

Automatic Speech Transcription in AAL Solutions

Alexander Sorin¹, Ron Hoory¹,

¹ IBM Haifa Research Lab, Haifa University Campus, Mount Carmel, 31905 Haifa, Israel
{sorin, hoory}@il.ibm.com

Abstract. Voice is a valuable channel for extracting information on user's context, intentions and needs within an AAL solution. Automatic speech transcription (speech-to-text) is an inevitable step in semantic analysis of speech content. In this paper we consider challenges, limitations and perspectives of speech transcription application in AAL solutions based on and generalizing the experience gained in HERMES EU FP7 project.

Keywords: speech transcription, conversation, acoustic model, language model, adaptation, topic detection, semantic analysis

1 Introduction

Video, audio, biometrical and physical sensors are employed in various AAL projects. They are used in order to capture the user's input directly or to sense the user's context and deduce user's intentions and needs from his/her interaction with the environment. Voice, as an inherent and the most natural means of human interaction, is a valuable information channel for this purpose. One traditional application of this idea is the voice-enabled HCI (Human Computer Interface), which is successfully deployed in various areas and can be adopted in AAL. Another (complementary) direction is the development of systems that listen to ambient conversations of the user with other people and extract from them information relevant to the services provided by specific AAL solutions. It is important to note that related ethical and privacy aspects should be respected in such a solution. This area is less researched and significantly more challenging compared to the voice dialog based HCI.

Extraction of meaning that can be utilized by the system from speech must be preceded by automatic speech transcription (i.e. converting the speech signal to text). In general speech transcription is the most challenging automatic speech recognition (ASR) task¹, though the level of difficulty varies across speech transcription applications.

A speech transcription system employs a combination of acoustic and language models (AM and LM respectively) in order to determine the sequence of words that best fits an observed audio signal. The AM, that typically contains a huge number of parameters, is built by an offline *training* process that requires large volumes of

¹ Examples of other ASR tasks are recognition of constrained spoken input in automatic dialog systems (e.g. flight booking) and voice enabled directory lookup, e.g. voice dialing.

speech audio data along with manual transcripts. The training data should contain speech uttered by many speakers. The LM models the probability of any combination of n consecutive words, e.g. four words. Huge text corpora are required for the LM building.

The first commercial (and also the least challenging) application of speech transcription was desktop dictation (e.g. IBM ViaVoice Dictation product). Substantial advances in ASR technology achieved within the last decade made it possible to approach transcription of spontaneous conversational speech [1]. Examples of emerging applications in this area are transcription of a contact center phone calls and broadcast TV programs. These applications involve relatively narrow lexical domain and audio signals with high signal-to-noise ratio (SNR) level recorded by a close-talking microphone.

The accuracy of speech transcription is measured by Word Error Rate (WER), the percentage of replaced, deleted and inserted words compared to the actually uttered text. The transcription accuracy varies in a wide range across the applications. Two polar examples give an idea of the range. A WER of 5% is achievable in the desktop voice dictation. Conversely, the best ASR system developed in the CHIL EU FP6 project for transcription of meeting recordings made by distant table-top microphones achieved WER of 46% [2].

It is worth noting that state-of-the-art speech transcription systems are not robust to a mismatch between training and operation conditions. I.e., the training data used for acoustic and language modeling should match well the acoustic (microphone, noise) and linguistic (vocabulary, word statistics) aspects of the target application. Otherwise the accuracy drops down dramatically. Furthermore, operation environments targeted by certain applications (e.g. meetings transcription) establish a challenge to the state-of-the-art technology regardless of the availability of appropriate training data.

Below we discuss challenges, limitations and perspectives of speech transcription application in AAL solutions generalizing our experience gained in HERMES EU FP7 project [3].

2 Case Study: Speech Transcription in HERMES EU FP7 Project

The HERMES project develops a personal cognitive support system for the elderly population. In the HERMES system speech transcription enables important services supporting numerous user scenarios. To this end, transcripts of conversations are used for content based search and semantic analysis including topic segmentation and categorization. The audio can be recorded either by a PDA operating the HERMES client SW or by stationary wall-mounted microphones installed at the user's home. A close-talking wireless headset microphone is also used in the project, mainly for research purposes. Spanish is the target language for speech transcription in the HERMES proof-of-concept prototype.

The R&D work on speech transcription was preceded by a data collection task. Spoken audio data from 50 potential elderly users comprising 55 hours of conversations and 12 hours of read-out newspaper articles has been recorded. The

recording was carried out simultaneously by all types of microphones and manually transcribed. This data is used for adaptation of the ASR system and for evaluations.

