Use case : a mobile speech assistant for people with speech disorders

Agnieszka Bętkowska Cavalcante and Leszek Lorens

Gido Labs sp. z o.o.
Romana Maya 1, 61-614 Poznań, Poland
{agnieszka.betkowska,lech.lorens}@gidolabs.eu

Abstract

Speech is one of the most natural means of communication, but can be a challenge for people with speech disorders. In particular, they often have severe difficulties with correct articulation, which is one of the main causes of intelligibility issues for interlocutors. If the speech disorders also coincide with serious motor disabilities, this leads to overwhelming limitations: a person is not able to communicate by voice, nor is able to use common alternative communication methods requiring precise movements. To address this situation, researchers have advocated the use speech recognition technology. However, existing speaker independent speech recognition systems typically fail to recognize distorted speech, and they are often not tailored to individual users' necessities. Against this background, in this study we propose a novel platform that enables nontechnical users to create speech assistants tailored to their needs and disabilities. As a concreate example, we discuss a mobile phone assistant created and tested by a person with cerebral palsy and with distorted explosive speech. This assistant, which is controlled by voice, enables its user to read emails and other text messaging services, to send predefined messages to the people from a contact list, and to perform phone calls. In laboratory conditions, the performance of the speech recognition system tailored to this particular user was above 99%.

1. Introduction

Large IT players such as Google and Apple have been investing massively in human-machine interfaces, and in recent years particular emphasis has been devoted to the development of speech interfaces (speech recognition/understanding and speech systems). Existing cloud-based solutions developed by these companies have been a large success in the field because they are consistently achieving good recognition rate for modal speech. However, they still have poor performance when used by speakers whose speech is impaired for various reasons. Unfortunately, a similar phenomenon occurs with human listeners: humans can easily understand modal and undisturbed speech, but they have difficulties in understanding disordered speech because the linguistic content and paralinguistic information in the uttered message is degraded (Tran et al., 2010). If speech disorders coincide with serious motor disabilities, this can lead to overwhelming limitations: a person is not able to communicate by voice, nor is able to use common alternative communication methods requiring precise movements.

Owing to the large spectrum of speech distortions and articulation problems of speakers, devising speaker independent recognition systems targeting disordered speech is very difficult (Rosen and Yampolski, 2000). Adaptation of existing state-of-the-art speech recognition systems to the speech of users with speech impairments often does not offer satisfactory improvements (Havstam et al., 2003),(Rosen and Yampolski, 2000). Hence, to simplify the original problem, researchers have focused on the development of user-dependent dysathric speech recognition systems (Fager et al., 2010). Most of the proposed systems are created in laboratory conditions, with pre-defined limited vocabulary and grammar. They typically require the final user to take part in training sessions at universities or company premises. By taking into account that the mobility of people suffering from dysarthric speech is often limited, and their ability to record many speech samples in one session is questionable, such conditions limit the practical use of these system to few users and few speech recognition
tasks.

2. Objective of the paper

The objective of this paper is to introduce the concept of Personal Speech Assistants, a platform that enables nontechnical users to create their own speech recognition systems tailored to their needs and disabilities. Using this platform, users with disabilities can develop at home their own speech models. They become able to record the necessary training data, define appropriate vocabularies and language models for specified tasks, and train the models. All tasks are performed with no or little knowledge of speech technology. These models are then used in mobile applications that act as the user's own speech assistants for the given tasks. In particular, we discuss a mobile phone assistant created and tested by a person with cerebral palsy and with distorted, explosive speech. The application, controlled by voice, enables this user to read emails and other text messaging services, to send predefined messages to the people from a contact list, and to perform phone calls. The application builds upon the sphinx recognition engine (Huggins-daines et al., 2006), and it uses speech models prepared by the user. In real conditions, the recognition rate was above 84% for action commands and above 94 % for a list of navigation commands. The application was also tested in terms of time necessary to accomplish given tasks as compared with traditional manual input. Although the recognition was not perfect, voice input allowed the user to reduce the time for each action by about 30%.

