Joint Application of Speech and Speaker Recognition for Automation and Security in Smart Home

Kong Aik Lee, Anthony Larcher, Helen Thai, Bin Ma, Haizhou Li

Joint Application of Speech and Speaker Recognition for Automation and Security in Smart Home

Kong Aik Lee, Anthony Larcher, Helen Thai, Bin Ma, and Haizhou Li

Human Language Technology Department, Institute for Infocomm Research
Agency for Science, Technology and Research (A*STAR), Singapore
{kalee,alarcher,nthhthai,mabin,hli}@i2r.a-star.edu.sg

Abstract

This paper describes the deployment of speech technologies in *STARHome*, a fully functional smart home prototype. We make use of speech and speaker recognition technologies to provide three voice services, namely, voice command for controlling home appliances, voice biometric for entrance-door access control, and service customization (speaker-loaded command control). Voice applications for *STARHome* have been designed to deal with short utterances and low SNR.

Index Terms: home security and automation, command control, smart home, speaker recognition

1. Introduction

While the term *smart home* might have different definitions, the most common interpretation is that it is a home equipped with various technologies which anticipate and respond to the needs of the occupants intuitively. To this end, the *STARHome*, a fully functional 180 square meters smart home prototype located at the *Fusionopolis*, Singapore, has been developed to facilitate the cross-fertilization of ideas and technologies catering to the lifestyle of modern urban living [1]. Speech technologies are found very much applicable to smart home, for example, for automatic transcription of TV programs to give subtitles in the original or other languages. Speech command is a more convenient and natural way of controlling home appliances. Of particular interest in this *Show & Tell* paper is the joint application of speech and speaker recognition technologies in smart home.

Voice appears to be a natural way to interact with automated systems as it conveys different types of information, which to our concern are the message and the speaker identity. The *STARSpeaker* recognition engine described in this paper aims to incorporate command control and voice biometric for home automation and security. It makes use of two mature speech technologies, speech recognition and speaker authentication, to provide three voice services. These services include command control, service customization and entrance-door access control, as whole could be seen as a unique voice command service with an increasing level of security.

Over the past few decades, speaker recognition engines have seem to reach high accuracy in the context of conversational telephony speech as evaluated in the past few NIST speaker recognition evaluations (SREs) [2]. The error rate had dropped below 5% using speech segments of two and a half minutes for enrolment and test. This might be useful for application where telephone calls are monitored and screened for suspects over an immense amount of recordings. However, additional constraints have to be imposed for the ergonomic use of voice biometrics in daily context. The most critical being the duration of utterances which, in most cases, have to be restricted to be 3 seconds or less. On top of this is the low signal-to-noise ratio (SNR) in home environment.

2. Application scenario and system architecture

As mobile devices, like smart phones and tablet PCs, become more and more common, we propose to use such devices as an alternative to the more challenging far-distance microphone. An added advantage is that these mobile devices also come with touch-screen displays, which allows pass phrases to be prompted to the users. We could also exploit the touch-screen capability to implement the so-called *push-to-talk* feature, which greatly simplify the design of the voice activity detection (VAD). Since mobile devices are hand held, the distance from mouth to the microphone will always be constrained within 60 cm. These mobile devices are connected to the *STARSpeaker* server via Wi-Fi link, which is now very common in domestic use. Fig. 1 shows the system architecture of the *STARSpeaker* server. The application scenario is explained as follows.

A graphical user interface (GUI) is available on the server for the user to enroll and change settings like the list of commands and sentences required for authentication. For enrollment, the new user accesses the server GUI and set his/her Id. The server activates the enrollment procedure on the smart-phone which guides the user all along the enrollment step. In the targeted application scenario each user will be ask to record 3 times a set of 30 short phrases. This set of 30 sentences can easily be customized through the server GUI. A fast enrollment scenario has been designed for the *INTERSPEECH Show & Tell* demonstration. In this version, each user is requested to pronounce three short sentences twice.

When entering the *STARHome*, the user launch the corresponding service through an ergonomic Android© application and read the sentence displayed on the smart-phone screen. The authentication result is then displayed on the smart-phone screen and the door open whether the test is positive. Inside the house, command control is activated by a *push-to-talk* feature available on the smart-phone.

3. Speaker recognition using very short utterances

Using voice biometrics for security and home automation involves several ergonomic constraints including a drastic limitation of speech duration for recognition. Despite the high accuracy reached by text-independent speaker recognition engine in recent benchmarking evaluations, they exhibit a pronounced accuracy knee when limiting the speech duration below 20 seconds.

