Mobile and Personal Speech Assistant for the Recognition of Disordered Speech

Agnieszka Bętkowska Cavalcante and Monika Grajzer
Gido Labs sp. z o.o.
Email: a.b.cavalcante@gidolabs.eu, 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. Moreover, in the systems targeted for disordered speech, the recognition performance is highly diminished by the environmental factors related to the disease. This limits the practical applicability of these solutions. To overcome this problem, we propose a Mobile and Personal Speech Assistant (mPASS) – a platform providing the users with a set of tools, which enable to intuitively create their own speech recognition system corresponding to their needs and capabilities. Our long term vision is that handicapped users, without computer science and artificial intelligence knowledge, will use this platform to design at home their own speech recognition system, tailored to the domain, vocabulary, and language they find most useful. As a result, a personalized speech recognizer will be created, which can be used with diversified speech-based applications.**

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

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]–[3]. Hence, a significant group of people is not able to use many voice-controlled state-of-the-art technology advances, which could support independence in handling their daily activities.

It is estimated that 1.3% of the population encounters significant difficulties in speech-based communication [4]. 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 [3]. 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 [3][5][6]. There have been several attempts to design a speaker-dependent dysarthric speech recognition systems [1]–[9], but they were trained mainly in the laboratory environments. Only a few of them were created and tested in real usage scenarios [2][4] with the limited achieved performance, which was not sufficient for the practical implementation [4].

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.

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 handicapped users. By maximizing their comfort, more speech samples can be collected and, at the same time, users' motivation to work with the system is improved. Moreover, 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.

This paper is organized as follows: Section II provides a brief overview of related work, while Section III depicts design challenges that are driven by the analysis of previous approaches. Sections IV and V present the mPASS solution and its architecture. The preliminary results are discussed in Section VI and Section VII concludes the paper.

II. RELATED WORK

In the recent years, an increased attention has been put towards the design of disordered, in particular dysarthric,

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-480-0

speech recognition systems (dysarthria is the key group of speech disorders) [1]–[9]. The investigated related works 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) [1][5][6]. 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, they 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 [4][5]. Substantially, the diminished performance did not allow for practical exploitation, as concluded from the the year-long project VIVOCA [4].

III. 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 [5]. These factors have significant impact on the design of an ASR system as a whole.

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) 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, often handicapped, people, 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 as well as an important success measure.

IV. 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. The 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.

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 not available to the disordered speakers thus far. 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, exploited in converters transforming (possibly disordered) speech to text or to a synthesized speech, and many more. Some examples of such end-point applications, currently being developed by us to showcase the capabilities of the mPASS technology, are:

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 aims to accomplish. They also constitute the differences between our approach and the related works.

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.

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

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-480-0

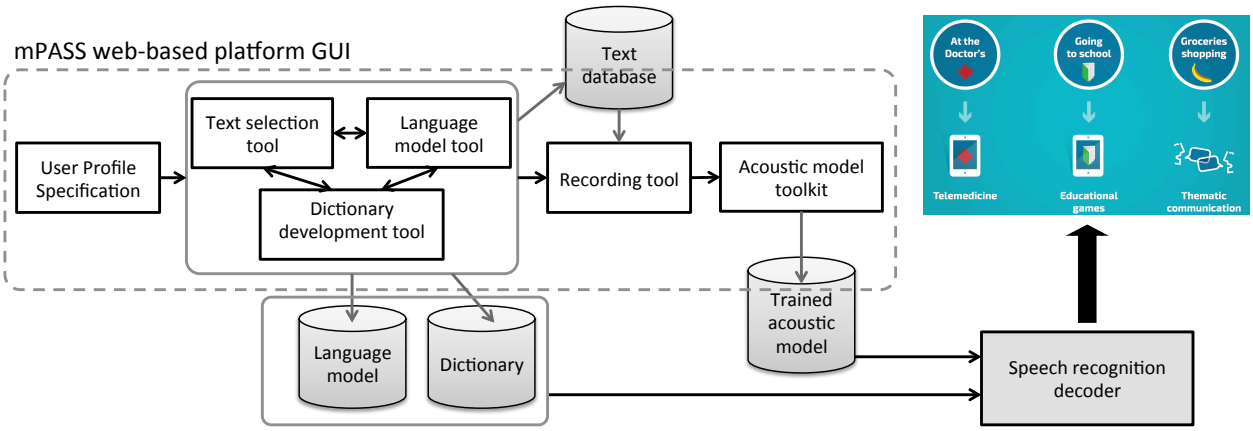

Figure 1. Mobile Personal Speech Assistant architecture – an overview.

can be adjusted to the user's health condition, motivation and other factors. Such an approach also minimizes the problem of reduced performance in case of systems which were trained in the silence conditions, but are used in the environment with existing background noise.

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.

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.

V. 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 the 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. 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.
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.
   a) **Text selection tool**: it is equipped with several phonetically balanced and phonetically rich texts for polish language. They have been created based on a well-known poems and short stories for children in order to make them easy to pronounce by the disordered-speech users. It is also possible to create the text automatically based on the existing dictionary and language model.
   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.
   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, initial examples are provided automatically). 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.
3) **Recordings**: The user records selected texts and/or word lists. There is a minimal suggested number of recordings specified. In addition, the system also 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

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-480-0

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 – the user will be given a real-time feedback information.

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.
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.

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.

From the user perspective, the recording tool functionality is the most important part of the mPASS platform. It is, however, also 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 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.

VI. Preliminary results

Initial performance trials were 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 [10] to a recipient from a phone contact list. User-defined voice commands are used to control the application. The user recorded 8 messages of his own choice (e.g., “I will be back in 1 hour”) and several action commands (“up”, “down”, “OK”, etc.) - all together 21 phrases, 30 times each. The ASR was using the HMM-based acoustic model. The recognition performance obtained in the laboratory environment was approx. 99%, whereas in a real environment (home/office) on average 84%. 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, including the time lost for necessary repetitions when recognition errors occurred, with the time needed for the same action to be completed by using the regular touch input (the person controls mobile phone installed on a wheelchair with his chin). The results were averaged over 20 trials – they presented 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 (Table I). Substantially, the user assessed a voice-controlled mobile speech assistant as the preferred option, which is the most important success measure.

TABLE I. Comparison of the time required to complete an action with a voice-controlled and manual entry

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

The initial trial presented above constitutes a first proof-of-concept evaluation. At this stage, the obtained performance results cannot be directly compared to the ones presented in the related works, since they were gathered for different usage scenario and with different ASR system, especially with regard to the selected vocabulary. However, in general, the system performance reached very high levels for the laboratory environments, often higher than those reported in related works. They were accompanied with a very promising outcome for the real usage environments, which was rarely achieved before. More detailed performance evaluation of the ASR systems created with the mPASS platform, based on the database of recordings collected from another 7-10 users, is a part of the future work.

VII. Conclusion and Future Work

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. Our findings revealed that the voice-controlled input was perceived as up to 49% better than 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.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-480-0

REFERENCES

[1] 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.

[2] M. S. Hawley et al., "A speech-controlled environmental control system for people with severe dysarthria," Medical Engineering & Physics, vol. 29, no. 5, 2007, pp. 586–593.

[3] 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.

[4] M. S. Hawley et al., "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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-480-0