun to recognize speech?* and *It's fun to wreck a nice beach?*, sound the same. A statistical language model assigns a probability to a sequence of words by means of a probability distribution. It should be pointed out that language modeling is not related only to speech recognition, but also MT, PoS-tagging and Information Retrieval (IR). Both unigram and n-gram language models are used in IR to calculate the probability of recognizing one or more words in a corpus. The *SRI Language Modeling Toolkit*[3] is a toolkit for building and applying statistical language models.

On the other side, acoustic models 'compile' audio speech recordings and their transcriptions into statistical representations of the sounds that make up each word. Thus in order to create an acoustic model, necessary resources are i) audio recordings of speech, ii) their text transcription, and iii) a compiler to create statistical representations of sounds. Acoustic models are not related only with acoustics and phonetics, but also with gender and dialect difference among speakers.

History

As far as the history of speech recognition is concerned, in the early 1960s IBM developed and demonstrated *Shoebox*, a forerunner of today's voice recognition systems. This innovative device recognized and responded to 16 spoken words, including the ten digits from "0" through "9". *Shoebox* was operated by speaking into a microphone, which converted voice sounds into electrical impulses. Today IBM offers various products[4] to connect mobile and speech middleware. To give two examples, *IBM EmbeddedViaVoice*[5] technology can be used for small mobile devices and automobile components. Some of its features are that it recognizes more than 200,000 spoken words across 14 languages. *WebSphereVoice Server*[6] includes ASR, text-to-speech (TTS) software and speaker verification, and can be integrated with other software and hardware telephony products.

In 1984 *Apricot Portable Computer* was the first PC with built-in speech recognition, licensed from Dragon Systems, Inc. The company Dragon Systems was founded in 1982 and its technology was acquired in 2005 by Nuance Communications. *Dragon NaturallySpeeking*[7] ASR system works today with most Windows applications, has *Nuance* Text-to-Speech (TTS) technology, and reaches up to 99% ASR accuracy[8]. The latest features, as of version 11, show more intuitive user interface (elimination of distractions),improved accuracy (15% greater than previous version), and faster performance. Another product from Nuance Communications is *Dragon Dictate* which is used currently by Macintosh. Earlier Macintosh recognizers were by MacSpeech, a company acquired by Nuance Communications in 2010.

In the mid-1990s, Microsoftdevoted resources to speech and language processing. Today *Microsoft Tellme*[9] speech technologies enable speech interaction with PC, phone, car, and TV. Microsoft Windows's ASR system can be used in command and dictation. Its Speech Application Programming Interface (SAPI) allows the use of speech recognition and synthesis within Windows applications; the latest version 5.4 ships in Windows 7.

Current tools

A supplier exclusively of speech technologies for the automotive and mobile industries is SVOX[10]. It is worth mentioning that Google uses SVOX to assist with pronunciation on its web services Google Translate and Google Dictionary[11]. LumenVox[12] is another provider of ASR software that allows users to control their telephony or computer application using their voice. Some FOSS ASR tools follow:

- *CMU Sphinx*[13]: this tool is developed by the Carnegie Mellon University (CMU). The Sphinx-4 framework includes three primary modules: *FrontEnd*, *Decoder*, and *Linguist*. The *FrontEnd* comprises one or more parallel chains of replaceable communicating signal processing modules, so that it takes one or more input signals and parameterizes them into a sequence of *Features*. The *Linguist* generates the SearchGraph that is used by the decoder during the search; it translates any type of standard language model, along with pronunciation information from the dictionaryand structural information from one or more sets of acoustic models. The language model module of the *Linguist* provides word-level language structure, while the acoustic model module provides a mapping between a unit of speech and a hidden Markov models (HMM) that can be scored against incoming features provided by the *FrontEnd*. The *SearchManager* in the Decoder uses the *Features* from the *FrontEnd* and the SearchGraph from the *Linguist* to perform the actual decoding, generating *Results* (Walker et al., 2004).
- *Hidden Markov Toolkit*[14] (HTK): it is a portable toolkit for building and manipulating HMMs. HMMs are used in ASR and synthesis to aid, among others, in disambiguating homographs. The first version of HTK was in 1989 by Steve Young.
- *Julius*[15]; Julius(Lee & Kawahara, 2009) has been developed as a research software for Japanese large vocabulary continuous speech recognition (LVCSR); currently is developed by Nagoya Institute of Technology. It is a decoder based on word N-gram and context-dependent HMM. To run ASR with *Julius* one should prepare a language and an acoustic model. It offers, though free Japanese and English language and acoustic models.
- *Simon*[16]: it has been developed since 2007 by ‘simon listens’, a non-profit organization for research and apprenticeship. It does not ship with any speech models, but it provides an end-user interface to create them. HTK then compiles the speech model. *Simon* can be used with all languages and dialects and one can even mix languages. The current release can be used to set up command-and-control solutions especially suitable for disabled people.

Other FOSS of speech recognition can be found at other webpages as well[17].

Last but not least, another research field of ASR regards the development of silent, subvocal speech recognition. In 2004 NASA Ames Research Center[18] found that small, button-sized sensors, stuck under the chin and on either side of the ‘Adam's apple’ could gather nerve signals, and send them to a processor and then to a computer program that translates them into words. As application areas they suggest spacesuits, noisy places like airport towers, or even traditional voice-recognition programs to increase accuracy.

Machine Translation

The definition of Machine Translation (MT) is the automatic translation of text or speech from one natural language (source language) into another (target language) by means of computer software. In 1955 the first MT experiment between Georgetown University and IBM took place; it translated 60 sentences from Russian into English. In 1966 the ALPAC report (Pierce et al., 1966) stated that “there is no immediate or predictable prospect of useful MT”. Although the report’s impact was that the USA government reduced its funding for MT, the research projects in the USA were extended. The MT systems SYSTRAN (1970-currently) in the USA and EUROTRA (1978-1992) in Europe were some of the first MT successful initiatives. Rule-based MT (RBMT) approaches were the first approaches, whereas today the high availability of mono, bi- and multilingual corpora allow for statistical MT (SMT) and example-based MT (Nagao, 1984). Today MT systems typically combine RBMT with SMT. Through the years MT systems have been faster and with higher accuracy. Lately many FOSS MT tools have made their appearance.

Current tools

There is a plethora of MT systems nowadays, both proprietary and open-source. A good distinction between these two can be found in Forcada (2006: 2). Some of the commercial MT systems today are SYSTRAN[19] (earlier RBMT – now hybrid), PROMT[20] (earlier RBMT – now hybrid), SDL Language Weaver[21] (SMT), AppTek TranSphere[22](RBMT), IBM WebSphere Translation Server[23](hybrid).