Our aim in this project is to be able to perform recognition using utterances less than 3 seconds, at the same time achieving an error rate below 5% for both miss and false acceptance. In order to deal with very short duration we chose to exploit text-dependency. Indeed, modeling both speaker

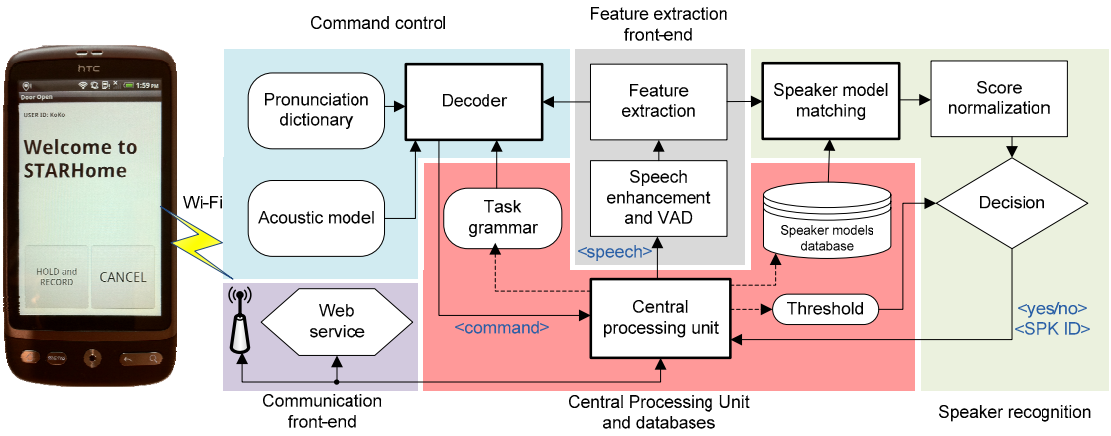

Figure 1: *The STARSpeaker recognition system. Five major components are the feature extraction front-end, communication front-end, command control, speaker recognition, and the central processing unit. The smart phone serves as the user interface with a microphone for recording and a display allowing pass phr controls home appliances via Web services.*

characteristics under specific phonetic context and temporal structure of speech has been shown to greatly improve performances [3]. Text-dependent speaker verification approaches usually belong on two families. The first one, based on a LVCSR systems which acoustic parameters are adapted to the user, usually requires a large quantity of adaptation data. The second family makes use of Dynamic Time Warping (DTW) to measure the similarity of temporal structure between speech segments. Thus DTW approaches can not take benefits of the whole speech material available from a given speaker.

The *STARSpeaker* recognition engine performs a progressive specialization through a three-layer architecture and takes advantage of all the speech material available from the speaker while harnessing the temporal structure of short speech utterances. This specific three-layer architecture originally proposed in [4, 5], as depicted in Fig. 2, allows the recognition engine to reach high accuracy in this challenging context.

4. Command control and service customization

We define two types of commands – specific and general. A specific command has a designated action to be carried out, for examples, *door open*, *light on*, *TV on*, *channel one*, etc. On the other hand, a general command is always open-ended and requires more information, in our case, the speaker identity for the system to respond to the user. Examples of general commands are, *play music*, *watch movie*, *call mum*, etc. By linking a pre-set list of preferences to the speaker identity, the general commands are customized to some specific action, like playing some music preferred by the user.

The command recognizer is built on top of a LVCSR system. The acoustic models were trained using English with Singapore accent. The HMMs are continuous density GMM tied-state triphones with clustering performed using phonetic decision trees. Since the base system is general purpose, adding a new command involves only modification to the pronunciation dictionary and the task grammar.

When a general command is recognized, the *STARSpeaker* system automatically performs an open-set identification task by authenticating the speaker when belonging to the *STARHome* enrolled users.

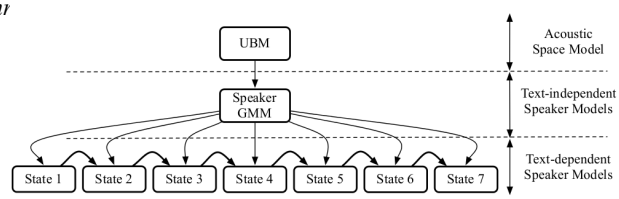

Figure 2: *Overview of the three layer acoustic architecture.*

5. Conclusions

We have reported the core elements of an on-going two years research project with emphasis on commercialization of speech and speaker recognition technologies. By restricting some flexibility, we found that speech technologies could be the right technologies for home security and automation.

6. Acknowledgements

This project is funded under HOME2015 Phase 2 Program, Science and Engineering Research Council (SERC), A*STAR Singapore.

7. References

[1] The STARHome Project. Available: http://www.starhome.sg

[2] A. F. Martin and J. S. Garofolo, "NIST speech processing evaluations: LVCSR, speaker recognition, language recognition," in *Proc. IEEE Workshop on Signal* Processing *Applications for Public Security and Forensics*, 2007, pp. 1-7.

[3] J. H. Neeland, J. W. Pelecanos, R. D. Zilca, and G. N. Ramaswamy, "A study of the relative importance of temporal characteristics in text-dependent and text-constrained speaker verification," in *Proc. IEEE ICASSP*, 2005, pp. 653-656.

[4] A. Larcher, J. -F. Bonastre, and J. S. D. Mason, "Reinforced temporal structure information for embedded utterance-based speaker recognition," in *Proc. INTERSPEECH*, 2008, pp. 371-374.

[5] J. -F. Bonastre, P. Morin, and J. -C Junqua, "Gaussian dynamic warping (GDW) method applied to text-dependent speaker detection and verification," in *Proc. EUROSPEECH*, 2003, pp. 2013-2016.