Speech-to-Speech Translation in an Assisted Living Lab

Dimitra Anastasiou
Computer Science/Languages and Literary Studies
University of Bremen
Bremen, Germany
00491637435527
anastasiou@uni-bremen.de

ABSTRACT

In this paper we describe speech processing in the Bremen ambient assisted living lab (BAALL); how it is currently and how it can be improved in the future with the implementation of Machine Translation (MT) technology. Multimodal human-machine interfaces (HMI) are not only preferred by the elderly and people with cognitive or physical impairments, but in many cases speech or touch interaction is the only option. We are experiencing the aging population phenomenon today and thus living in a safe, autonomous, independent, and convenient way in a domestic environment is desired more than ever. Pervasive computing is regarded as the next generation of computing and multimodal interfaces are indeed pervasive today. We argue for multilinguality in assisted living environments by means of development of a speech to speech translation (SST) system with the distinction – from traditional SST systems – of giving the output in a source (user's natural language/mother tongue) and not a target language.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Dictionaries, Linguistic Processing, I.2.7 [**Natural Language Processing**]: Speech recognition and synthesis, I.2.9 [**Robotics**]

General Terms

Performance, Design, Reliability, Experimentation, Human Factors, Languages, Theory.

Keywords

Assisted Living, Machine Translation, Speech Processing.

1. INTRODUCTION

There have been many research initiatives and advances in Ambient Assisted Living (AAL) environment the last years, see AALIANCE project, AAL Joint Programme and so on. AAL is concerned with intelligent assistant for a better, healthier and safer life in the preferred living environment through the use of Information and Communication Technologies (ICT). ICT help make AAL viable, efficient, and effective. In addition and not exclusively in AAL, multimodality, the seamless combination between different modes of interaction, from visual to voice to touch, has the following advantages according to [1]:

i. Improves accessibility to the device;
ii. Improves accessibility by the user;
iii. Offers improved flexibility, usability, and interaction efficiency.

Human-computer interaction (HCI) is a basic concept in ICT and speech recognition and synthesis intuitive realizations of HCI. When, in addition, users speak in their mother tongue, and not in English or an artificial language, they feel more confident, intuitive, autonomous, free, and friendly towards intelligent devices. [2] mention that "older persons targeted by AAL technologies especially need more easy-to-use methods to interact with inherently complex supporting technology". We believe that when users express themselves in their own natural language/mother tongue and devices/robots act after recognizing this natural language is an easy-to-use, intuitive, and natural method of HCI.

In this paper we focus on speech processing implementation in the Bremen ambient assisted living lab (BAALL) based at the University of Bremen; it is an apartment suitable for the elderly and people with physical or cognitive impairments. We present how is speech processing currently and how it can be improved in the future with a speech to speech automatic translation system. The paper is laid out as follows: in chapter 2 we provide some state-of-the-art about speech technology i) in assistive environments and ii) in relation to robotics, both with focus on vocabulary and grammar. Section 3 describes BAALL in general, an assistive wheelchair called "Rolland", and a dialog case scenario (3.1). In 3.2 we discuss our own motivation and contribution which will be exemplified in section 4. We summarize and conclude the paper in chapter 5.

2. RELATED WORK

In the next paragraphs we present some relatively (last decade) recent work related to dialog systems in AAL but also considering robotics in general.

To start with, in 2002 [3] stated characteristically that "situations humans helping mobile robots to find their way or to complete tasks while engaging in a dialogue are expected to become more widespread as robots begin to appear in domestic environment". This holds true as the following work shows.

[2] 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 dialogue are that i) nothing can be forgotten, ii) users are more concentrated on the information, and iii) easier communication with the technical system (p.18). Apart from that, the authors carried out an Automatic Speech Recognition (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 and female persons (p.25).

[4] described their project which is about deploying 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 dialogue management system they implemented uses the speech recognition 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 problems [4] faced in their experiments was the following: the longer and more continuous the speech is, the more recognition errors there are. They concluded that short and distinct phrases help to improve the precision of the speech recognition.

[2] 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 envisage four architectural levels: acquisition, analysis, planning, and activation level. In the analysis level, there is an Automatic 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 dialogue with relation to robotics but not in assistive environments. [5] developed a prototype to study integration of speech dialogue 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, Text-To-Speech application, XEmacs application, and 3D Robot application. The ASR application they used is Microsoft Speech API 5.1 software development toolkit (SAPI); they used SAPI in command mode (it can be also used in dictation 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 conclude that dialogue 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 dialogue came natural. Here it should be pointed out that [6] also mentioned that constraining language is a plausible method of improving recognition accuracy.

SAPI 5.1 was also used by [7] to develop a speech system to design robot trajectories that would fit with CAD paradigms. Apart from controlled language, another method to improve speech recognition accuracy is the use of an artificial language instead of a natural language. [8] 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.

[3] developed *Godot*, a mobile robot platform for the interface between a sophisticated low level robot navigation and symbolic high-level spoken dialogue system. The dialogue 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. The speech synthesizer they used was *Festival*.