As for FOSS MT tools, we distinguish between three categories of free MT:

- Open-source MT systems[24]:
  - o *Apertium*[25]: a RBMT system;
  - o *Cunei*[26]: an EBMT system;
  - o *Joshua*[27]: an SMT system;
  - o *Moses*[28]: an SMT system, specifically decoder;
  - o *OpenLogos*[29]: the open source version ofthe commercial MT system LOGOS;
  - o *OpenMaTreX*[30]: an EBMT system based on the marker hypothesis (Dandapat et al. 2010).
- Online translators of proprietary MT systems;
  - o PROMT;
  - o SYSTRAN.
- Online MT services:
  - o *Caitra*[31] (Koehn, 2009): this tool offers the possibility to the user to select the best candidates and postedit the MT output.

- Google *Translate Toolkit*[32]: it is a free translation service that provides instant translations of words, sentences, and webpages between 57 different languages; it is an SMT approach.
- Microsoft® *BingTranslator*[33]: it offers an online MT service currently supporting 35 languages. It also has a "Translate-and-Speak", a TTS functionality.
- SDL *FreeTranslation*[34]: different language combinations are supported by different companies, such as SDL Enterprise Server, PROMT, MTLabs, and Language Weaver.
- Yahoo *Babelfish – Translator*[35]: its technology is based on SYSTRAN.

ASR-MT Combination

SMT is the approach of MT systems used in the ASR-MT combination, because SMT distributes probabilities and provides hints to the ASR system. Also, confusion network decoding has been the most successful approach in combining outputs from multiple MTs (Rosti et al. 2008).

The combination of ASR-SMT systems has two usages: i) translation dictation and ii) spoken language translation. In the former case, the MT system provides hints to the ASR system as to what the translator is likely to utter when translating a source text, while in the latter there is spoken audio input in source language (SL) and transformation to written text or spoken audio in target language (TL). One of the first efforts about ASR-SMT combination was by Brown et al. (1996), while others like Paulik et al. (2005), Khadivi et al. (2006), Reddy et al. (2007) followed. We particularly look at three more recent approaches:

- Désilets et al. (2008) evaluated productivity gains of hybrid ASR-SMT systems for translation dictation. They conducted an experiment with eight professional translators dictating into French using Dragon *NaturallySpeaking* 8. They did not find productivity gain at a baseline 11.7% word error rate (WER), however, they found that dictation with better ASR systems with WER of 4% or less, would experience statistically significant productivity gains of 25.1% – 44.9% Translated Words per Minute.
- Gales et al. (2007) tried out two different approaches of speech to text (STT) system combination and its impact on MT performance: i) hypothesis combination (alignment of hypotheses against a 'base' hypothesis and conversion into a consensus network) and ii) cross adaptation (hypotheses are obtained from the output of other STT systems). Although hypothesis combination (i) gave lowest error rates, the use of cross-adaptation was found to be a "safer combination scheme for translation" (p.1280).
- Rosti et al. (2008) described an incremental alignment method to build confusion networks based on the translation edit rate (TER) algorithm. They used a confusion network as the reference. The algorithm finds the minimum edit distance between the hypothesis and the reference by considering all word arcs

between two consecutive nodes in the reference as possible matches for a hypothesis word at that position. Then shifts of blocks of words, which have an exact match somewhere else in the network, are tried in order to find a new hypothesis word order with a lower TER.Karakos and Khudanpur (2008) extended the algorithm of Rosti et al. (2008) and concluded that systems combined should be of comparable quality to have gains. In 2009 Rosti et al. proposed flexible matching using WordNet (Fellbaum, 1998)to find all possible synonyms and words with identical stems in a set of hypotheses.

Speech synthesis

Speech synthesis or text-to-speech (TTS) is the production of speech (acoustic waveforms) from text (Jurafsky & Martin: 2009: 249). Firstly, the input text is converted in a phonemic internal representation (i.e. text analysis) and then this internal representation is converted into a waveform (i.e. waveform synthesis). Von Kempelen built between 1760 and 1790 the first full sentence speech synthesizer (not the chess-playing hoax *Mechanical Turk*). Since the early 1980s many computer operating systems have included speech synthesizers. TTS applications are mainly used in hands and/or eyes-busy situations, such as automobile navigation. Apart from that, synthesized speech helps also people with visual or speech impairments. TTS is used, for example, by blind people through screen readers, software applications that interpret what is being displayed on the screen. In addition, TTS can be used in education to help foreign language learning.

TTS technology can be categorized into concatenative and formant. The former means that the synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Formant synthesis does not use human speech samples at runtime, but uses additive synthesis and an acoustic model to create artificial speech. In the next paragraphs we refer exclusively to some speech TTS software; for those tools who include both speech recognition and synthesis, look at previous chapter “Automatic Speech Recognition-Current tools”.

Current tools

Proprietary TTS software includes:

- *Cepstral*[36]: it offers synthetic voices; in demo voices not only the rate (fast/slow speech), but also the pitch (high/low) and the effect, like *Split Personality* or *Spacetime Echo* can be selected.
- *Natural Reader*[37]: it has natural sounding voices and can convert written text into audio files such as MP3 or WAV; it offers free versions for Windows and Mac users, but there are paid versions with more free natural voices available, pronunciation editor, and conversation control.
- *SpokenText*[38]: it offers a trial for 7 days and the supported download format is MP3 only, while in the paid versions it is Audio Book and Multiple MP3 as well.

- *Talking Clipboard*[39]: it uses natural sounding synthetic voices (SAPI 5 compliant) and can convert them to MP3 or WAV audio files; it has 30-day free trial.

Apart from proprietary TTS software, many FOSS TTS tools exist and some of them follow below:

- *eSpeak*[40]: it uses a "formant synthesis" method; this allows many languages to be provided in a small size.
- *Festival*[41]: it is developed by the University of Edinburgh and offers a general framework for building speech synthesis systems as well as including examples of various modules. *Festival* currently supports British and American English and Spanish.
- *Festvox*[42]: this project is part of the work at CMU's speech group; it is firmly grounded within *Festival* and *Flite* (see below).
- *Flite*[43]: Flite (festival-lite) is a small, run-time synthesis engine developed at CMU and primarily designed for small embedded machines and/or large servers. It is designed as an alternative synthesis engine to *Festival* for voices built using the *FestVox*[44] suite of voice building tools.
- *FreeTTS*[45]: it is based upon *Flite*. Some possible uses of *FreeTTS* are Java Speech API JSAPI 1.0 synthesizer, remote TTS Server that can work with a speech/telephony system, desktop TTS engine, or downloadable web application.*FreeTTS* provides support to import voice data directly from *FestVox*.
- *Mary*[46]: it was originally developed as a collaborative project of Language Technology lab from the German Research Center for Artificial Intelligence (DFKI) and the Institute of Phonetics at Saarland University. Version 4.3 supports German, British and American English, Telugu, Turkish, and Russian. It comes with toolkits for quickly adding support for new languages and for building unit selection and HMM-based synthetic voices.
- *YAKiToMe!*[47]: this is a web service where one can copy and paste text into a box, select a topic, and the text will be converted to audio. Then the Podcast Library page displays how long users have to wait before their audio file is ready for listening. *YAKiToMe!* gives free access to AT&T Natural Voices(R) and Windows voices. MT with Babelfish is also possible as a potential next step.