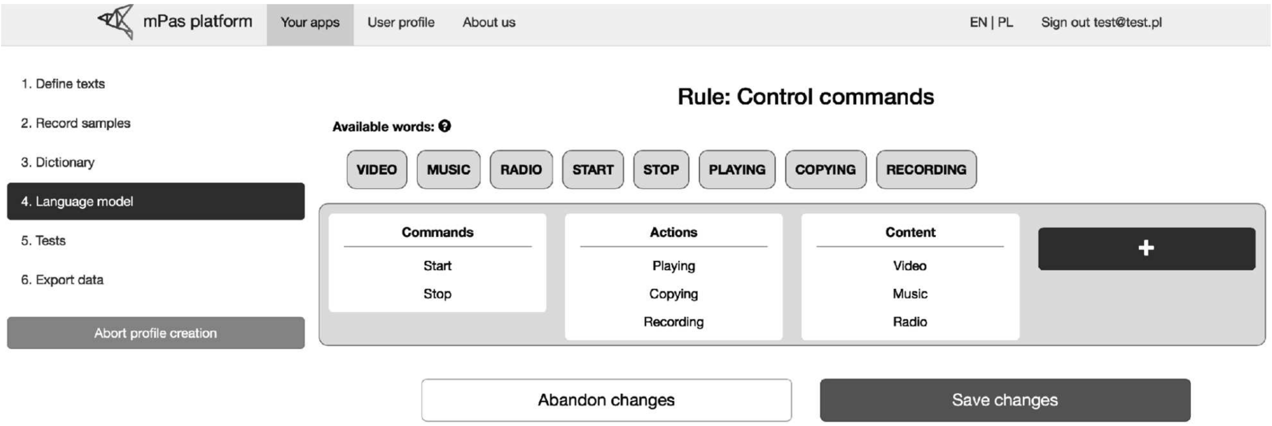


Figure 1: mPAS platform

3. Description of the personal speech assistent platform

3.1. mPAS platform

The main idea of the proposed mPAS platform (mPAS - mobile and personal speech assistant) is to guide nontechnical users through the processes of designing and building automatic speech recognition (ASR) systems, without requiring expert knowledge, especially in speech recognition technology. To simplify the whole process, we implement the platform as a web application. By doing so, users do not need to install complex software. All that is required is a computer with Internet connection, a fairly modern operating system, and a microphone. Having in mind the end-user, we put strong emphasis on user-centric interface designs, which enable users with various disabilities (such as poor sight, inability to read, poor motor skills, among others) to use the proposed system.

Briefly, the mPAS platform consists of the following main tools (Figure 1):

- Text creation/selection tool: this tool allows users to select texts from databases (provided by the mPAS platform or shared by other users), to download and share new texts, and to give suggestions on which text to use for a specific tasks (for example, which text is good to create phone-based speech recognition system). The text can be associated with pictures and with sound files providing pronunciation guidelines to nonreading users.

- Recording tool: the purpose of this tool is to enable users to record selected texts and to check/correct recordings of speech samples. The text is displayed in large fonts, and it is associated with large pictures together with voice examples. The recording session can be stopped/resumed at any time, so the recording process can span over a long period, which enables users to stop whenever they are tired. Recordings are stored at a server.

- Dictionary selection/creation tool: In this tool users select a dictionary (a set of possible words to be recognized) from a database provided by the platform or shared by other users. The dictionary can be also created automatically from a given text, with the words added or modified manually by the user. Transcriptions are done automatically based on the selected application.

- Language model selection/creation tool: the language model can be chosen from a database, or/and it can be created automatically from selected text (in the form of the statistical language model or grammar rules). Grammar rules can also be created manually by the user through graphical interfaces. The tool guides through the correct choice of the language model by taking into account the specific task.

- Acoustic model creation tool: based on the data collected by other tools, the acoustic models are created automatically. At the moment, HTK tools (Young et al., 2006) and CMU sphinx (Huggins-daines et al., 2006) are used as the standard training and decoding schemes. However, the mPAS platform can also be easily integrated with new speech recognition systems such as research-level systems based on support vector machines.

We emphasize that the proposed platform is not limited to specific speech recognition systems (e.g., command based systems, continuous speech recognition systems, etc). We note that the performance of these systems typically depend on the amount of training data available, hence one of the main ideas of the platform is to ease and to increase the amount of data collected, irrespective of the speech recognition system being used. By doing so, it is envisioned that users will design speech recognition systems for small tasks at first, which are fast to create and whose recognition performance are expected to be good (depending also to the level of user's speech impairment). By achieving high performance in small tasks, we expect to motivate users to record more data and to keep improving their speech recognition systems.

