ALADIN: The self-taught vocal interface

Jort F. Gemmeke
ESAT-PSI, KU Leuven, Kasteelpark Arenberg 10, 3001, Leuven, Belgium
email: jgemmeke@amadana.nl

Abstract

We describe a demonstration of a novel vocal user interface (VUI), ALADIN, which is trained through usage, by demonstrating spoken commands with manual controls. It works by mining the vocal commands to find recurrent acoustic patterns corresponding to words or phrases that constitute elements of the user's commands. The demonstration consists of the VUI controlling a 3D environment with home automation devices such as lights and doors, as well as an actual home automation setup. Additionally, a tablet interface is provided for feedback and manual control.

1 Introduction and motivation

These days, Vocal User Interfaces (VUIs) are firmly rooted in everyday life, with ample examples such as talking to your smartphone, vocal interfaces for home automation or directing your navigation device by voice while driving. In this work, we describe a demonstrator of the technology developed in the ALADIN project[1], which aims at building a VUI that is trained by the end-user himself based on his demonstrated actions and spoken commands [1, 2].

Most state-of-the-art VUIs rely on large vocabulary Automatic Speech Recognition (ASR) to convert the spoken commands into text, and then use language processing techniques to convert the transcribed commands into actions, feedback or queries for more information. Other VUIs may opt for an ASR system with more restricted vocabulary and grammar, adapted to the device. These VUIs have in common that they are trained in advance on large amounts of speech material and interaction data, with possibly an adaptation to the end-user's speech or behavior. For some users however, such as those speaking under-resourced languages, dialects or those with a speech impairment, the VUI may still not be able to offer usable performance.

[1]This research was funded by the IWT-SBO project ALADIN (contract 100049).

In the proposed VUI, we take a novel approach by learning the actionable words, subwords or phrases from the user's speech in a completely language-independent, speaker-specific way. This is done by mining the spoken commands together with demonstrated actions to find recurrent acoustic patterns corresponding to parts of commands. By not using ASR and mapping audio directly to actions, we avoid the use of possibly incorrect or unavailable pronunciation models, vocabularies and language models. Moreover, since training is done by the end-user through usage, the VUI can adapt naturally to changing speech or environments.

The proposed approach was designed for a specific target audience: physically impaired people with restricted limb motor control (c.f. Figure 1), who also suffer from (often severe) dysarthria. Their speech is often dialectic and may change over time due to a progressive muscle disease. We refer the reader to [3, 4] and the references therein to related work on adapting conventional VUIs to dysarthric speech.

In the remainder of this work, we briefly discuss the employed techniques in Section 2 and describe the implementation and components of the demo in Section 3. We conclude with a discussion of future plans in Section 4.

2 Scientific and technical description

In this section we give a brief description of the technique employed by the VUI. For a more in-depth treatment we refer the reader to [2, 5]. We distinguish two phases: a training phase and a usage phase. In the training phase, a command is learned by giving the desired vocal command, followed by demonstrating the action with the manual control. For example, the user could give the vocal command "Turn on the television" together with pressing the standby button on the television remote control.

A vocabulary finding module uses a semantic frame description of manual controls together with the vocal commands to find words or phrases that constitute the user's commands. The main challenge here is that since the VUI operates without prior speech knowledge, we work without segmentation and without knowledge of word order. The

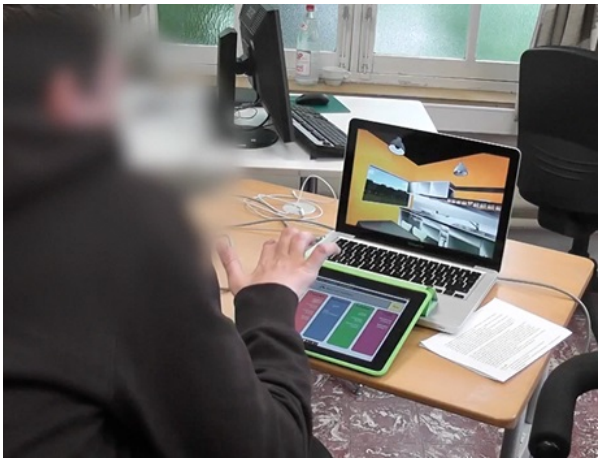

Figure 1: Target user with the proposed VUI. The user has spastic hand movements and impaired speech. The laptop screen shows the immersive 3D environment that can be controlled with spoken commands. The tablet provides feedback and allows manual control.