It should be pointed out that various companies sell different voices for different locales. Some examples are A&T Natural Voices, Nuance Real Speak and so on. Instead of buying voices, as aforementioned, building new voices is available through CMU's *FestVox* project. Voices can also be personalized: in EMIME project Kurimo et al. (2010) used the voice of the speaker in the input language to utter the translated sentences in the output language (see following “Speech-to-Speech Translation”).

Speech-to-Speech Translation

The European Language Resources Association (ELRA) defines the goal of Speech-to-Speech Translation (SST) as follows:

> The goal of the Speech-to-Speech Translation (SST) is to enable real-time, interpersonal communication via natural spoken language for people who do not share a common language. It aims at translating a speech signal in a source language into another speech signal in a target language.

Speech-to-speech translation (SST) systems have been developed mostly for small hand-held mobile devices to facilitate spoken communication between people speaking different languages. Some research projects which developed SST systems are TC-STAR (2004-2007), LC-STAR (2002-2005), NESPOLE! (2000-2002), TONGUES (2000-2002), and Verbmobil (1996-2000). In addition, the research system MATRIX (1998) and the commercial NEC (2005) in Japan, and Digital Olympics (2006) in China were successful speech translation systems. Also, the Speechlator (part of BABYLON project, 2003) was a two-way system working with Arabic as TL on a consumer PDA.

More recently, some tools which offer SST systems are the following:

- *Google's Conversation Mode*[48]*:* in 2010 it was announced that in Android operating systems there will be a conversation mode, where users speak, *Google Translate* translates their speech, and reads the translation out loud (currently that is available for English-Spanish). Google states that because this technology is in alpha, factors like regional accents, background noise, or rapid speech might be possible limitations. Google develops software for translator phone[49] by analyzing chunks of speech, and translating them in their entirety.
- *Jibbigo*[50]*:* it is an SST application for mobile devices. Its available language pairs are Chinese-English, French-English, German-English, Iraqi Arabic-English, Japanese-English, Korean-English, Spanish-English, and Tagalog-English. *Jibbigo* licenses theTTS from the company SVOX (see ‚Automatic Speech Recognition'). It was built on two decades of the scientific research in speech and language processing at *Mobile Technologies*. Mobile Technologies and Jibbigo maintain a strong research collaboration with *InterACT*, the International Center for Advanced Communication Technologies at Carnegie Mellon University and Karlsruhe Institute of Technology, Germany.
- *Phraselator*[51]*:* it is developed by Voxtec for use by the U.S. Military ; it has aided the needs of thousands of service personnel in Iraq, Afghanistan, and Southeast Asia.
- *SpeechTrans:* AppTek's system is deployed on computers, wearable machines, and telephony servers[52]; the input is recorded through a telephone channel or microphone, and recognized using different ASR systems and tuned to either microphone- or telephone-quality speech. The recognized utterances are normalized using SMT based on finite state automata. The output is then translated by a hybrid MT (RBMT and SMT).
- *Trippo VoiceMagix*[53]: it uses speech technology by Nuance Communications; Speech input is US English and supports 27 languages.

Dialog systems

Dialog systems are widely used to decrease human workload in telephony applications, for example in call centres. The functionality provided by dialog systems is known as interactive voice response (IVR).

A speech dialog system consists of three components: speech input, internal processing, and speech output. The input has two functions: ASR and language understanding. The internal processing includes dialog management, while the speech output includes rendering and response. More precisely, a typical[54] dialog system workflow contains the following stages:

- ASR or gesture recognition.
- Text analysis (tagging and parsing) by a Natural Language Understanding unit.
- Semantic analysis by the dialog manager, along with a task manager that has knowledge of the specific task domain. When the input is malformed or inaccurate, the manager has to decide if it is necessary to request a more specific user input or if it is possible to extract the missing accuracy from knowledge resources (Will, 2010).
- Output generation by the dialog manager (language and/or gesture generation).
- Output rendering using an output renderer, which may include TTS engine.

As far as the quality of dialog systems is concerned, Will (2010: 60) states that the main cause of failure in speech dialogs is the difficulty of recognizing context-sensitive information in full sentence utterances. Context-sensitive information is manageable by consulting knowledge sources which are divided into dialog history and task record.

As seen in the stages i. and iv. above, dialog systems have also employed other modes for communication for input and output, such as gestures, and we will examine that in the next chapter 'Gesture Recognition and Localization'.

Gesture Recognition and Localization

According to McNeil (1992), gestures are the movements of the hands and arms that we see when people talk. The meanings of 'gesture' as an action performed to convey a feeling or intention, or an action performed for show in the knowledge that it will have no effect are not considered here. However, for our purposes we do not focus only on hands and arms movements, but we include eye tracking as another part of gesture recognition as well. Gestures is a mode of communication in multimodal systems and they emphasize or express more accurately the content of the spoken language. Gestures are used, among others, when verbal communication is not possible or limited, i.e. in road works (very noisy) or airfield. Moreover, older people are not very familiar with interaction with mouse and keyboard or small buttons in the remote controller and that is why they prefer gesture interaction (Burkhardt et al. 2011).

McNeil’s (1992) categorization of gestures is the following:

- *Iconics:* they bear a close formal relationship to the semantic content of speech.
- *Metaphorics:* they are pictorial like iconics but the pictorial content presents an abstract idea rather than a concrete object or event.
- *Beats:* they are so named because they look like beating musical time. They always have the same form regardless of the content.

As for speech-gesture combination, McNeil (1992) points out that speech and gesture must cooperate to express a person’s meaning. Although they are closely linked, he refers to some of their distinct differences. Gestures are global because the direction is from whole to part; in language instead, the words are parts and when they are combined, they create a whole. Gestures are also synthetic because one gesture can combine many meanings, while in language the relationship of words to meaning is analytic, that means that distinct meanings are attached to distinct words. One interesting observation about the syntactic relationship between speech and co-speech gestures in multimodal grammar is explored by Fricke (2009). She claims that co-speech gestures are able to instantiate the syntactic function of an attribute within verbal nominal phrases and can semantically modify the nucleus of the nominal phrase. Speech-gesture alignment is being researched currently, among others, by the University of Bielefeld (project SFB 673-B1[55]).

