Proof-of-concept Evaluation of the Mobile and Personal Speech Assistant for the Recognition of Disordered Speech

Agnieszka Bętkowska Cavalcante

Gido Labs sp. z o.o.
ul. Romana Maya 1
60-133 Poznan, Poland
Email: a.b.cavalcante@gidolabs.eu

Monika Grajzer

Gido Labs sp. z o.o.
ul. Romana Maya 1
60-133 Poznan, Poland
Email: m.grajzer@gidolabs.eu

***Abstract*—Recently, Assistive Technologies tend to exploit speech-based interfaces as a means of communication between humans and machines. While they perform very well for normal speech, their efficacy is very limited for people suffering from a variety of speech disorders, especially in the presence of environmental factors related to the disease. To overcome these issues, we have proposed a Mobile and Personal Speech Assistant (mPASS) – a platform providing the users with a set of tools, which enable to intuitively create their own automatic speech recognition system (ASR) corresponding to their needs and capabilities. The system can be designed at users' home and tailored to the domain, vocabulary, and language they find most useful. The personalized speech recognizer can be used with diversified speech-based applications. The initial results depict the baseline performance of ASR systems created with the mPASS platform and help to identify the most accurate system set-up. Moreover, a proof-of-concept field trial shows that the mPASS speech recognition system was successfully used in the voice-controlled application, achieved high recognition accuracy and was identified by the user as better than the traditional touch input.**

Keywords–dysarthric speech recognition; personal speech assistant; speech recognition for assistive technologies; mPASS platform evaluation.

I. INTRODUCTION

The ability to speak, communicate and exchange thoughts is one of the fundamental needs of human beings. Unfortunately, it cannot be sufficiently satisfied in case of people suffering from a variety of speech disorders. As a result, communication situations, which are natural part of everyday activities, can become a formidable obstacle requiring help of an accompanying person. In addition, current technological achievements in the fields of ambient and assisted living, control of smart devices, smart homes, etc. tend to exploit speech-based interfaces as a core means of communication between humans and machines. Moreover, motor functions impairments, which call for the use of Assistive Technologies, are very often associated with speech production problems. Standard automatic speech recognition (ASR) systems, targeted for regular speakers, perform very poorly for people with speech disorders [1]–[4]. Hence, a significant group of people is not able to use many voice-controlled state-of-the-art technology advances, which could support their independence in handling daily activities [1].

It is estimated that 1.3% of the population encounters significant difficulties in speech-based communication [5]. The ability to use speech-based interfaces would significantly improve the lives of people suffering from speech impediments, in particular those with accompanying motor skills disorders. However, there are many diversified speech disorders and it is very challenging to design a single ASR system, which could recognize the impaired speech in each particular case [4]. Traditional methods of constructing ASR systems, used with success for normal speakers, fail in such a task – they require large-scale databases, which are not feasible to be created for disordered speech. Adaptation of standard ASR systems to the disordered speech led to the very limited system performance [4][6][7]. There have been several attempts to design a speaker-dependent dysarthric speech recognition systems [2]–[10], but they were trained mainly in the laboratory environments. Only a few of them were created and tested in real usage scenarios [3][5] with the limited achieved performance, which was not sufficient for the practical implementation [5]. Moreover, speech recognition systems usually require a lot of recordings to initially train the classifier, while collection of even a few speech samples may be very challenging in case of people with severe speech disorders and accompanying other diseases [1][3][4][6][7].

The design of a disordered speech recognition system with a good recognition performance for diversified speech impediments is very challenging. In order to increase the practical application of disordered speech recognizers in Assistive Technologies, we present a concept of a mobile and Personal Speech Assistant (mPASS) – a platform providing the users with a set of tools for building an ASR system, which is tailored to their speech disorders, needs, and capabilities. The mPASS toolchain is designed for non-technical user – the expert knowledge, in particular the knowledge about speech recognition, is not required. One of our key goals is a user-centric interface design allowing to use the platform by people with motor functions impairments and other disabilities. The user can choose the scope, in which he/she wishes to use the system, record training samples, and create personalized speech recognizer, which can be later used as a core engine for different speech-based endpoint applications. In case of people with severe motor disorders and/or accompanied intellectual disabilities the help of a user's carer or other person can be mandatory to operate the system, however, the technical background of such a person is not required [1].

The mPASS platform is to be exploited at users' home.

Therefore, the users are not obligated to attend long recording sessions at a remote location, which is a significant obstacle for the people with disabilities. By maximizing their comfort, more speech samples can be collected and, at the same time, users' motivation to work with the system is improved. In addition, the samples are recorded in the environment in which the ASR system will be later used – this should increase the recognition performance. Such an approach was never practised for a disordered speech thus far. By realizing this idea, we envision that we will be able to engage in our study many users, who will create different types of ASR systems, addressing diversified needs and being successfully used in many practical deployments [1].

In this paper we will also present the initial results depicting the baseline performance of ASR systems created with the mPASS platform for the group of 8 users with diversified speech impairments. They provide meaningful information, which will help to improve the design and accuracy of future disordered speech recognition systems. Moreover, we will present the results of a proof-of-concept field trial where an ASR system created with the mPASS platform was successfully used in the voice-controlled application.

This paper is organized as follows: Section II provides a brief overview of related work, Section III depicts lessons learned from the analysis of these approaches, which helped to come up with the user and system requirements for the mPASS platform, while Section IV presents the summary of identified design challenges. Sections V and VI present the mPASS solution and its architecture. The preliminary results are discussed in Section VII and Section VIII concludes the paper.

II. Related work

In the recent years, an increased attention has been put towards the design of disordered, in particular dysarthric, speech recognition systems (dysarthria is the key group of speech disorders) [2]–[12]. Two key initiatives in this field are the STARDUST project with its continuations (SPECS and VIVOCA) [3][5][13]–[17] and the Universal Access project [8][18]. The STARDUST system is based on the recognition of selected commands and, in latter versions, also phrases being a chains of words from the trained vocabulary. The system was developed mainly for the environmental control systems. Interestingly, the objectives were to teach users how to better articulate words towards increased speech recognition performance. Thus, the methodology here is opposite to the common ASR systems adaptation, where the recognizer tries to adjust itself to the particular user articulation and its variability. The most recent investigations revealed, however, that the system performance in realistic usage conditions have not met user requirements and therefore was not perceived as practically applicable [5]. This suggests that the proposed methodology was not sufficiently effective.