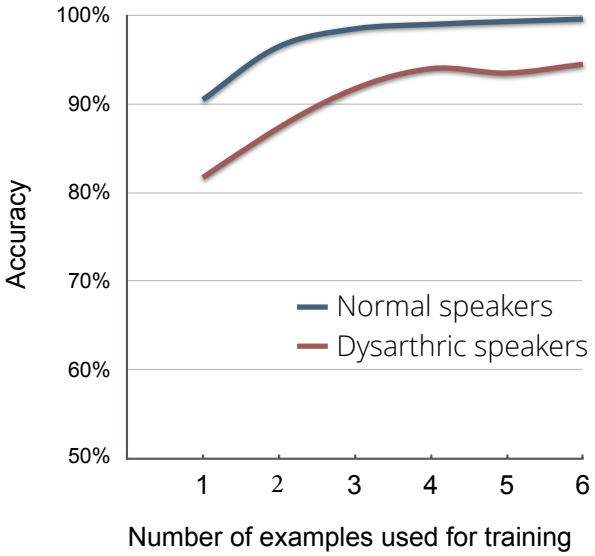

Figure 2: Evaluation of average action recognition accuracy over speakers as a function of the number of examples given of each command. The recognition task consists of 30 home automation commands.

vocal command is processed into low level (spectrographic) and intermediate level (utterance-based) acoustic representations. Th word finding is based on non-negative matrix factorization (NMF), which decomposes the utterance-based representations into a low rank multiplication of recurrent acoustic units representing subwords, words or phrases, and their activation across sentences. For example, the learned 'vocabulary' could become "Turn on" and "the television".

In the usage phase the most likely command is induced from a vocal command by factorizing it's utterance-based representation using the acoustic units found in training. With acoustic units mapped to elements of semantic frames during training, the recognized action is then send to the device. The effectiveness of this approach has been evaluated on a database with spoken home automation commands [5], and is visualized in Figure 2.

3 Implementation and use

3.1 Demo components

The demo consists of multiple components, shown in Figure 3. The VUI runs on a dedicated device such as a laptop. It takes speech input either from a far-field microphone (for example the laptop's built-in microphone) or a wireless close-talk microphone. A second component, possibly running on a separate device, is an immersive 3D environment, which simulates a home environment and allows control of doors, blinds and lights (26 commands in total). Movement in the 3D environment is controlled with a game controller, which also operates the push-to-talk functionality of the VUI.

The third component consists of a companion application running on a tablet, which provides feedback to the user and allows manual (touch) control of the home automation devices. The application was specifically designed to be operated with impaired users, with large buttons and controls that activate on touch release, rather than touch presses. To increase the interactivity of the demo, a final component consists of an actual home automation setup with four individual controlled power outlets connected to colored lights (not shown).

3.2 Implementation

The 3D environment is implemented in Unity. The VUI is currently implemented in Matlab. The VUI is trained by repeating every command (30 in total) three times in both a noisy and a quiet environment. Both sessions lasted only 5 minutes each.

The VUI communications with the tablet and the 3D environment via a network connection. The home automation component communicates with the VUI using a KNX bus.

3.3 Media

We provide a website [6] with video's which demonstrate an earlier version of the demo, the prospective end-users, and the current training process.

Figure 3: Demo setup. In the top-right corner, the interactive 3D home environment is shown. The VUI is running on the laptop, with the companion tablet shown to the right. The demo staff operating the VUI is using a wireless microphone and a game controller. A far-field microphone and actual home automation setup are also supported (not shown).

4 Conclusions and future developments

We describe a demonstration of a vocal user interface (VUI) which can be trained through usage, by demonstrating spoken commands with manual controls. The demonstration consists of the VUI controlling a 3D environment with home automation devices such as lights and doors, as well as an actual home automation setup. Additionally, a tablet interface is provided for feedback and manual control.

While in the current demonstration, there is a separate (short) training phase and usage phase, future development will see an integration of these two phases: The system will keep updating itself when a new training item or correction is provided by the user.

References

[1] The ALADIN project website. [Online]. Available: http://www.esat.kuleuven.be/psi/spraak/projects/ALADIN/

[2] J. F. Gemmeke, J. van de Loo, G. De Pauw, J. Driesen, H. Van hamme, and W. Daelemans, "A self-learning assistive vocal interface based on vocabulary learning and grammar induction," in *Proc. INTERSPEECH*, 2012, pp. 1–4.

[3] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proc. INTERSPEECH*, 2012.

[4] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria," *Medical Engineering & Physics*, vol. 5, no. 29, pp. 586 – 593, 2007.

[5] B. Ons, N. Tessema, J. van de Loo, J. F. Gemmeke, G. De Pauw, W. Daelemans, and H. Van hamme, "A self learning vocal interface for speech-impaired users," in *Proc. Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2013, pp. 73–81.

[6] [Online]. Available: http://www.aladinspeech.be/media