3.2. An example : mobile phone assistant

To make the system as practical as possible, we discussed the needs and expectations towards the system with potential users (people with speech disorders), their caregivers, and their therapists. Based on this survey, we identified many applications that could ease the daily life of people with speech impairments. One of them is a mobile phone assistant for people with motor disabilities and severe articulation dysfunctions. The goal of this application is to control mobile phones via voice in order to enable its user to communicate with other people via emails, SMSs, and phone calls. In its simplest version, the application should assist the user in emergency situations. Many caregivers pointed out that in case something happens to them, the application should enable the user with speech impairment to call for help. In advanced versions, the application should allow the user to send predefined emails and SMSs to their caregivers and also to people from contact lists. The application can be adapted to the needs of the given person by allowing the user to predefine list of possible messages, and the phone numbers of the main caregivers.

The graphical user interface (GUI) of the application was designed to minimize the chance of executing wrong commands and to be easy to use for users with impairments. The main GUI contains only 8 buttons representing 8 main actions (see Figure 2). Whenever the system recognizes a spoken command, the corresponding button receives focus. Only when the user confirms the choice by saying “OK” the action is executed. The screen contains one additional button showing the current status of the ASR (automatic speech recognition) system (processing data/listening), which gives the user hints on when to speak or to wait.

When the SMS/email option is selected, a list of contacts is presented to the user, which chooses the desired number/email by navigating the list with simple commands (up/down/OK) (Figure 3). Later a list of predefined messages is shown (Figure 4), and the user can choose the message in similar manner or by reading it. Next, by saying the command down, the user sets the focus on the button send, and he or she executes the action “send” by saying OK (Figure 5).

When the incoming SMS/email action is chosen, the system checks for new sms/emails, and it displays the list of senders. The user can once again navigate the list with simple commands (up/down/OK) to choose the name of the sender. Next, the messages from the chosen person are displayed on the screen. To return to the main screen, the user navigates to the button "exit," by saying "down" and he or she confirms the "exit" command by saying OK.

4. Experimental results

4.1. Mobile Speech Assistant Specification

We now turn our attention to the results obtained by a mobile phone assistant created and tested by a person with cerebral palsy and with distorted, explosive speech. The user operates his mobile phone, installed on a wheelchair, by using his chin, and his goal is to use the mobile speech assistant to simplify the communication with people (and, in particular, with his caregiver) during his unassisted walks. Having this purpose in mind, the user defined 8 messages:

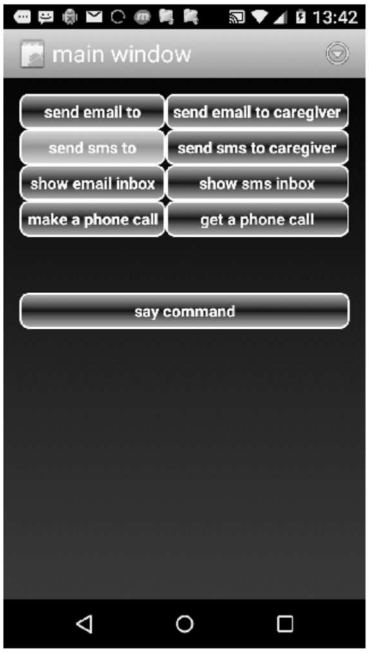


Figure 2: The main screen with 8 buttons representing available actions.

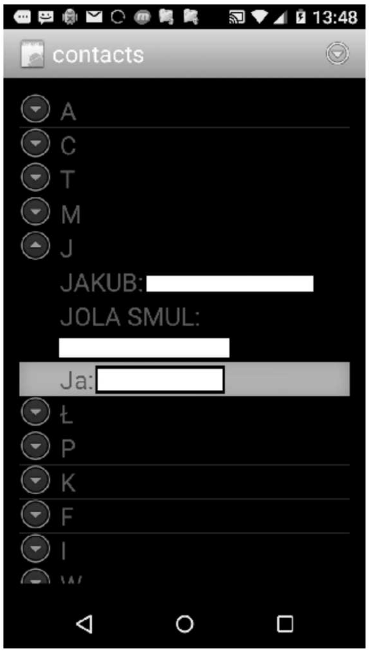
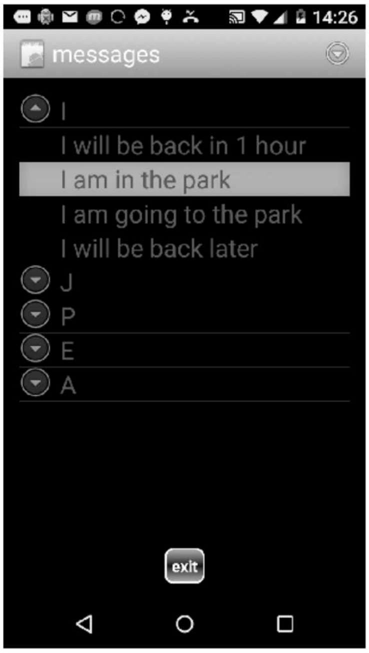


Figure 3: A contact list screen. Users can navigate through the list with commands "up", "down","OK".