The second key initiative is the Universal Access project. It is the only system we have approached that was investigating also a possibility of more complex, phoneme-based, recognition, which is more challenging than word-based recognition. The project was focused particularly on the design of new speech recognition techniques, allowing for good performance with dysarthric speakers in large-vocabulary ASR systems. Nevertheless, the final performance results were often far from the levels achieved for normal speakers, especially in case of severe dysarthria. However, the recognition by the ASR system was still able to outperformed human listeners. This suggests that well designed systems for people with dysarthria can bring a significant improvement in the communication with others. Universal Access system was developed in a laboratory environment.

In general, the investigated related works [2]–[10] were mainly targeting the limited-vocabulary, discrete speech recognition systems focused on the command and control target applications. The final dysarthric speech recognition system was task specific and could have been used only with one, selected, speech-based application. This assumption was driving the methodology selection and ASR system set-up. A common practice was also to use the speech recognizers designed for natural speakers and adapt them to dysarthric speech (e.g., Dragon Dictate, Swedish solution Infovox or traditional models based on the Hidden Markov Model (HMM) solutions) [2][6][7]. The performance of these recognizers was limited, especially in case of severe speech impairments. Although, in general, the top performing systems presented 80-90% of accuracy, those results were obtained in the laboratory conditions. The trials conducted in more realistic environment revealed that the external factors (such as background noises) significantly degraded the investigated systems to unacceptable levels [5][6]. Substantially, the diminished performance did not allow for practical exploitation, as concluded from the year-long project VIVOCA [1][5][19].

III. Lessons learned and user requirements

Based on the detailed analysis of the related works, we have identified a list of user-related factors that the authors of other solutions perceived as important. They are particularly significant in case of disordered speech recognition systems, where user-related factors highly influence the achievable performance. The lessons learned during the analysis of the available literature helped to improve the mPASS system design. Hence, our observations constitute a set of tips and user requirements that the mPASS system should fulfill in order to increase its practical usage potential:

1) **The process of training, testing and using the system:** The speech recognition system should be trained and tested in the environment similar to the targeted environment of final speech-based application in which the recognizer will be used. This allows to catch and model the factors related to background noise, microphone type, sounds produced by the access technology interface and a person himself/herself, which were identified as very important [5]–[7]. Moreover, only those interfaces should be used, which are known to the users (e.g., a mouse dedicated for the people with motor skills disorders). This eliminates possible errors and frustration, which could result from using new, unknown access technology [6]. Additionally, it was also depicted, that although combined audio-visual interface is beneficial, the users encountered problems when both audio and visual information was available simultaneously [13]. Furthermore, gamification and similar technologies can positively influence user motivation and commitment [6]. From this perspective also the ability to train, test and use the system at user's home or school is of primary importance, since the need

for travelling to the training sessions can be a significant obstacle [6]. As reported in the literature, most participants of the system trials encountered problems with long training sessions. Hence, the system should use short training sessions and allow users to take breaks whenever necessary. Good and stable results were observed after longer work with the system – the process could have been spanned across the period of several days or weeks, but longer breaks (e.g., related to serious health issues) had negative impact on the performance [6]. Thus, systematic work with the system is important.

2) **Speech recognition techniques and system features:** The core speech recognizer technologies as well as other supporting techniques should be diversified and aligned to the severity of speech disorder, existing motor skill impairments (if any) and user needs. In particular, from the articulation perspective, the system should offer solutions capable of dealing with [4][7]: decreased intelligibility, limited articulation of some (or many) phonemes, explosive speech, slow speech, presence of additional sounds other than speech (loud, involuntary pause sounds, loud breathing, etc.) and unnaturally long pauses between words (disfluencies). Professional dysarthria assessment tests can be helpful in identifying particular problem a user encounters. Feedback information given to the user about the appropriateness of a produced sound volume, quality of recordings, background noises disturbing the system, speech recognition performance, etc., is meaningful.
3) **Problems encountered by the people with dysarthria:** The work with an ASR system was a new experience to the users – they had to get familiar with the technology and the interface. Hence, often the first training sessions were very fluctuating and some time was required to achieve a stable state [3][4][6][20]. Moreover, motor skill disorders are often associated with speech disorders. This influences mainly the interface design, but also the microphone usage – there were several problems reported with headsets and, thus, it is recommended to use stationary microphone [5][7]. In general, for people with motor skills impairments, the need for interaction with the system should be kept minimal (button pressing actions and similar). In spite of the interface improvements towards enhancing user comfort, in many cases the help of accompanying person can be necessary, at least during training phase.
4) **Selection of training material:** Selection of a text material, which should be recorded, can be challenging. It should be aligned uniquely to each user's needs, since predefined training sets can be difficult to pronounce for many users due to their particular speech impairment. In many studies (e.g., [3][6]) the users were allowed to change frequently misclassified or unrecognized words (in command-based systems) to different ones. Moreover, it should be allowed for the users to articulate their "version" of a given command, even if it is pronounced significantly different than by natural speakers (however, the "own" version has to be always the same). Considering the above findings, the selection of training words and/or sequences should be combined with the creation of language model and dictionary for the ASR system.

IV. Design challenges

The analysis of related works led to the conclusion that the system performance in normal, practical usage situations is influenced by the degree of speech disorder and motor functions impairments, environmental factors (e.g., noises), system access technology design, etc. User motivation was also thoroughly depicted by other researchers as a crucial element of a successful system usage. From the performance perspective, it was assessed as even more important than a degree of speech impairment – better motivated users with severe disorder can train the system better than less motivated ones with milder disorder [6]. These factors have significant impact on the design of an ASR system as a whole [1].

Taking into consideration the outcomes of the related works, it turns out that the challenges in the design of such a system for the disordered speech focus on two factors [1]:

1) the core speech recognition technology, which calls for the development of new techniques targeting disordered speech, especially with regard to acoustic modelling
2) disability-oriented, user-centric system design, taking into account the user needs, which allows for a comfortable usage in the presence of accompanying difficulties

Usually, the second factor is perceived as much less important, especially at the research stage of product development, and it does not influence performance. However, when designing the system for the demanding and diversified group of people with disabilities, its importance becomes equally relevant as the technical excellence of the core speech recognition technology. Hence, our goal is to address both these challenges and come up with a solution, which would conveniently combine novel research outcomes with the user-centric design. Substantially, we also perceive a positive practical verification of a solution as a key challenge and an important success measure [1].