Moving forward from a theoretical view of gestures to gesture-based software, we will refer to some gesture recognition software. It can be used for various purposes, such as to transcribe the symbols represented through gestures into text, help with sensors in patient rehabilitation, identify pointing/spatial instructions, control interactions within video games (see Xbox 360-Kinect), or domotic appliance devices, and so on. Also facial gestures and eye tracking may control cursor motion or focus on elements of a display. Eye tracking particularly in the field of translation has been researched by O’Brien (2010).

Current tools

Some 3D gesture recognition software and their main functions are presented below:

- *Android Gesture Recognition Tool* (Neßelrath et al. 2011): it allows recording hand movement gestures by exploiting the accelerometers of an Android smart phone. It makes it possible to create gesture training sets.
- *CamSpace*[56]: this platform can detect human gestures and turns everyday products (like can, bottles, boxes, etc) or objects into computer controllers that can operate new or existing games and applications. Use for educational games or advertizing lets advertisers and advertising agencies create motion games and experiences based around advertisers' products and user hand gestures.
- *GestureTek*[57]: it is a custom 3D Depth Sensing Prototype System for Gesture Control.

- *HandVu*[58]: vision-based Hand Gesture Recognition and User Interface (UI); it detects the hand in a standard posture, then tracks it and recognizes key postures – all in real-time and without the need for camera or user calibration.
- *IISU* by SoftKinetic[59]: enabled with a depth sensing imager, it allows end-users to watch their video images or full-body 3D avatar while interacting in real time with computer-generated characters and devices. Compatible with all major 3D depth-sensing devices, *IISU* is available now to developers of interactive digital entertainment, serious games, interactive marketing solutions, and consumer electronics applications.
- *Throw and Tilt* (Dachselt & Buchholz, 2009) combines sensor-enabled, gesture-based mobile phone interaction with large displays. More precisely, throwing gestures can be used to transfer personal media data such as music, pictures, and map locations to a large screen. An even more advanced way is to seamlessly transfer the whole UI between various devices.

A full list of gesture recognition software can be found at various webpages[60].

Gesture Localization

We turn our attention now to gesture localization and how it is related to gesture recognition. In our opinion, even if gesture recognition is performed with high accuracy, the meaning might be not understood by some users, as this gesture is not used as such in a specific locale. In addition, users would prefer some gesture recognition software against some other for the reason that they want to do only gestures which are 'semantically allowed' in their culture and hence are not offending in another locale.

Some examples of different gestures in different locales follow: the 'thumbs up'[61] gesture means that everything is good in America and many European countries, whereas it is rude in Asian and Islamic countries. Men holding hands or kissing is considered normal in some Asian and Islamic countries, while it would possibly mean homosexuality in most westernized countries. Also, the curled finger gesture[62] means telling someone to come to you in America and England, but is considered rude in Japan and death in Singapore. Prolonged eye contacted is considered rude in Japan, Korea, and Thailand, while it is preferred in North America and Europe[63]. It is worth mentioning that some gestures are sex-dependent, i.e. men and women do a different gesture to express the same meaning. Examples of sex-dependent gestures are "thank you" and "goodbye" in Zimbabwe. The International Society for Gesture Studies (ISGS) investigates, among others, the roles and organization of gesture in face-to-face conversation, universal and cultural aspects of gesture, and gesture's relationship to thought and language. For more information about gesture recognition for culture specific interactions see Rehm et al. (2008).

After showing some culture-specific gestures, we define gesture localization as follows:

> Gesture localization is the adaptation of gestures to a target locale in order to transfer the same meaning as in the original locale.

Gestures need localization, i.e. being interpreted and accordingly, recognized by software because different gestures exist in different locales and the same gesture might have a different meaning in different locales. Gesture localization should be always taken into consideration when gesture recognition software is developed, so that the software can be used by other locales at least to the same extent as in the original locale. The curled finger gesture, for instance, to tell a robot to navigate to the user, would have not found any acceptance by users in Singapore.

Spatial gestures

Co-speech spatial gestures appeal to our research particularly, as our application area is an ambient assisted living lab and a wheelchair which is navigated through speech interaction. These gestures usually make the navigational intention of the user more perspicuous. Breslow et al. (2010) distinguish spatial gestures into gestures used for thinking (cognitive gestures) and gestures used to help express predetermined ideas (linguistic gestures). The former help to determine what to say in a spatially complex domain, while the latter what we have determined to say.

Nickel and Stiefelhagen (2006) recognized with their real-time vision system pointing gestures in human-robot interaction by visual tracking of head, hands, and head orientation. Based on hands' motion, they decomposed a pointing gesture into three distinct phases and modeled each phase with a dedicated HMM. They followed a multi-hypotheses approach and achieved 60% relative error reduction.

Burkhardt et al. (2011) described some intuitive and easy gestures for the elderly. They made an experiment where test subjects were asked to make gestures with the WiiMote controller to scroll (right, left, up, down), zoom, renew, and navigate in a relational database. The results showed that the more complex the tasks are, the more (in number) and more intuitive the gestures became. This, in turn, means bigger effort in implementation or training.

Speech and Gesture in Assisted Living

In this chapter we focus on ambient assisted living (AAL) environments and related work regarding dialog and gesture-based systems in AAL. Then we refer specifically to the Bremen Ambient Assisted Living Lab (BAALL[64]) and its human-robot speech interaction.

AAL is concerned with intelligent assistant systems of assistance for a better, healthier and safer life in the preferred living environment through the use of Information and Communication Technologies (ICT). Human-computer interaction (HCI) is indispensable part in ICT and speech recognition and synthesis are intuitive realizations of HCI. There have been many research initiatives and advances in AAL environment the last years, for example, AALIANCE project[65], AAL Joint Programme[66] and so on. Goetze et al. (2010) mention characteristically: "older persons targeted by AAL technologies especially need more easy-to-use methods to interact with inherently complex supporting technology".

One motivation for AAL initiatives is the demographic change due to the aging population phenomenon. According to the German Federal Statistical Office[67], in 2060 14.2% will be 80 years old or older. Also, almost double amount of people will be retired. The ever more increasing amount of older people who need care makes the need for a safer, healthier, and more independent lifestyle essential. The aging phenomenon is reality and technology should help increase the autonomy of the elderly and their active participation in society.