- jestem w parku (I am in the park)
- będę za godzinę (I will be back in 1 hour)
- już jestem (just arrived)
- jesteś w domu (are you at home),
- pilne - pomocy (please, help me)
- rozładowana bateria (empty battery)
- wrócę poźniej (I will be back later).



I

I will be back in 1 hour

I am in the park

I am going to the park

I will be back later

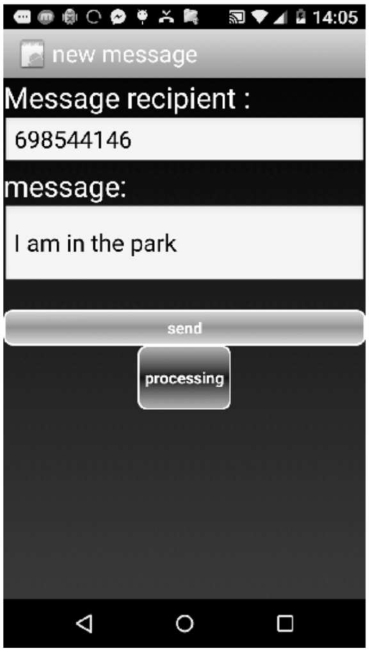
J

P

E

A

Figure 4: A list of predefined messages. Users can speak the message or navigate through the list with commands "up", "down", "OK"



Message recipient :

698544146

message:

I am in the park

Figure 5: The summary screen. The "OK" command executes the "send" action

- jadę do parku (I am going to the park)

Together with action commands, the user can utter 21 different phrases. The mobile speech assistant was implemented in Java with Android SDK and CMU sphinx packages. It was tested on the second generation of a Motorola Moto G smartphone (Android 5.0.2 Lollipop at the time of the testing), and Nvidia shield Tablet.

4.2. Experimental Results

For the training and evaluation of the mobile speech assistant, each phrase was recorded by the user 30 times. The recorded samples were digitized at 16000 Hz sampling rate, and they were analyzed at a 25 msec frame period with 10 msec shift. Mel Frequency Cepstral Coefficents (MFCCs) parameters consisting of 12 static features, energy, 12 Δ features, Δ energy, 12 $\Delta\Delta$ features, and $\Delta\Delta$ energy were used as the input features. The vocabulary contains 26 entries. Each phoneme was represented as monophone HMM with 3 states. Due to the limited training, 1 to 8 mixtures per state were used.

For the system evaluation in laboratory conditions, we divided the database into two sets. A fixed number (2,4,6,8, and 10) of examples of each phrase was choosen randomly as the training set, and the remaining examples were used as the evaluation set. The procedure was repeated 20 times, and the results were averaged.

For the system evaluation in more realistic conditions, we tested the execution of the application in two ways:

1. From the set of allowed commands, over 30 commands were spontaneously chosen and spoken to the application operating on the mobile platform. Based on the correct/incorrect execution of the commands, we evaluated the overall recognition performance. The test was performed not only by the above-mentioned speaker, but also by a subject without any speech impairments.

2. The performance of the application was also evaluated in terms of successfully accomplished actions, and in terms of its usefulness to the user. In particular, four actions were evaluated:

   - send SMSs to the chosen person from a contact list with the text "I am in the park" or "I will be back in one hour",
   - send an email to the chosen person from a contact list with text "I am in the park" or "I will be back in one hour",
   - send SMSs to a caregiver with text "I am in the park" or "I will be back in one hour",
   - send an email to a caregiver with text "I am in the park" or "I will be back in one hour",

   We measured the time required to accomplish the action by using mobile phone assistant and by using a traditional touch input.