V. mPASS approach – mobile and personal speech assistant

To address the above challenges, we propose a platform, which allows *non-technical users* to build their own speech recognition systems, tailored to their particular needs and speech disorders. Our vision is that disabled users, without computer science and artificial intelligence knowledge, will use the mPASS platform to define the domain, vocabulary, and language that is most useful for them in order to communicate effectively with the outside world. They will then train their own ASR system and adapt it to their individual way of speaking. The mPASS system allows to create different types of speech recognizers, at different levels of complexity, ranging from small-vocabulary, command-based systems, to dictation-based systems with different vocabulary sizes for the recognition of sentences and phrases. More complex systems are envisioned for people with mild and moderate speech disorders, since the users with severe speech disorders usually do not use speech in such broad contexts [1].

The personalized speech recognizer can be used later on with many diversified speech-based applications. The proposed mPASS platform is available on a desktop computer as a web-based application providing tools for creating user- and task-dependent speech recognition systems. The models created and trained with this application can be than ported to a mobile

device and used in the final speech-based application of interest (where the models for the disordered speech need to substitute or complement the ASR models for the natural speech). Hence, the speech recognizer built by using the mPASS toolchain *can be used with many different speech-based applications*, which were thus far not available to the users with speech disorders. Those applications are widely exploited in the environmental control systems, command-and-control systems (e.g., to steer some home appliances with voice commands), control of mobile device functions, converters transforming (possibly disordered) speech to text or to a synthesized speech, and many more [1]. Some examples of such end-point applications, currently being developed by us to showcase the capabilities of the mPASS technology, are [1]:

1) dictation-based, task-specific application allowing to "translate" impaired speech during a conversation in a restaurant, bank, at the doctor's office, etc.
2) educational game, targeted for autistic users, aiming at helping them in speech therapy classes
3) mobile communication application for users with very severe speech disorders and motor skills impairment (the user exploits a few sounds he/she can produce to control an image-based "communication book")

Having in mind the identified challenges, we present below the key objectives the mPASS system aims to accomplish. They also constitute the differences between our approach and the related works [1].

In contrary to other approaches, the process of building a disordered speech recognizer with mPASS should be *automated* and should limit the need for external help to minimum. Since the influence of practical usage constraints is tremendous, they should drive the system set-up [1].

The ASR system should be created *at user's home* and a training process can span across longer period of time, if necessary. Thus, the time spent on training the recognizer can be adjusted to the user's health condition, motivation and other factors. In addition, such an approach also minimizes the problem of reduced performance in case of systems trained in the silence conditions, which are used in the environment with existing background noise [1]. With the mPASS platform the recordings are to be made in the environment in which the system is then to be used. As a result, it is possible catch and model the factors related to background noise, microphone type, sounds produced by the access technology interface and a person himself/herself. This enables to create ASR systems, which will better deal with such environmental conditions.

Finally, the mPASS toolchain is intended to allow for the *exploitation of existing resources*, which are proved to be good for creating speech recognition systems. *Novel approaches* are to be provided only where necessary, e.g., while building acoustic models for dysarthric speech, where we are developing a new method of the dysarthric speech recognition based on the modified speech classification methods [1].

At this stage the targeted language is polish, however the platform by design is language-agnostic and could be used for building speech recognizers for other languages as well [1].

VI. System architecture

The mPASS platform guides the user through the steps required to build the speech recognition system (Figure 1). During the process, the user follows on-screen instructions. The core part of the platform is a web-based application – a client side is implemented by using *AngularJS* framework and the server side is based on the *Node.js* framework. The voice is captured by the HTML5 function *getUserMedia*. The client and the server exchange data in the JSON format. The speech recognition system trained with this application is then incorporated with a target speech-based application, on a mobile or embedded device [1]. The below steps present how the process is organized and which consecutive actions are expected to be executed by the user:

1) The user has to create a *profile*, which is strictly related to the level and scope of the envisioned system usage (e.g., command-based, recognition of sentences, continuous speech). There can be different profiles created for the same user, each targeting different kind of speech recognizer for different tasks (e.g., containing vocabulary/training sets for controlling TV, going to doctor's office, restaurant, etc.). Based on the selected system level, the baseline speech unit is automatically defined as word, syllable or phoneme [1].
2) **Creating texts to be recorded, dictionary and language model**: These elements are usually combined and they influence each other. For instance, in command-based ASR systems it could be most convenient to start with a vocabulary, while for the other ones it could be better to start with a set of texts for recording. The mPASS toolchain further guides through the next steps, including support for intuitive creation of language model and dictionary. The final relation between text selection, dictionary and language model is proposed automatically [1].
   a) **Text selection tool**: it is equipped with several phonetically balanced and phonetically rich texts for polish language. They have been created by us based on a well-known poems and short stories for children in order to make them easy to pronounce by the users with disordered speech. The phonetic balance of the recorded text should be duly considered, especially while building more sophisticated systems [13]. It is also possible to create the text automatically based on the existing dictionary and language model [1].
   b) **Dictionary tool**: Dictionary contains the list of words that the system will be able to recognize. It can be created either manually or by extracting words from the texts selected previously for recording or from the language models defined by the user. It is also envisioned that the dictionary tool will automatically suggest additional entries that could maximize ASR performance. For that purpose the dictionary will be analyzed by the mPASS platform in terms of length of the words, phonetic differences between them, and others. There is also an option to substitute frequently unrecognized words with their synonyms based on the user input or automatic suggestion from the mPASS system [1].
   c) **Language model tool**: The purpose of this tool is to create grammar or statistical *n*-gram language models. In the first case the user is supported to manually create grammar rules via dedicated interactive graphical interface (technical knowledge is not necessary at this step). Grammar consists of a set of rules that define the possible combination of words in the dictionary. The related mPASS tool enables to create such rules –