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

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

Apart from speech processing in one language only, multilingual support in assistive environments is the next important step towards natural and intuitive HCI and telemedicine, as well. Thus we now focus on multilinguality and speech to speech translation systems. Undoubtedly, translation facilitates communication and can often save lives (see [9, 10]). Much important information is not translated into other languages apart from the lingua franca English in many domains in the 2/3 parts of the world. Apart from translation of written language, interpretation of spoken language is also crucial in the medical domain. [10] 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. Listeners were recruited from assisted-living facilities in small to mid-sized cities within the Midwest. 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 [11] shows 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. In this multilingual environment misinterpretation of care stuff instructions can lead to a wrong dosage of medicine which can have dramatic impact. Patients have to fully understand the medical instructions; thus both translation and interpretation plays an important socioeconomic, apart from a communicative role in AAL.

Speech to speech translation systems are developed nowadays mostly for mobile telephony (see Google's *Conversation Mode*, *jibbigo*, Trippo *VoiceMagix*). A definition of speech-to-speech translation is noteworthy. According to European Language Resources Association (ELRA),

> "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."

The distinction between our designed SST system and other SST systems is that the output is in the source and not in the target language (see 3.2). MT is used in order to translate the input into the language of the grammar developed for the machine, but the machine speaks back in the natural language of the user.

3. BAALL

At the University of Bremen, at the German Research Center for Artificial Intelligence (DFKI), there is a Bremen Ambient Assisted Living Lab (BAAL) which is an apartment suitable for the elderly and people with physical or cognitive impairments. The BAALL contains all standard living areas (home office, bedroom, bathroom and dressing area, living and dining room, kitchenette). The apartment was constructed according to the design-for-all principle according to the Casa Agevole at the Sta. Lucia research hospital in Rome (see [13]). A wheelchair called "Rolland" equipped with two laser range-sensors, wheel encoders, and an onboard computer serves mobility assistance in BAALL.

In BAAL natural interaction with the users is taken in serious consideration, through special devices for compensating special limitations, but foremost by emphasizing spoken dialogue. This is the environment where future ASR-MT combination systems can be applied. In the next sections we present the current speech processing in BAALL (3.1) and how it can be improved in the future (3.2).

3.1 Current Speech Processing

The user can interact by speech with Rolland in BAAL to navigate through the living areas. Rolland's navigation assistant and natural language interaction technology is integrated, as [13] 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 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 Italian.

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

i. Users with physical impairments can speak instead of clicking on buttons in their assisted environment;
ii. Users feel more friendly and socially integrated when speaking to Rolland;
iii. Rolland can intelligently make second steps (connecting pizza with kitchen) saving time and sparing tiresome single instructions.

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

i. Grammar is in German and English only;
ii. Grammar is minimal; thus the instructions limited;
iii. Rolland is not intelligent enough to remember his original position so that he returns back or follow a sequence of events (after Mario finishes his pizza, wants to wash his hands, so Rolland brings him to the bathroom);
iv. Locale/Context awareness is missing (user has to say *turn the kitchen light off* even if you are in the kitchen).

3.2 Our Contribution

Our contribution is to cover the aforementioned limitations of monolinguality, minimalism, and activity simplicity. Multilinguality can be achieved by developing a unique platform combining an ASR system, an MT system, and a text to speech (TTS) system. All systems will be free/open-source. Using the platform, the HCI workflow would be the following:

i. Users speaks in their mother tongue and speech is transcribed to text through ASR;
ii. MT of existing grammar of Rolland into another natural language – user's language;
iii. Rolland gives the output (if any needed or wanted) back in the users' natural language though TTS.

In addition to the advantages we mentioned in 3.1, the advantages of our contribution are that is exclusively based on open-source software, so it is accessible and affordable. Apart from that, users 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.

Minimalism is faced by lexical, morphological, and syntactic extensions to the grammar. These extensions will cover morphologically rich and syntactically flexible languages. The lexical vocabulary entries (which lead to more tasks that Rolland will be able to do) can be arbitrarily extended, as MT caters for automatic translation. Last but not least, activity simplicity can be with activity-based ontologies. More information about how we combat monolinguality can be found in the next chapter 4. The other two items are outside the scope of this paper.

4. PLATFORM

We develop a unique platform where the systems *CMU Sphinx*, *Microsoft Bing Translator*, and *FreeTTS* are used. The reasons we selected these tools are because i) they are all open source, ii) well known, and iii) used by other researchers and applications (see section 2). This unique platform is initially implemented in Rolland. The individual components are mentioned below.

4.1 Speech recognition

Speech recognition maps from an acoustic signal to a string of words. Non-verbal components of speech, such as pausing and rhythm, intonation, new vs. given info, length of utterance, the modality effect, auditory suffix effect, pose challenges in ASR.

The ASR tool Sphinx [14], current version 4, is a tool developed by the Carnegie Mellon University (CMU). It is a flexible, modular, and pluggable framework to help foster new innovations in the research of hidden Markov model recognition systems.