To start with, in 2002 Theobalt et al. stated that "situations humans helping mobile robots to find their way or to complete tasks while engaging in a dialog are expected to become more widespread as robots begin to appear in domestic environment". We present following recent advances which prove this expectation true:

- Ivanecky et al. (2011) argue for using mobile phones specifically to control home devices, first because the recorded speech signal has good quality, as the mobile phone is acting as a close-talk microphone and second, the set of the commands for the house control is relatively small (usually around 50). They made an experiment installing the entire system into real houses and asking the users to talk as they wish (out of grammar). The results showed action accuracy 91.23%, taken into account also that almost 30% of utterances were spoken by a non-native speaker.
- Krieg-Brückner built at DFKI, University of Bremen, the BAALL which is an apartment suitable for the elderly and people with physical or cognitive impairments (see Krieg-Brückner et al., 2010). A wheelchair called "Rolland" serves mobility assistance in the BAALL. Rolland is equipped with two laser range-sensors, wheel encoders, and an onboard computer. In BAALL natural interaction with the users is taken in serious consideration, and thus it is enabled through special devices for compensating special limitations.
- Goetze et al. (2010) described technologies for acoustic user interaction in AAL scenarios. They designed and evaluated a multi-media reminding and calendar system as a part of a personal activity and household assistant for acoustic sound pick-up, processing, enhancement, and analysis providing functionality for acoustic input and output of assistive systems. The authors examined whether users prefer a suggested structured dialog or a free input of speech. In the latter case, the participants were asked to provide input commands to the system without a structured dialog. The free input of appointments was preferred by 58%, the structured by the rest. On the one side, some reasons for preferring free input are i) input is more familiar, ii) higher flexibility, iii) more individual, and iv) less complicated. On the other side, reasons preferring the structured dialog are that i) nothing can be forgotten, ii) higher concentration on the information, and iii) easier communication with the technical system (p.18). Apart from that, the authors carried out an ASR performance study having as training set both male and female speakers of different age and hearing loss. The results showed that the ASR performance was lower for the older persons and for female (p.25).

- Becker et al. (2009) deployed sensors into an assistive environment. They state that "the speech interface is the easiest way for the user to interact with the computer based service system". The dialog management system they implemented uses the ASR engine *Sphinx* combined with the *Cepstral* and *Festival* speech synthesis. The authors' reason of selecting *Sphinx* was the requirement lack of the speaker to 'train' the system. The problem they faced in their experiments was that the longer and more continuous the speech is, the more recognition errors there are. Thus they concluded that short and distinct phrases help to improve the precision of the speech recognition.
- D'Andrea et al. (2009) described a multimodal pervasive framework based specifically at the grammar level and developed a methodology for defining a formal grammar and inductive mechanisms to generate rules for synthesizing grammars. They envisaged four architectural levels: acquisition, analysis, planning, and activation level. In the analysis level, there is a speech recognizer, gesture recognizer, and speech synthesizer. The multimodal input is parsed according to rules included in the Multimodal Grammar Repository.

Now we present some work on dialog with relation to robotics but not in assistive environments.

- Motallebipour & Bering (2003) developed a prototype to study integration of speech dialog into graphical interfaces. Their goal was to make a robot able to understand spoken language instructions and perform simple tasks. The instructions were used within a restricted domain and that had as benefits that the speech vocabulary and the number of natural sentences are limited and the prototype can be integrated into existing (computer-aided design) CAD software. They used ASR application, Action Logic application, TTS application, XEmacs application, and 3D Robot application. The ASR application they used is Microsoft SAPI 5.1 in command mode. The command node used Context Free Grammar (CFG) grammars to recognize single words and short phrases. The CFG format in SAPI 5 defines the structure of grammars and grammar rules using XML. The authors concluded that dialog sentences by three non-native English speakers were recognized with good accuracy using SAPI 5 and that although the grammar and set of used words were limited, the test subjects felt that the dialog came natural. Here it should be pointed out that Rosenfeld et al. (2001) also mentioned that constraining language is a plausible method of improving recognition accuracy. SAPI 5.1 was also used by Haage et al. (2002) to develop a speech system to design robot trajectories that would fit with CAD paradigms.
- Mubin et al. (2010) developed ROILA (Robot Interaction Language) in order to improve the accuracy of speech recognition and to make it learnable for a user. Initially based on Toki Pona, they conducted a i) phonological and ii) morphological overview of natural languages in order to create ROILA which consists of 16 letters, four parts-of-speech and four pronouns. Their results are higher accuracy compared to English for a relatively larger vocabulary, although the acoustic model of *Sphinx* is primarily designed for English.

- Theobalt et al. (2002) developed *Godot*, a mobile robot platform for the interface between a sophisticated low level robot navigation and symbolic high-level spoken dialog system. The dialog component used Discourse Representation Structures. They used the off-the-shelf *Nuance 7.0* speech recognizer. The grammar they used is compiled from a linguistic unification grammar and includes semantic representation. They used the speech synthesizer *Festival*.

We draw some conclusions from some of the above related work:

- Speech vocabulary and training set of words/sentences (corpus) are limited, but give less error rates;
- Long and continuous sentences are difficult to be recognized;
- Free input of speech rather than structured dialog is preferred by test subjects.

Now we present two studies showing how important is multilinguality in AAL. Undoubtedly, translating vital information can often save lives (see Anastasiou & Schäler, 2010). Translation in medical domain is crucial and as telemedicine is one aspect of AAL, that shows that multilingual support in AAL environment is indispensable. Burda (2005) made an experiment with twenty native speakers of English, ages 62 to 91, who listened to words and sentences produced by native speakers of English, Taiwanese, and Spanish. Participants transcribed the words and sentences and rated speakers' comprehensibility and accentedness using separate 7-point Likert-type scales. Listeners performed the most poorly on items spoken by the native Spanish speaker. A report by Santo Pietro and Ostuni (1997) denoted that at least 30-40% of direct care staff in health care settings are from backgrounds other than native English-speaking Euro-American, while 90% of the residents are native English-speaking Euro-American. Lack of information in a local language or misinterpretation of care staff instructions can lead to a wrong dosage of medicine which can have dramatic impact. Hence translation plays, apart from a communicative and socioeconomic, also a life-saving role in AAL.

Current Dialog in the Bremen Ambient Assisted Living Lab

In Bremen Ambient Assisted Living Lab (BAALL) natural interaction with the users is enabled through spoken dialog with special devices. However, the dialog is currently performed only in German and English. Having this as motivation, we want to move from a monolingual setting to a multilingual one. BAALL is the environment and Rolland the implementation device of our designed platform.