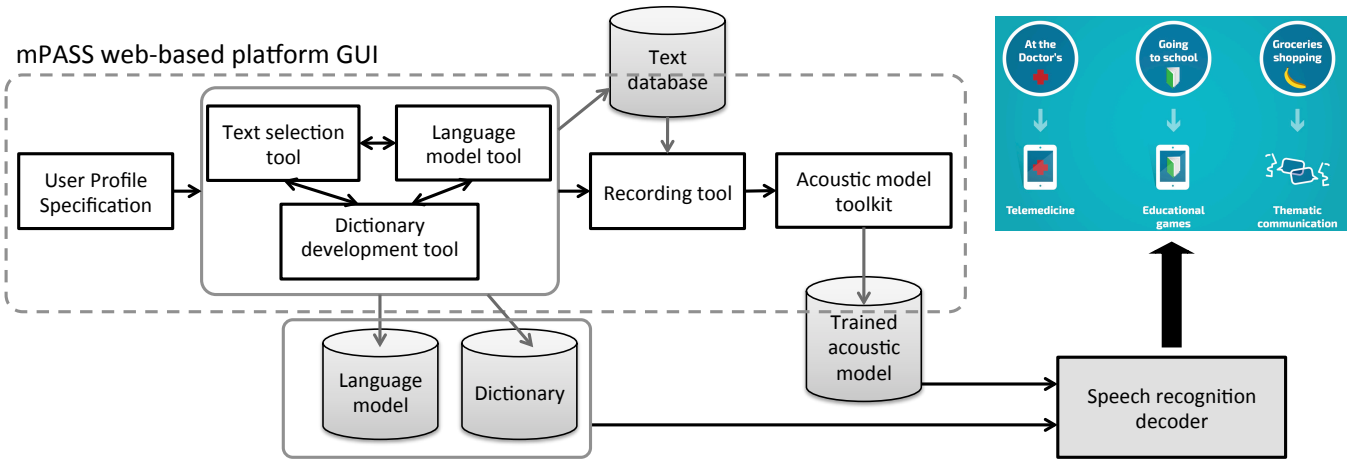


Figure 1. Mobile and Personal Speech Assistant architecture – an overview [1].

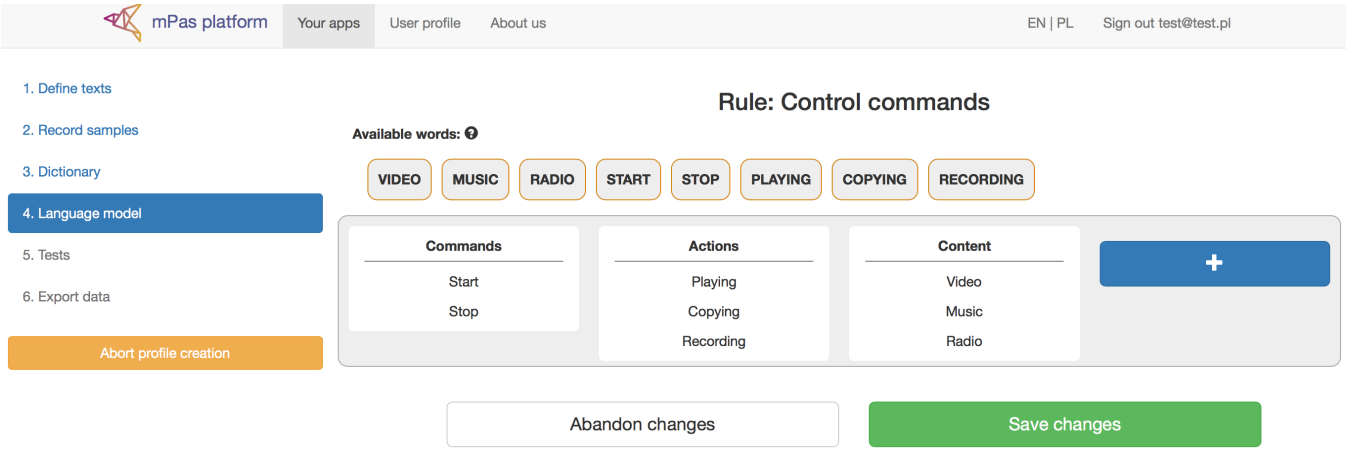


Figure 2. The screen from the mPASS web-based application [22].

the first example is created automatically based on the training texts or by using some predefined examples. The user is then allowed to extend or modify it. The user does not need to understand the methods and formats of grammars, but can intuitively follow the tool's suggestion. mPASS will also automatically verify and, if necessary, correct the convergence between dictionary and language model, so that the latter does not contain words that are not present in the dictionary and vice versa. Figure 2 depicts an exemplary screen from the grammar-based language model creation in the mPASS web-based application.

Alternatively, the mPASS system can automatically modify the pre-loaded generic statistical *n*-gram model for a given language, in order to align it to the scope of the desired ASR system. The statistical *n*-gram model specifies the probability of particular *n*-gram sequences.

3) **Recordings**: The user records selected texts and/or word lists. There is a minimal suggested number of recordings specified. In addition, the system gives a possibility to add new recordings at a later time, pause and resume the recording sessions. The tool also allows to play additional audio information on the attached headphones. The supplementary audio-visual information is supposed to help people with intellectual disabilities, visual impairments, children, etc. We also aim to supply the tool with mechanisms allowing for monitoring and potential correction of wrong recordings [1] – the user will be given a real-time feedback information presenting the recorded waveform and whether the required volume is achieved. Hence, feedback information will refer to the appropriateness of a produced sound volume, quality of recordings, background noises disturbing the system, speech recognition performance, etc.
4) **Training the acoustic model**: This step is an automated background process. Only experienced (developer-type) users are allowed to change some of the parameters, e.g., choose different methodologies/techniques, such as HMMs or Support Vector Machines (SVMs). We are also developing our novel acoustic modelling methods, which will be included in the mPASS system [1].
5) The obtained acoustic model, dictionary and language model are then *exported* to be used in the desired target speech-based application. Optionally, the initially created acoustic model can be later on extended based on additional recordings collected while creating other user profiles for different contexts [1].

TABLE I. DESCRIPTION OF USERS, WHO PROVIDED VOICE SAMPLES COLLECTED WITH THE MPASS PLATFORM

User	Speech disorder	Severity of disorder	Age	No. of sessions
User 01	dyslalia	low	child	10
User 02	sigmatism, devoicing of phones	moderate	youth	27
User 03	dysarthria	moderate	youth	20
User 04	dysarthria	severe	youth	22
User 05	dysarthria	low-moderate	adult	30
User 07	dysarthria, praxic functions disorders	low	youth	12
User 12	dyslalia	low-moderate	child	10
User 13	dysarthria, prosodic disorders	low	youth	21

All recordings, recorded texts, dictionaries and language models are stored in a database. The user may wish to share them with others (if agreed) in order to help develop better ASR systems for the other users in the future [1].

From the user perspective, the recording tool functionality is the most important part of the mPASS platform. It is, however, also the most vulnerable to possible errors – wrong recordings, additional background noise and other factors affecting the recorded material will directly influence the acoustic model and its performance. Hence, in order to tune our interface design and system features to real user needs, we have performed initial recording sessions with several users having diversified speech disorders: one adult with explosive speech and associated motor impairments, 4 teenagers presenting variable levels of dysarthria and 4 healthy children 3-6 years old with impaired speech typical to their age. Those trials helped to improve the system design and obtain an initial database used by us for the evaluation of acoustic modelling techniques. Currently, the key components of the mPASS platform are implemented and it can be used for further evaluation [1].