4.2.1. Results

The results of the system performance in the laboratory conditions were ranging from 83% to 99%. As expected, the larger the number of training samples, the better the final recognition results tends to be. In the case of 10 examples, the system achieved 99% of accuracy for HMM with 4 mixtures per state. The worst results were achieved when only 2 samples were used, but the recognition accuracy could still be acceptable for users with strong motor disabilities.

Commands	Recognition results
Action	81%
List controls	88%
Pre-defined messages	80%
Total	84%

Table 1: The recognition performance in real environment (home/office). User with distorted explosive speech spoke 82 commands to the mobile speech assistant running on the smartphone.

Commands	Recognition results
Action	100%
List controls	89%
Pre-defined messages	96%
Total	94%

Table 2: The recognition performance in real environment (outdoors with relatively strong wind). Healthy subject spoke 72 commands to the mobile speech assistant running on the smartphone.

The results of the application evaluation in real conditions are presented in Table 1. The healthy subject spoke 72 commands, out of which 68 commands were recognized correctly (Table 2). The recognition results were better in home conditions, because fewer sources of nonstationary noise were present. The user with speech impairments spoke 82 commands, and the final recognition result was 84%. The performance of the system was lower than in the case of the healthy subject owing to the additional, involuntary sounds made by the user with speech impairments (e.g., loud breathing, noise made by wheelchair, among others).

Action	Time spoken input	Time manual input
Send sms to caregiver	31 sec	56 sec
Send sms to a person from contact list	53 sec	56 sec
Send email to caregiver	33 sec	65 sec
Send email to a person from contact list	60 sec	65 sec

Table 3: Time necessary to accomplish actions by using the speech assistant and by using traditional input.

We also measured the time necessary to finish predefined actions by using our application, and compared this time with the conventional manual input (see Table 3). In the case of the healthy subject, sending SMS to a person from contact list required on avarage 15-22 sec, and the manual input was slightly faster for short messages than voice input. This result is explained by the fact that the application allows only up and down movements within the list of contacts, while manual input enable users to choose directly the list item. In contrast, with the user with speech impairments, each action required at least 30 sec (including time lost with recognition errors) when speech assistant was used. Using manual input, the user required more time to accomplish the task. In particular, sending SMSs or emails to the caregiver, the time gain was above 45%. This result could be further improved by reducing the number of recognition errors occuring during the use of the application.

5. Conclusions

In this paper, we presented the concept of Personal Speech Assistants, a platform that enables nontechnical users to create their own speech recognition systems tailored to their needs and disabilities. As a specific example, we developed a mobile speech assistant that was operated by a person with cerebral palsy and with distorted, explosive speech. The proposed application was able to recognize 99% of spoken commands in laboratory conditions, but its performance degradated to 84% when used in real environments. This degradation in performance was caused by the additional involuntary sounds made by the user with speech impairments (e.g., loud breathing, noise made by wheelchair, among others) and other existing non-stationary noises. More precisely, the application treated these sources of noise as part of the speech signal, so better voice activity detection algorithms are required.

The proposed application was evaluated in terms of successfully accomplished actions and also in terms of its usefulness to the user. Experimental results showed that the average time of finishing a task using voice input is from 5% to 49% faster than using manual input. In addition, even when the execution was similar, the user indicated mobile speech assistant as the most comfortable option.

6. Acknowledgments

This work was part of the project "Mobile and Personal Speech Assistant" financed by the National Centre for Research and Development (NCBiR) within the framework of the LIDER IV program

7. References

Fager, S. K., D. R. Beukelman, T. Jakobs, and J.-P. Hosom, 2010. Evaluation of speech recognition prototype for speakers with moderate and severe dysarthria: A preliminary report. *Augmentative and Alternative Communication*, 26(4):267–277.

Havstam, C., M. Buchholz, and L. Hartelius, 2003. Speech recognition and dysarthria: a single subject study of two individuals with profound impairment of speech and motor control. *Logopedics Phonatrics Vocology*, 28(2):81–90.

Huggins-daines, David, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alex I. Rudnicky, 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *in Proceedings of ICASSP*.

Rosen, K. and S. Yampolski, 2000. Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative and Alternative Communication*, 16(1):48–60.

Tran, V.-A., G. Bailly, H. Lœvenbruck, and T. Toda, 2010. Improvement to a nam-captured whisper-to-speech system. speech communication. *Speech Communication*, 52(4):314–326.

Young, S. J., D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, 2006. *The HTK Book Version 3.4*. Cambridge University Press.