As a baseline system to build on, we have chosen the Spanish transcription system that has been developed by IBM for transcription of European Parliamentary Sessions Speeches in TC-STAR EU FP6 project. The baseline system is based on IBM Research Attila ASR toolkit. Hundreds of hours of manually transcribed audio and text corpora containing 43 billion words have been used in TC-STAR for training the acoustic and language models respectively. WER of 8% has been achieved by this system in the TC-STAR evaluation [4].

Acoustic and linguistic aspects of the ASR task targeted in HERMES (i.e. distant microphone, two speakers, spontaneous informal speech) and in TC-STAR (i.e. close-talking microphone, single speaker, consistent language) significantly differ from each other. This discrepancy results in a high degree of mismatch between the baseline system training conditions and testing conditions stemming from the HERMES requirements. It was anticipated that the mismatch associated with the newspaper read-out speech captured by the close-talking microphone would be moderate while the conversations recorded with the PDA would pose a high degree of mismatch². The evaluation results presented in Table 1 support this assumption and reveal the separate influence of the acoustic and linguistic aspects.

Table 1. WER achieved by the baseline system on HERMES data.

	Close-talking microphone	PDA microphone
Readout	WER=24%	WER=41%
Conversation	WER=48%	WER=68%

Both acoustic and language models should be tailored to the HERMES pertinent data in order to improve the accuracy of the conversation transcripts.

Offline supervised AM adaptation to the voice of certain speaker, referred to *speaker enrolment*, is well aligned with the personal nature of the HERMES system. The adapted AM can be used for transcription of speech uttered by the enrolled speaker, e.g. the primary user of HERMES system or his/her relative³. The enrolment adaptation of the baseline AM performed on the PDA readout data also adapts the model to the acoustic channel and background noise to a limited degree.

The training text corpus derived from the conversational data collected in HERMES consists of 150,000 words which is by far less than needed for building a decent LM. For the study purposes we have built a new LM from this corpus and mixed it with the baseline LM in order to circumvent the overfitting effect of training on the limited amount of data. The effect of the AM and LM adaptation is demonstrated by Table 2. The acoustic and language model adaptation steps outlined above led to 20.5% relative WER reduction. Our further on-going activity includes complete re-training of the baseline AM on the bulk of the multi-speaker conversations and attempts to

² The work on the data recorded by the wall-mounted microphones is at an initial stage. Hence we do not address this data type in the paper.

³ This approach must be underpinned by speaker segmentation and identification technology which is addressed within HERMES project but is beyond the scope of this paper.

identify external sources (e.g. web) of relevant text corpora for the language modelling.

Table 2. WER reduction on PDA-recorded conversations by AM and LM adaptation.

	Baseline AM	Speaker enrolled AM
Baseline LM	WER=68%	WER=64%
Conversation LM	WER=60%	WER=54%

3 Discussion and Conclusions

Typically an AAL solution sets up highly challenging conditions for speech transcription including use of distant microphones, non-stationary background noise and open lexical domain. Availability of the relevant speech data for ASR training is limited, since massive data collection is usually unaffordable in a multidisciplinary AAL R&D project. Keeping in mind the state of the art in ASR technology, it means that the accuracy of conversational speech transcription achievable within an AAL system in WER terms is expected to be in the range of 50% - 40%.

This level of WER is perceived as high. However it is worth noting that the WER measure is somewhat pessimistic. It penalizes nominal errors like small modifications of suffixes and prefixes and other errors that do not affect the text comprehension. The workflow envisioned within an AAL system is aiming at a shallow semantic analysis (e.g. topic segmentation and classification) where speech transcription is followed by identification of certain keywords and analyzing their statistics in the transcripts. In order to optimize the performance of the semantic analysis the analyzer should be trained/developed based on the actual ASR transcripts. Moreover, the meta-information usually produced by an ASR system can be utilized by the semantic analysis to improve its performance. This information includes alternative choices, word confidence measures and phonetic transcript. This approach is implemented in the spoken information retrieval system used by HERMES [5].

References

1. Kingsbury, B., Mangu, L., Saon, G., Zweig, G., Axelrod, S., Goel, V., Visweswariah, K., Picheny, M.: Towards domain-independent conversational speech recognition, In *EUROSPEECH-2003*, Geneva, Switzerland (2003)
2. <http://www.springerlink.com/content/f573675230lk3185/>
3. HERMES EU FP7 project website: <http://www.fp7-hermes.eu/>
4. Ramabhadran, B., Siohan, Olivier, Mangu, L., Zweig, G., Westphal, M., Schulz, H., Soneiro, A.: The IBM 2006 speech transcription system for European parliamentary speeches", In *INTERSPEECH-2006*, Pitsburg, PA, USA (2006)
5. Mamou, J., Ramabhadran, B., Siohan, O.: Vocabulary Independent Spoken Term Detection, in *SIGIR-2007*, Amsterdam, Netherlands (2007)