VII. PRELIMINARY RESULTS

A set of initial experiments has been conducted with the use of voice samples recorded with the mPASS platform. The goal was to:

1) evaluate the recognition accuracy with regard to the selection of the basic recognition unit as either phoneme, word or syllable
2) evaluate 3 acoustic modelling methods based on Hidden Markov Models (HMMs), Support Vector Machines (SVMs) and Structured SVMs
3) perform an initial field trial conducted by a single user, who was using the voice-controlled application for sending SMS-es and e-mails

This initial evaluation constitutes a proof-of-concept evaluation of the mPASS approach, provide important insights towards the most accurate set-up of speech recognition systems for the disordered speech and present the borderline of the achievable performance. During the course of further research we will investigate further improvements of the mPASS platform, especially with regard to acoustic modelling, which should further improve the recognition accuracy.

A. System performance for diversified recognition units

In each speech recognition system the recognizer is trained with regard to particular recognition unit, which can be specified as either phoneme, word or syllable. The aim of experiments reported in this section was to verify, which recognition unit would be the most suitable in case of the disordered speech recognition.

We have investigated 4 different variants of basic recognition units:

- word: each word is represented by an HMM model with particular number of states equal for each word, e.g., the word "nine" is represented with a single HMM model with *n* states and the word "ten" is represented with another HMM model with *n* states, where *n* is an arbitrary selected value adjusted experimentally (i.e., *nine* $\rightarrow$ *nine*, *ten* $\rightarrow$ *ten*).
- phoneme: each phoneme is represented by a single HMM with 3 states; each word in a dictionary is represented as a sequence of phonemes, however, different words can include the same phoneme (i.e., *nine* $\rightarrow$ *n a y n*, *ten* $\rightarrow$ *t E n*) (phonemes are represented in SAMPA notation).
- phoneme/word: each phoneme is represented by a single HMM with 3 states; each word in a dictionary is represented as a sequence of phonemes, however, each phoneme can be present only once, e.g., *nine* $\rightarrow$ *n_1 a_1 y_1 n_11* (HMM for a word with 12 states), *ten* $\rightarrow$ *t_2 E_2 n_2* (HMM for a word with 9 states). This system can be envisioned as a word-based system, where HMM for each word is represented with different number of states.
- syllable: each syllable is represented by a single HMM with particular number of states equal for each syllable and each word is represented as a sequence of syllables; different words can have the same set of syllables (i.e., *nine* $\rightarrow$ *nine*, *eleven* $\rightarrow$ *e lev en*).

For the needs of initial experiments, the recordings database was collected by means of the mPASS platform. It contains the speech samples of 8 users with diversified speech disorders. The group consists of 2 pre-school children with impaired speech typical to their age and 6 persons with variable speech disorders and other dysfunctions. Their speech impairments were described and characterized by the language therapist. The short characteristic of each user is presented in Table I. All users were recording the training sessions containing numbers from 1 to 10. The speech samples were recorded at users home. The sessions were then divided into 3 sets: training set, development set and test set. Hence, each set was containing several sessions with 10 samples each

TABLE II. PERFORMANCE COMPARISON OF HMM- AND SVM-BASED DISORDERED ASR SYSTEMS CREATED WITH THE MPASS PLATFORM

User	word	phoneme	phoneme/word	syllable
User 01	97.38% (26 states)	94.92%	**98.19%**	94.98% (11 states)
User 02	90.25% (7 states)	**94.58%**	94.33%	93.00% (8 states)
User 03	89.50% (6 states)	91.41%	**93.24%**	90.46% (10 states)
User 04	62.25% (28 states)	54.94%	62.84%	**65.85% (21 states)**
User 05	91.37% (13 states)	94.10%	**96.47%**	91.81% (12 states)
User 12	66.96% (19 states)	75.41%	**79.45%**	69.65% (16 states)
User 13	84.98% (19 states)	90.07%	**92.45%**	81.09% (15 states)

(numbers 1-10).

The HMM-based ASR systems were created for the users identified in Table I. In case the basic recognition unit was set to word or syllable, it was necessary to define the most suitable number of HMM states, which would be allocated per recognition unit. It was done experimentally by evaluating the recognition performance for HMM models containing 3 to 27 states, and for different number of mixtures – 1, 2, 4 or 8 per state. The best results were selected for the comparison and are depicted in Table II. For the other 2 cases (i.e., phoneme and phoneme/word), the number of HMM states was variable, depending on the length of a particular word.

Evaluation of the obtained recognition results reveals that the phoneme/word basic recognition unit provides the best outcome for most of the users. For User 02 better values were obtained for a phoneme-based system, however the differences in comparison to the phoneme/word version are negligible. In case of User 04, the best performance was obtained for a syllable-based system – most likely phoneme as a basic unit is too small to properly reflect variability of different units in case of severe speech disorders (more specifically, the syllable is not only longer, but it also always has a vowel, which makes recognition easier). Nevertheless, for both User 04 and User 02 the word/phoneme-based system performs still relatively well. Hence, this recognition unit should be recommended for the small-size ASR systems. Additionally, such a recognition unit is also used in the popular, open-source CMU Sphinx speech recognition software [21]. In case of more complex speech recognition system the word/phoneme-based model could be substituted with a phoneme-based systems in order to reduce system complexity.

Following the results presented in Table II, the mPASS platform will select the default recognition unit as phoneme/word, however will also automatically evaluate other options in case the recognition accuracy will not be satisfactory. In such a case the platform will also automatically specify the most accurate number of HMM states, where necessary.

B. System performance for diversified acoustic modelling techniques

The key acoustic modelling techniques were selected for the initial performance evaluation of disordered speech recognizers created with mPASS platform, namely HMMs and SVMs. HMM method is traditionally used in many ASR systems – it aims to model the speech recognition process as a sequence of most probable states of the hidden Markov process. The SVM method is a promising solution, which exploits discriminative supervised machine learning technique to classify observed speech samples into the most probable classes (labels) representing the final output. One of the variants of this technique, using the additionally structured label space, is a Structured SVM methodology. It will also be evaluated in the performed study. The SVM-based techniques were not widely exploited for the speech recognition thus far. Hence, their comparison to traditional HMM-based methodologies will provide meaningful insights into the future development of improved acoustic modelling techniques.