4.2 Machine Translation

Machine Translation (MT) concerns the automatic translation of text or speech from one natural language into another by means of computer software. MT has three approaches: rule-based MT, example-based MT, and statistical MT (SMT). SMT is the approach of MT systems used in the ASR-MT combination, because SMT distributes probabilities and provides hints to the ASR system. Challenges of MT in general, include idiomaticity (spoken language, discourse, idioms, metaphors) and data challenges (domain specific, string length). A current initiative towards speech to text (STT) system combination and its impact on MT performance have been taken by [15]. Although hypothesis combination gave lowest error rates, the use of cross-adaptation was found to be a "safer combination scheme for translation" (p.1280).

Bing Translator [16] is a free translation service that provides translations of words, sentences, and webpages; it is a hybrid approach (RBMT and SMT) and for some language pairs it uses the technology of *SYSTRAN*. Initially, we compared some MT systems (demo of *SYSTRAN, Bing, Google Translate, CAITRA*) and saw that for our purposes a hybrid MT approach fits best.

4.3 Speech synthesis

Speech synthesis/text to speech (TTS) refers to the production of speech (acoustic waveforms) from text.

FreeTTS [17] is a speech synthesis system, based upon *Flite*, a small run-time speech synthesis engine developed at CMU. *Flite* is derived from the *Festival* Speech Synthesis System from the University of Edinburgh and the *FestVox* project from CMU.

4.4 Future development

Future prospects concerning our platform include evaluating other free/open source software (FOSS) STT, MT, and TTS systems, to the already selected ones. After that, we plan to follow the cross-adaptation for STT system combination to evaluate later the MT quality output. Moreover, training of acoustic models with voices of older people as training sets is among our future experiments.

5. SUMMARY and CONCLUSION

Multimodality is important in many different aspects of everyday life, particularly in AAL and domotics. In 2030 26% of the population will be over 65 years; thus the need for a safer, healthier, and more independent lifestyle for the elderly and people with impairments is essential. The Bremen Ambient Assisted Living Lab (BAALL) is an apartment suitable for the elderly and people with physical or cognitive impairments. We referred to some related work about dialogue and robotics and/or AAL. Then the current speech processing in BAALL was briefly described with more focus on ourcontribution. For the time being, this includes a unique platform combining FOSS ASR, MT, and TTS systems. Multilingual support in HCI facilitates not only independence, intuition, and user-friendliness, but is often necessary to avoid dramatic medical accidents.

6. ACKNOWLEDGMENTS

Our thanks for the helpful comments and feedback of the reviewers to revise this paper. We gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center SFB/TR 8 Spatial Cognition.

7. REFERENCES

[1] D' Andrea, A., D' Ulizia, A., Ferri, F., and Grifoni, P. 2009. A multimodal pervasive framework for ambient assisted living. In *Proceedings of the PETRA '09*, (Corfu, Greece, June 9-13, 2009).

[2] Goetze, S., Moritz, N., Appell, J.E., Meis, M., Bartsch, C., and Bitzer, J. 2010. Acoustic user interfaces for ambient-assisted living technologies. *Inform Health Soc Care*. 2010 Sep-Dec; 35(3-4):125-43.

[3] C. Theobalt, J. Bos, T. Chapman, A. Espinosa-Romero, M. Fraser, G. Hayes, E. Klein, T. Oka, R. Reeve. 2002. Talking to Godot: Dialogue with a Mobile Robot. In *Proceedings of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems* (Lausanne, Switzerland), 1338-1343.

[4] Becker, 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* (Corfu, Greece, Jun 9-13, 2009).

[5] Motallebipour, H. and Bering, A. 2003. *A Spoken Dialogue System to Control Robots*. Technical report, Dept. of Computer Science, Lund Institute of Technology.

[6] Rosenfeld, R. Olsen, D. Rudnicky, A. 2001. Universal Spech interfaces. *Interactions* 8(6) (2001).

[7] 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, September 25-27, 2002), 247-252.

[8] 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, August 16-18, 2010) 6233/2010, 250-256.

[9] Anastasiou, D. and Schäler, R. 2009. Translating Vital Information: Localisation, Internationalisation, and Globalisation, *Journal Synthesis*.

[10] Burda, A. and Hageman, C.F. 2005. Perception of accented speech by residents in assisted-living facilities. *Journal of Medical Speech - Language Pathology*.

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

[12] HLT Evaluation Portal, Speech-to-Speech Translation: http://www.hlt-evaluation.org/article.php3?id_article=35

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

[14] CMU Sphinx, http://cmusphinx.sourceforge.net/

[15] 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., and Messaoudi A. 2007. Speech Recognition System Combination for Machine Translation. In *Proceedings of the IEEE ICASSP* (Honolulu, Hawai, April 15-20, 2007), 1277-1280.

[16] Bing Translator: http://www.microsofttranslator.com/

[17] FreeTTS: http://freetts.sourceforge.net/docs/index.php#what_is_freetts