The user can interact through speech with Rolland in BAALL to navigate through the living areas. Rolland's navigation assistant and natural language interaction technology is integrated, as Krieg-Brückner et al. (2010) discussed in the following case scenario: Mario is sitting in his wheelchair at the desk of his home office. He says to Rolland 'I'd like to eat a pizza'. The wheelchair reaches the kitchen and replies to the user 'OK, I m going to take you to the kitchen'. When Mario is in the kitchen, he asks 'Where is the pizza?', Rolland replies 'The pizza is in the fridge. I am taking you to the fridge'. A snippet of the grammar that enables (part of) the above scenario follows:

```
<ESSEN>= <pizza object> |
    a <pizza object>;
<TASK>= I want to eat <ESSEN> |
    I would like to eat <ESSEN>;
```

At the moment the grammar is available in German and English and Rolland can speak back in German, English, and (some) Italian.

The advances of speech interaction with Rolland in BAALL are the following:

- Users with physical impairments can speak instead of clicking on buttons;
- Users feel friendlier when speaking to Rolland;
- Rolland can intelligently make second steps (connecting pizza with kitchen) saving time and sparing tiresome single instructions.

The limitations we see in Rolland and BAALL are the following:

- The ASR system used is proprietary.
- Grammar is in German and English only; that means that users can speak only in these two languages.
- Grammar is minimal; thus the instructions limited.
- Rolland is not intelligent enough to remember his original position so that he returns back or follow a sequence of events (e.g. after Mario finishes his pizza he wants to wash his hands, so Rolland brings him to the bathroom).

Our Experiment

We believe that when users express themselves in their preferred natural language/mother tongue and devices/robots act after recognizing this natural language is not only easy, but also intuitive, and natural method of HCI. Users feel more secure and friendlier towards intelligent machines.

In order to mitigate the limitation ii. we mentioned above, we develop a platform where ASR, MT, and TTS systems are combined. It is a distinct SST system, as the difference between our designed SST system and other traditional SST systems is that the output is in the source language (usually mother tongue of the user) and not in the target language. Grammar minimalism is another limitation (iii) and this can be mitigated through addition of more tasks, apart from navigational ones. Controlling devices, like turning the TV on, for example, is another application where the SST system can be used. Making the grammar lexically rich (with morphological and syntactic extensions) will lead to a wider spectrum of tasks Rolland can undertake. These tasks can then be well structured in activity-based ontologies, so that complex activities (limitation iv.), but also their sequence can be well managed/tracked and controlled. In the next paragraphs we focus only on multilinguality provided by our platform; the other two items are outside the scope of this paper.

Our platform combines ASR, MT, and TTS tools. This unique platform is initially tested in the wheelchair Rolland and later can be implemented in intelligent devices in assisted living environments, such as beds, kitchen drawers, and electronic appliances.

The workflow of the HCI using the specific platform is the following:

- User speaks and/or makes a co-speech gesture, and speech and gestures are converted into text;
- Language identification and MT technology to translate the existing grammar in the language of the user;
- Rolland and intelligent devices act based on the user's input;
- Machine speaks back into the user’s language.

The advantage of multilingual support in BAALL is that users can speak in their own mother language; that means that Rolland can be used in a crosslingual setting, not only by German or English native speakers or speakers who can speak German/English. Advantages related to sentiments like speech naturalness, freedom, and security are counted too.

Discussion and Conclusion

In this paper we presented the state-of-the-art of automatic speech recognition and speech synthesis, dialog systems, and speech-to-speech translation. Gesture examples, recognition software, and meaning of gesture localization were also furnished. Then we referred to ambient assisted living environments and relevant related work considering speech and gesture support.

The elderly and people with physical or cognitive impairments is a group which needs special care and specific, easy-to-use technology should be developed to meet their needs. The elderly suffer often from hearing loss, speak with pauses or unclearly, and typically have impaired vision. All these factors limit the usage of the technology compared to other people. They cannot read the text written under buttons in TV remote controllers, for instance. Although, there are remote controllers designed today for the elderly with bigger buttons and text in bigger font, speech recognition software would facilitate their interaction. They can say, for example, the command to the TV ‘turn on the TV’. The challenge for ASR systems here would be the prosody of the elderly. ASR software works usually with clear voice and short and distinct sentences which is often not the case with elderly’s speech. Thus in our future prospects belongs to train acoustic models with voices of older people; we will use *VoxForge* project for this purpose.

In addition, gestures can be part of human-computer interaction. Making a gesture like showing with the hand to the right, Rolland will turn right; also a ‘tick in the air’ could turn the TV on. In controlling devices particularly, the limitation is that users have to learn specific gestures and software has to be trained to recognize them. We saw that the more complex the tasks become, the more intuitive the gestures. Moreover, the cultural background of gestures should be taken into account when developing or adapting recognition software. We plan to conduct a survey regarding how people in different locales make gestures for various purposes (spatial or device controlling).

In this paper the current speech processing in BAALL was described together with its limitations which are lock-in propriety software, monolinguality, grammar minimalism, and activity simplicity. Our contributions to eliminate the limitations are usage of FOSS, a unique speech-to-speech translation system, lexical grammar enrichment, and activity-based ontologies. Against monolinguality specifically, we designed a unique platform combining FOSS ASR, MT, and TTS systems. Its distinction from a typical SST system is that the intelligent device does not speak in the language the input is translated into, but in the user's language. MT is used only to translate the existing grammar in the user's language. Our speech translation platform is an effort towards provision of multilingual support in intelligent environments.

Future prospects concerning our platform include testing various FOSS STT, MT, and TTS systems to select the best one which fits our needs. After that, we plan to follow the system combination of ASR systems to see whether they perform with higher accuracy than when used individually. We will follow the cross-adaptation approach, as Gales et al. (2008) found this more appropriate.

Multimodality and multilinguality are two different aspects in AAL which can be combined though to make not only human-human but also human-computer interaction efficient. In AAL devices should be reachable, navigation should happen without crashes, and all this at high speed and low cost. Ivanecky et al. (2011) stated that the current ASR systems in AAL are either reliable, but complex and very expensive, or inexpensive, but unreliable. As far as speed is concerned, turning the TV on through speech or gesture should not take longer than clicking on a button.

Multilingual support in human-computer/robot interaction in AAL facilitates not only independence, intuition, and user-friendliness, but is often necessary to avoid dramatic medical accidents in case of language misunderstanding between care staff and patients.

Navigation of people on a wheelchair through speech interaction, sensor-based frameworks, speaking calendars, showing a website in a large display by throwing gestures and so on are today reality and not research goals any longer. Technology is advancing, but it should be made affordable and applicable in many domains, including AAL. We hope for future research initiatives and projects developing multimodal and multilingual applications in assistive environments. We presented an initial contribution of a distinct speech-to-speech translation system with the goal of a language-independent dialog management tied with localized co-speech gestures initially applied on a wheelchair and later on other devices in ambient assisted living environments.

We gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center SFB/TR 8 Spatial Cognition - Subproject I3-SharC.

Bibliography

Anastasiou, D., Schäler, R. (2009), « Translating Vital Information: Localisation, Internationalisation, and Globalisation », in Journal *Syn-thèses*.

Becker, E., Le, Z., Park, K., Lin, Y., Makedon, F. (2009), « Event-based experiments in an assistive environment using wireless sensor networks and voice recognition » in *Proceedings of the PETRA '09*.

Breslow, L. A., Harrison, A. M., & Trafton, J. G. (2010), « Linguistic Spatial Gestures » in D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling,* Philadelphia, PA: Drexel University, pp. 13-18.

Brown, P.F., Chen, S.F., Della Pietra**,** S. A., Della Pietra, V. J., Kehler A. S., Mercer R. L. (1994),« Automatic speech recognition in machine-aided translation » in *Computer Speech & Language,* Volume 8, Issue 3, July 1994, pp. 177-187.

Burda, A., Hageman, C.F., (2005), « Perception of accented speech by residents in assisted-living facilities » in *Journal of Medical Speech - Language Pathology*.

Burkhardt, D., Nazemi, K., Stab, C., Wichert, R., Breyer, M., Fellner, D. W. (2011), « Natürliche Gesteninteraktion mit beschleunigungssensorbasierten Eingabegeräten in unterstützenden Umgebungen » in *AAL Kongress*, Berlin, Germany.

D' Andrea, A., D' Ulizia, A., Ferri, F., Grifoni, P. (2009), « A multimodal pervasive framework for ambient assisted living » in *Proceedings of the PETRA '09,* Corfu, Greece.

Dachselt, R., Buchholz, R., (2009), « Natural Throw and Tilt Interaction between Mobile Phones and Distant Displays » in *CHI 2009*, *The Human-Computer Interaction Archive*, April 4–9, 2009, Boston, Massachusetts, USA, pp. 3253-3258.

Dandapat, S., Forcada, M. L., Groves, D., Penkale, S., Tinsley, J., Way, A. (2010), « OpenMaTrEx: A Free/Open-Source Marker-Driven Example-Based Machine Translation System» in*Proceedings ofthe7th International Conference on Natural Language Processing (IceTAL),* Reykjavík, Iceland.

Désilets, A., Stojanovic, M., Lapointe, J.F. et al. (2008) « Evaluating Productivity Gains of Hybrid ASR-MT Sytems for Translation Dictation » in *IWSLT Workshop*, Hawaii, pp. 158-165.

Fellbaum, C. (1998),*WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA.

Forcada, M. (2006), « Open source machine translation: an opportunity for minor languages » in *Proceedings of the 5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority languages", 5th International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, pp. 1-6.

Fricke, E., (2009) « Multimodal attribution: How gestures are syntactically integrated into spoken language » in *GESPIN 2009: Gesture and Speech in Interaction*, Poznań.

Gales, M.J.F., Liu, X., Sinha, R., Woodland, P.C., Yu, K., Matsoukas, S., T. Ng, Nguyen, K., Nguyen, L., Gauvain, J-L, Lamel L., Messaoudi A. (2007), « Speech Recognition System Combination for Machine Translation » in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, Hawai, 1277-1280.

Goetze, S., Moritz, N., Appell, J.E., Meis, M., Bartsch, C., and Bitzer, J. (2010), « Acoustic user interfaces for ambient-assisted living technologies » in*Inform Health Soc Care*, 35(3-4), pp. 125-143.

Haage, Schötz, S., and Nugues, P. (2002), « A prototype Robot Speech Interface with Multimodal Feedback » in *Proceedings of the 2002 IEEE Int. Workshop ROMAN,* Berlin, Germany, pp. 247-252.

Ivanecky, J., Mehlhase, S., Mieskes, M. (2011), « An Intelligent House Control Using Speech Recognition with Integrated Localization » in *AAL Kongress*, Berlin, Germany.

Jurafsky, D., Martin J.H. (2009), *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, 2nd edition, Prentice-Hall.

Karakos, D., Khudanpur, S. (2008), « Sequential System Combination for Machine Translation of Speech » in *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology* (SLT-08), Goa, India.

Khadivi, S., Zens, R., Ney, H. « Integration of Speech to Computer-Assisted Translation using Finite-State Automata », in *Proceedings of COLING/ACL* 2006, Sydney, Australia.

Koehn, P. (2009), « A Web-Based Interactive Computer Aided Translation Tool », *Proceedings of the ACL Interactive Poster and Demonstration Sessions,* Singapore*.*

Krieg-Brückner, B., Röfer, T., Shi, H., Gersdorf, B. (2010), « Mobility Assistance in the Bremen Ambient Assisted Living Lab » in*GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, 23 (2), Verlag Hans Huber, pp. 121-130.

Kurimo, M. et al. (2010), « Personalising speech-to-speech translation in the EMIME project », in*Proceedings of the ACL 2010 System Demonstrations*, Uppsala, Sweden.

Lee, A., Kawahara, T. (2009), « Recent Development of Open-Source Speech Recognition Engine Julius » in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (APSIPA ASC).

Manning, C., Schütze, H. (1999), *Foundations of Statistical Natural Language Processing,* MIT Press. Cambridge, MA: May 1999.

McNeill, D. (1992), *Hand and Mind – What Gestures Reveal about Thought*, The University of Chicago Press, Chicago, London, 1992.

Motallebipour, H., Bering, A. (2003), « A Spoken Dialogue System to Control Robots » *Technical report*, Dept. of Computer Science, Lund Institute of Technology.

Mubin, O., Bartneck, C., and Feijs, L. (2010), « Towards the Design and Evaluation of ROILA: A Speech Recognition Friendly Artificial Language » in *Proceedings of the 7th IceTAL* (Reykjavik, Iceland) 6233/2010, pp. 250-256.

Nagao, M. (1984), « A Framework of a Mechanical Translation between Japanese and English by Analogy Principle » in Elithorn, A.; Banerji, R. (Eds.), *Artificial and Human Intelligence*, Amsterdam, North-Holland, 173-180.

Neßelrath, R., Lu, C., Schulz, C. H., Frey, Jochen, Alexandersson, J. (2011) « A Gesture Based System for Context-Sensitive Interaction with Smart Homes »,in *Ambient Assisted Living, 4. AAL-Kongress 2011,* Berlin, Germany, January 25-26.

Nickel, K., Stiefelhagen, R. (2007), « Visual recognition of pointing gestures for human-robot interaction », in *Image and Vision Computing*, vol 25, Issue 12, pp. 1875-1884.

O'Brien, S. (2010). « Eye tracking in translation process research: methodological challenges and solutions » in *Copenhagen Studies In Language*, 38, pp. 251-266.