The experiments were performed for the database of recordings presented in Section VII-B. The speech recognition system was created and trained on the sessions from both – the training set and development set (this is driven by the objective to properly compare results between the HMM- and SVM-based systems, which will be described later on in this section). In case of HMM-based acoustic models, the basic recognition unit was phoneme/word, where each phoneme was represented by a 3-state HMM model with 1, 2, 4 or 8 mixtures – the particular value for each case was selected experimentally, depending on the size of the training material. For the training and recognition process the Mel-Frequency Cepstral Coefficients (MFCC) acoustic feature were used. The trained ASR system was then evaluated based on the recognition of samples from the test set.

For the speech recognition based on the SVM and Structured SVM acoustic models, the system is trained in the 2-step process. Firstly, the HMM-based model is trained, similarly as in the previous case, on the training set only. Secondly, this model is used for the recognition of samples from the development set, which allows for the computation of features from the log-likelihood feature space for these recordings. This input is then used to train the SVM or Structured SVM model. Finally, the acoustic model trained in this 2-stage process is used for the recognition of samples from the test set.

For both HMM- and SVM-based ASR systems, the above procedure was repeated 5 times, each time using different selections of sessions for training, test and development sets. The average recognition performance obtained in this experiment is presented in Table III. The Structured SVM model achieved the best performance in most cases or performed comparable to the reference HMM-based solution in the remaining cases. Hence, it can be perceived as the best solution among the 3 investigated ones. Interestingly, the pure SVM model performed poorly, often worse than the traditional HMM-based method. The advantages of the Structured SVMs over HMMs are particularly visible for the least-performing users, with the most severe speech disorders. This feature makes this technology an interesting alternative to HMM-based models in

TABLE III. PERFORMANCE COMPARISON OF HMM- AND SVM-BASED DISORDERED ASR SYSTEMS CREATED WITH THE MPASS PLATFORM

User	Training set size	Development set size	HMM	SVM	Structured SVM
User 01	4	2	86.68%	85.31%	**93.02%**
User 02	11	3	**93.10%**	89.36%	92.83%
User 03	10	3	**88.73%**	84.50%	88.61%
User 04	10	5	47.94%	47.21%	**49.51%**
User 05	15	5	98.33%	98.07%	**98.48%**
User 07	8	2	70.13%	68.06%	**74.53%**
User 12	6	2	66.59%	62.3%	**72.29%**
User 13	11	3	87.07%	79.31%	**89.41%**

TABLE V. RECOGNITION PERFORMANCE IN A MOBILE APPLICATION FOR THE SPEECH RECOGNITION SYSTEM CREATED WITH THE MPASS PLATFORM – COMPARISON BETWEEN A USER WITH DISORDERED SPEECH AND A HEALTHY SUBJECT

Action	User with disordered speech	User with normal speech
Action commands recognition	81%	100%
List control commands recognition	88%	89%
Pre-defined messages recognition	80%	96%
Average	84%	94%

the context of disordered speech recognition. Hence, as a part of future work we will be proposing a modification to the Structured SVM methodology in order to further increase the recognition accuracy.

System performance observed for User 01, 02, 03, 05 and 13 was high, exceeding 88% for all cases. The best result was as high as 98.48% for User 5. This is a very good outcome for disordered speech recognition. The worst performance was achieved for User 04, who has the most severe level of disarthria among all investigated cases. Lower accuracy was also observed for two cases – User 07 for who praxic functions disorder accompanies dysarthria making the speech disorder more complex and User 12 being a child with dyslalia. The low result of User 12, who has relatively low speech disorder, can be explained by the problems encountered during recording sessions, related to the age of the user – the recordings are of diversified sound volumes, the user was moving, etc. In general, however, the results present that the recognition performance is highly dependent on the severity of speech disorder of particular users, which confirms observations from the previous studies, e.g., [7]. In practice, the users with the most severe speech disorders would usually train the system with very limited, self-selected vocabulary, which would allow to introduce several control commands. The recognition accuracy in such set-up would likely be improved (in this trial we have pre-selected the vocabulary, so the user had to align to this selection, even if it was difficult for him/her to pronounce particular words). Nevertheless, our further works on this topic will focus on the new acoustic modelling techniques, which could improve recgnition performance for the most challenging group of users.

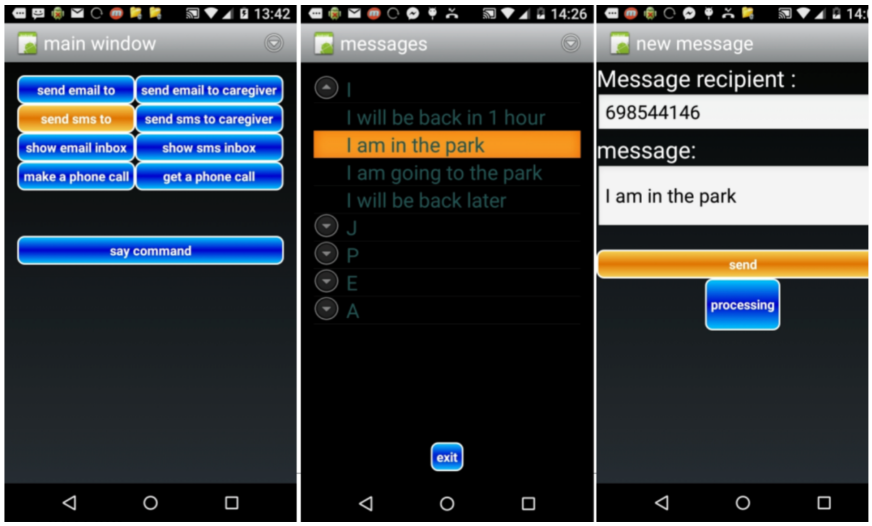
C. Proof-of-concept field trial

The initial proof-of-concept field trial was executed by the adult with explosive speech and cerebral palsy. With the mPASS platform, he created an ASR system for the exemplary voice-controlled mobile application, which allows to send an SMS or e-mail with one of predefined messages to a recipient from a phone contact list [22]. User-defined voice commands are exploited to control the application. Some of its screenshots are presented in Figure 3. The user recorded 8 messages of his own choice (e.g., “I will be back in 1 hour”), as depicted in Table IV, several action commands (“send”, “SMS to”, etc.) and list control commands (“up”, “down”, “OK”, etc.) – all together 21 phrases, 30 times each [1].

The ASR system was using the HMM-based acoustic model. The 3-state HMM model was used to represent each phoneme and the MFCC-based acoustic features were exploited. The basic system performance, with regard to accuracy of the recognition of the selected phrases and commands, was tested in a laboratory environment with the pocketsphinx speech recognition toolkit [21] (please note that the recordings were made at user’s home). For this purpose, the speech recognizer was trained with 2, 4, 6, 8 and 10 recording sessions selected randomly for each phrase. The remaining sessions (out of 30 collected) were used for testing. Each experiment was repeated 20 times and the results were averaged [22]. In case the training was performed for 10 recordings of each phrase, the recognition performance was 99%. It dropped to 82% when training on the smallest set of 2 sessions. Hence, for further

TABLE IV. MESSAGES SELECTED BY THE MOBILE APPLICATION USER, WHICH CAN BE SENT OVER SMS OR E-MAIL

No.	Message
1	I am in the park
2	I will be back in 1 hour
3	Just arrived
4	Are you at home?
5	Please, help me
6	Empty battery
7	I will be back later
8	I am going to the park



(a) The set of selected screensshots

(b) User controlling the computer and the application on a mobile device with his chin

Figure 3. An exemplary application using mPASS platform to dictate and send SMS-es and e-mails.

evaluation we have used an ASR system trained on randomly selected 10 recordings of each phrase/command.

The recognition performance under the real-usage conditions was evaluated with a disordered-speech user and a healthy user as a reference [22]. Such methodology is widely used for results comparison in case of speaker-dependent ASR systems, which are highly correlated with the context and vocabulary they are trained on. During the field trials both users were performing a given task with the dedicated application – they were sending an SMS and e-mail with a self-selected message. During the entire trial, the disordered-speech user spoke all together 82 commands (words or phrases), while the reference subject spoke 72 commands. The number of all spoken commands is smaller in case of healthy person, since less repetitions were necessary to complete a task. The overall recognition performance was evaluated based on the number of correctly recognized commands in relation to the total number of spoken commands.

The comparative results are presented in Table V. The disordered-speech subject was using the system in a home-office environment. Under such conditions, healthy subject achieved the accuracy close to 100%. Therefore, the experiment with a healthy person was also repeated in more demanding conditions – outdoors with relatively strong wind. These results are presented in Table V. The recognition accuracy in case of the user with disordered speech was on average 84%. It increased to 94% in case of normal speech, even though the user with unimpaired speech was testing the application in more demanding outdoor environment. Although the performance was lower in case of disordered speech, the achieved levels enabled to successfully control and use the application.

Additionally, we investigated performance measures related to the person's judgement of system's applicability and usability. We compared the time required to complete particular actions when using the dedicated voice-controlled application and the regular touch input (the person controls mobile phone installed on a wheelchair with his chin). In this measure we have also included the time lost for necessary repetitions when speech recognition errors occurred. The results were averaged over 20 trials and are given in Table VI. It can be observed that the voice-controlled version outperformed the manual entry for up to 49% – considering the time gain, which was observed with the voice input in comparison to manual input. Substantially, the user assessed a voice-controlled mobile speech assistant as the preferred option, which is the most important success measure [1].

TABLE VI. COMPARISON OF THE TIME REQUIRED TO COMPLETE AN ACTION WITH A VOICE-CONTROLLED AND MANUAL ENTRY [1]

Action	Voice input	Manual input	Gain
Send SMS to caregiver	31s	56s	45%
Send e-mail to caregiver	33s	65s	49%

D. Discussion

The initial trial presented above constitutes a first proof-of-concept evaluation. At this stage, the obtained results cannot be directly compared to the ones presented in the related works, since they were gathered for different usage scenarios and with different ASR systems, especially with regard to the selected vocabulary. However, we have also made an attempt to compare the proposed system with standard state-of-the-art speaker-independent speech recognition solution provided by Google (using the Google Speech API [23]). For this purpose, the disordered-speech user spoke several commands used during the proof-of-concept trial to the Google ASR system – 3 times each. Although this system is a top-performing speech recognition solution for normal speech, it was unable to recognize the disordered speech – in each single executed trial the Google system response was incorrect, leading to recognition performance of 0%. This constitutes

a confirmation that traditional ASR systems fail for disordered speech and further comaprisons with them cannot bring any additional information during the evaluation of disordered-speech recognition systems created with mPASS. Therefore, we have decided to compare the field trial results obtained for disordered speech with a healthy subject using the ASR system trained with the same methodology.

In general, the recognition performance of ASR systems created with mPASS reached very high levels for the laboratory environments. This finding is a consequence of the proposed mPASS system design, which allows to 1) create ASR systems with a scope corresponding to user expectations and capabilities, 2) collect speech samples in the environment in which the system is later used, 3) easily record the necessary number of speech samples in a convenient manner (the system automatically verifies, if it is necessary to collect additional samples in order to obtain the required recognition accuracy) and 4) use new SVM-based techniques for acoustic modelling. Substantially, the proposed ASR system also performed very well in real usage environment of home/small office – the proof-of-concept trials were concluded with a very promising outcome, which was rarely achieved before. However, we could also observe the drop of recognition accuracy in case of people with severe dysarthria. This effect was also widely observed in other trials reported in the literature and suggests that the scope of their ASR systems should be carefully adjusted to the sound pronunciation capabilities of these users. More detailed performance evaluation, including more complex ASR systems created with the mPASS platform, is a part of the future work. It is envisioned to be executed based on the database of recordings collected from another 7-10 users.

In addition, based on our observations, we have identified voice activity detection (VAD) functionality of the recording tool as one of the key challenges. In the trials performed to date the standard VAD techniques based on the analysis of the differences in volume level and signal to noise ratio often failed in case of disordered speech users. The level of additional involuntary sounds such as loud breathing, grunts, etc. and the noise occurring during recordings (e.g., sounds given by the computer access technology or a wheelchair) is frequently of significant volume. Hence, more sophisticated VAD techniques should be used to overcome such issues. Potential techniques can use machine learning technologies and reasoning based on the pre-recorded “silence patterns" for a given person. Such solutions are within a scope of our future research.

VIII. Conclusion

The mPASS system proposes a unique combination of an intuitive, user-centric system design with the top performing ASR tools. It provides an automated toolchain, which enables to easily follow the process of creating a speech recognition decoder. We believe that by using this technology the wide variety of users, with different speech impairments, will be able to build disordered speech recognition systems – tailored to their needs and achieving high recognition performance. Substantially, the users will be allowed to create and train the system at home environment. The initial results are very promising, especially taking into account a positive users’ feedback.