Paulik, M., Fügen, C., Stüker, S., Schultz, T., Schaaf, T., Waibel, A.(2005), « Document driven machine translation enhanced automatic speech recognition » in *Proceedings of InterSpeech*.

Pierce, J. R., Carroll, J. B., Hamp, E.P., Hays, D.G., Hockett, C.F., Dettinger, A.G., Perlis, A. (1966), *Language and Machines. Computers in Translation and Linguistics*, Report by the Automatic Language Processing Advisory Committee (ALPAC), Division of Behavioral Sciences, National Academy of Sciences, National Research Council.

Reddy, A. Rose, R. Desilets, A. (2007), « Integration of ASR and Machine Translation Models in a Document Translation Task », in *InterSpeech* 2007, Antwerp, Belgium.

Rehm, M., Bee, N., André, E., (2008) » Wave like an Egyptian: accelerometer based gesture recognition for culture specific interactions », in*BCS-HCI '08 Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction*, Volume 1.

Rosenfeld, R. Olsen, D. Rudnicky, A. (2001). **«** Universal Spech interfaces » in *Interactions* 8(6).

Rosti, A-V.I., Zhang, B., Matsoukas, S., Schwartz, R. (2008),« Incremental hypothesis alignment for building confusion networks with application to machine translation system combination », in *Proceedings of the Third Workshop on Statistical Machine Translation.*

Rosti, A-V.I., Zhang, B., Matsoukas, S., Schwartz, R. (2009)« Incremental hypothesis alignment with flexible matching for building confusion networks: BBN system description for WMT09 system combination task »in *Proceedings of the Fourth Workshop on Statistical Machine Translation.*

Santo Pietro, M.J., Ostuni, E. (1997), *Successful communication with Alzheimer's Disease Patients* - An In-service Manual. Boston:MA, Butterworth-Heinemann.

Theobalt, C. Bos, J. Chapman, T. Espinosa-Romero, A. Fraser, M. Hayes, G. Klein, E. Oka, T. Reeve. R. « Talking to Godot: Dialogue with a Mobile Robot» in *Proceedings of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Lausanne, Switzerland, pp. 1338-1343.

United Nations, (2002), *WORLD POPULATION AGEING : 1950-2050*,Department of Economic and Social Affairs Population Division, http://www.un.org/esa/population/publications/worldageing19502050/

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, Peter, Woelfel, J. (2004) « Sphinx-4 : A Flexible Open Source Framework for Speech Recognition »,*Sun Microsystems Technical Report*, No. TR-2004-139.

Will, T. (2007), *Creating a Dynamic Speech Dialogue*, VDM Verlag Dr. Müller.

Notes

1 We present only some and not all of the existing tools.

2 http://research.microsoft.com/en-us/projects/language-modeling/default.aspx

3 http://www.speech.sri.com/projects/srilm/

4 http://www-01.ibm.com/software/websphere/products/mobilespeech/

5 http://www-01.ibm.com/software/pervasive/embedded_viavoice/

6 http://www-01.ibm.com/software/pervasive/voice_server/

7 http://www.nuance.com/dragon/index.htm

8 ASR accuracy is rated with WER (word error rate) and speed is measured with the real time factor, but these are outside the scope of this paper.

9 http://www.microsoft.com/en-us/Tellme/default.aspx

10 http://www.svox.com/

11 http://www.svox.com/News-Items-Google-picks-SVOX-for-Translate-and-Dictionary-services-.aspx

12 http://www.lumenvox.com/

13 http://cmusphinx.sourceforge.net/

14 http://htk.eng.cam.ac.uk/

15 http://julius.sourceforge.jp/en_index.php

16 http://www.simon-listens.org/index.php?id=122&L=1

17 http://tldp.org/HOWTO/Speech-Recognition-HOWTO/software.html, http://htk.eng.cam.ac.uk/links/asr_tool.shtml, http://occidental.com.au/sr.html

18 http://www.nasa.gov/centers/ames/news/releases/2004/04_18AR.html

[19] http://www.systran.co.uk/

[20] http://www.promt.com/

[21] http://www.languageweaver.com/

[22] http://www.apptek.com/index.php/transphere-machine-translation-system

[23] http://www-01.ibm.com/software/pervasive/ws_translation_server/

[24] Aligners, language analysers, and evaluation tools are not considered here.

[25] http://www.apertium.org/

[26] http://www.cunei.org/

[27] http://joshua.sourceforge.net/Joshua/Welcome.html

[28] http://www.statmt.org/moses/

[29] http://logos-os.dfki.de/

[30] http://www.openmatrex.org/

[31] http://tool.statmt.org/

[32] http://translate.google.com/#

[33] http://www.microsofttranslator.com/

[34] http://www.freetranslation.com/

[35] http://babelfish.yahoo.com/

[36] http://www.cepstral.com/

[37] http://www.naturalreaders.com/index.htm

[38] http://www.spokentext.net/index.php?lang=en

[39] http://talkingclipboard.com/index.php

[40] http://espeak.sourceforge.net/

[41] http://www.cstr.ed.ac.uk/projects/festival/

[42] http://festvox.org/

[43] http://www.speech.cs.cmu.edu/flite/

[44] http://festvox.org/

[45] http://freetts.sourceforge.net/docs/index.php

[46] http://mary.dfki.de/

[47] http://www.yakitome.com/

[48] http://googleblog.blogspot.com/2011/01/new-look-for-google-translate-for.html

[49] http://technology.timesonline.co.uk/tol/news/tech_and_web/personal_tech/article7017831.ece

[50] http://www.jibbigo.com/website/index.php

[51] http://www.voxtec.com/phraselator/default.aspx

[52] http://www.apptek.com/index.php/speechtrans-speech-to-speech-translation

[53] http://www.cellictica.com/products.html

[54] Under typical we mean speech-to-speech, and not text-to-speech, or text-to-text dialogs.

[55] http://www.sfb673.org/projects/B1/

[56] http://camspace.com/

[57] http://www.gesturetek.com/3ddepth/introduction.php

[58] http://www.movesinstitute.org/~kolsch/HandVu/HandVu.html

[59] http://www.softkinetic.net/index.php?id=86

[60] http://www.sharewareconnection.com/titles/gesture-recognition16.htm

[61]

62

63 http://www.eruptingmind.com/communication-gestures-vary-different-cultures/

64 http://www.dfki.de/web/living-labs-en/baall-2013-bremen-ambient-assisted-living-lab-1?set_language=en&cl=en

65 http://www.aaliance.eu/public/

66 http://www.aal-europe.eu/

67 http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/EN/Navigation/Homepage__NT.psml