In the initial experiments we have investigated two types of acoustic models for the needs of disordered speech recognition. Our findings revealed that the Structured SVM method outperformed the traditional HMMs for the vast majority of cases. The performance of ASR systems created with the mPASS platform for 8 users allowed to reach high levels – often close to or higher then 90%. This is a very good result for disartric speech. Additionally, the comparison with HMMs shows that SVM-based techniques are an interesting methodology, which will be further investigated by us in the future. However, it was also observed that the achieved performance drops with the increase of speech disorder, which suggests that users with more severe speech impairments should align the complexity of their ASR systems to their capabilities. The mPASS platform, due to its flexibility, should allow to address this challenge accordingly. Moreover, the performance trials executed with the collected database of recordings allowed to investigate the most applicable system set-up with regard to the basic recogition unit selection. The results present that the recognition using a combination of phoneme and word would address well the variety of cases and speech disorders.

Additionally, the proof-of-concept field trial, with a dedicated voice-controlled mobile application, revealed a promising outcome. The speech-based input was assessed as up to 49% faster than the traditional manual input by a person with severe speech impediments and motor skills disorder. In the future, we plan to evaluate the mPASS platform with more users in several scenarios related to different mobile applications, which will be based on the ASR systems trained with mPASS. By using the proposed toolchain, we hope to achieve disordered speech recognition systems ready to be used in practical conditions with a variety of endpoint speech-based applications. Hence, our solution could be effectively exploited by people with speech impairments and assist them in their daily activities.

Acknowledgements

The presented work is financed by the National Centre for Research and Development in Poland under the grant no. LIDER/032/637/l-4/12/NCBR/2013. The authors would like to thank Michał Koziuk for his help with the system implementation and Leszek Lorens for his help with mobile application development and field trial organization.

References

[1] A. B. Cavalcante and M. Grajzer, “Mobile and Personal Speech Assistant for the Recognition of Disordered Speech,” in Proceedings of the Second International Conference on Smart Portable, Wearable, Implantable and Disability-oriented Devices and Systems (SPWID 2016), May 22-26, 2016, Valencia, Spain, 2016, pp. 6–10.

[2] S. K. Fager, D. R. Beukelman, T. Jakobs, and J.-P. Hosom, “Evaluation of a speech recognition prototype for speakers with moderate and severe dysarthria: A preliminary report,” Augmentative and Alternative Communication, vol. 26, no. 4, 2010, pp. 267–277.

[3] 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. 29, no. 5, 2007, pp. 586–593.

[4] K. Rosen and S. Yampolsky, “Automatic speech recognition and a review of its functioning with dysarthric speech,” Augmentative and Alternative Communication, vol. 16, no. 1, 2000, pp. 48–60.

[5] M. S. Hawley, S. P. Cunningham, P. D. Green, P. Enderby, R. Palmer, S. Sehgal, and P. O’Neill, “A voice-input voice-output communication aid for people with severe speech impairment,” Neural Systems and Rehabilitation Engineering, IEEE Transactions on, vol. 21, no. 1, 2013, pp. 23–31.

[6] C. Havstam, M. Buchholz, and L. Hartelius, "Speech recognition and dysarthria: a single subject study of two individuals with profound impairment of speech and motor control," Logopedics Phonatrics Vocology, vol. 28, no. 2, 2003, pp. 81–90.

[7] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," Augmentative and Alternative Communication, vol. 17, no. 4, 2001, pp. 265–275.

[8] H. V. Sharma, M. Hasegawa-Johnson, J. Gunderson, and A. Perlman, "Universal Access: Preliminary experiments in dysarthric speech recognition," in Proc. 10th Annual Conf. of the Internat. Speech Communication Association, 2009, p. 4.

[9] K. Caves, S. Boemler, and B. Cope, "Development of an automatic recognizer for dysarthric speech," in Proceedings of the RESNA Annual Conference, Phoenix, AZ., 2007, p. n/a.

[10] E. Rosengren, "Perceptual analysis of dysarthric speech in the ENABL project," TMHQPSR, KTH, vol. 1, no. 2000, 2000, pp. 13–18.

[11] AAC-RERC Project D2-B: Recognition of Dysarthric Speech. [Acccessed: 28 November 2016]. [Online]. Available: http://aac-rerc.psu.edu/index-28962.php.html

[12] ENAbler for computer-Based vocational tasks with Language and speech (ENABL). [Acccessed: 28 November 2016]. [Online]. Available: http://cordis.europa.eu/project/rcn/35115_en.html

[13] M. Parker, S. Cunningham, P. Enderby, M. Hawley, and P. Green, "Automatic speech recognition and training for severely dysarthric users of assistive technology: The STARDUST project," Clinical linguistics & phonetics, vol. 20, no. 2-3, 2006, pp. 149–156.

[14] STARDUST Project. [Acccessed: 28 November 2016]. [Online]. Available: http://spandh.dcs.shef.ac.uk/projects/stardust/

[15] Speech Technology Applications Toolkit - STAPTK. [Acccessed: 20 April 2015]. [Online]. Available: http://spandh.dcs.shef.ac.uk/projects/staptk/

[16] S. Judge and Z. Robertson, "Speech Driven Environmental Control System (SPECS): From Specification to Prototype," in Recent Advances in Assistive Technology and Engineering (RAATE) 2009, 2009, [Acccessed: 28 November 2016]. [Online]. Available: http://eprints.whiterose.ac.uk/10330/

[17] Speech-driven Environmental Control Systems (SPECS). [Acccessed: 28 November 2016]. [Online]. Available: http://www.sheffield.ac.uk/cast/projects/specs

[18] Universal Access Project. [Accessed: 28 November 2016]. [Online]. Available: http://www.isle.illinois.edu/sst/research/ua/

[19] VIVOCA Project. [Acccessed: 28 November 2016]. [Online]. Available: http://www.sheffield.ac.uk/cast/projects/vivoca

[20] A. Hatzis, P. Green, J. Carmichael, S. Cunningham, R. Palmer, M. Parker, and P. O'Neill, "An integrated toolkit deploying speech technology for computer based speech training with application to dysarthric speakers." in INTERSPEECH, 2003.

[21] CMU Sphinx - Open source speech recognition toolkit. [Accessed: 28 November 2016]. [Online]. Available: http://cmusphinx.sourceforge.net

[22] A. B. Cavalcante and L. Lorens, "Use case: a mobile speech assistant for people with speech disorders," in Proceedings of the 7th Language & Technology Conference, November 27-29, 2015, Poznan, Poland, 2015, pp. 192–197.

[23] Google Cloud Speech API. [Accessed: 29 November 2016]. [Online]. Available: https://cloud.google.com/speech/