Self-taught assistive vocal interfaces: An overview of the ALADIN project

*Jort F. Gemmeke[1], Bart Ons[1], Netsanet Tessema[1], Hugo Van hamme[1], Janneke van de Loo[2], Guy De Pauw[2], Walter Daelemans[2], Jonathan Huyghe[3], Jan Derboven[3], Lode Vuegen[1,4], Bert Van Den Broeck[1,4], Peter Karsmakers[1,4], Bart Vanrumste[1,4,5] **

[1]ESAT, KU Leuven, Leuven, Belgium
[2]CLiPS - Computational Linguistics Group, University of Antwerp, Antwerp, Belgium
[3]CUO | Social Spaces, iMinds, KU Leuven, Leuven, Belgium
[4]MOBILAB, Thomas More Kempen, Geel, Belgium
[5]iMinds, Future Health Department, KU Leuven, Leuven, Belgium
jort.gemmeke@esat.kuleuven.be

Abstract

This paper gives an overview of research within the ALADIN project, which aims to develop an assistive vocal interface for people with a physical impairment. In contrast to existing approaches, the vocal interface is trained by the end-user himself, which means it can be used with any vocabulary and grammar, and that it is maximally adapted to the — possibly dysarthric — speech of the user. This paper describes the overall learning framework, the user-centred design and evaluation aspects, database collection and approaches taken to combat problems such as noise and erroneous input.

Index Terms: vocal user interface, user-centred design, self-taught learning, speech database, dysarthric speech

1. Introduction

These days, Automatic Speech Recognition (ASR) is firmly rooted in everyday life, with ample examples such as talking to your iPhone using Siri, vocal interfaces for home automation or directing your navigation device by voice while driving. Still, voice control of the technology that we use in our daily lives is perceived as a luxury and more common interactions and input methods are often considered more suitable. A remote control, for instance, can be better suited for home automation because often, it is easier to push a button than to say a command.

Physically impaired people with restricted (upper) limb motor control, however, are permanently in the situation where voice control could significantly simplify some of the tasks they want to perform [1]. By regaining the ability to control devices in the living environment, voice control could contribute to their independence of living and their quality of life. Unfortunately, even state-of-the-art speech recognition systems are restricted in their vocabulary and grammar (mostly application determined), and offer little, if any, robustness to dialectic or dysarthric speech often encountered with disabled users.

While a solution to this lack of robustness might be to collect training material from a specific user and build custom acoustic and language models, in practice such effort is too great to make this a financially viable solution. The alternative currently used, adaptation of existing acoustic models, is limited to only very mild speech pathologies [2, 3, 4, 5, 6]. Moreover, the user's voice may change over time due to progressive speech impairments.

* Authors are ordered by institute.

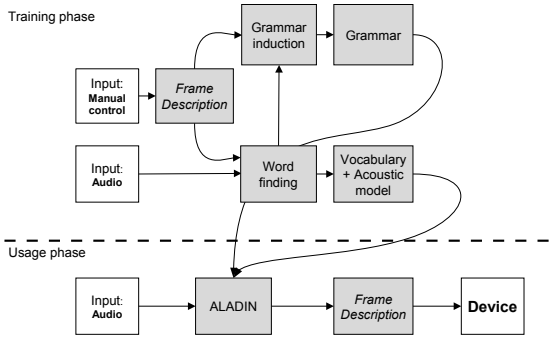

Figure 1: Overview of the ALADIN vocal interface framework. The white boxes indicate events or systems outside the ALADIN framework. The top panel shows the training phase, and the bottom panel indicates the usage phase.

In contrast, the ALADIN project aims to build a Vocal User Interface (VUI) that is trained by the end-user himself, which means it can be used with any vocabulary and grammar, and that it is maximally adapted to the — possibly dysarthric — speech of the user. This approach, while attractive, poses several challenges such as the lack of detailed annotation, since only the associated actions (such as a button push) are known, but not the words that are used to express the command. Other challenges include having to learn both acoustics and grammar from a (very) small number of examples, robustness against environmental noise or erroneous input from the end-users.

In this paper we give an overview of the ALADIN project [28]. Section 2 describes the overall learning framework, followed by a description of the user-centred design in Section 3. In Section 4 we outline the collection of several databases collected for evaluation, while in Sections 5 and 6 we describe the acoustic and language model induction from the user's speech, respectively. In Section 7 we discuss how we plan on valorising the developed technology and we end with our conclusions and plans for future work in Section 8.

2. Overall framework

A schematic overview of the ALADIN vocal interface framework is shown in Fig. 1. For now, we distinguish two phases: a training phase and a usage phase. In the training phase, a user is assumed to be able to still use a manual control, possibly with

some effort or through the help of a caretaker. A command is learned by giving the desired vocal command ("Input: Audio"), followed by demonstrating the action with the manual control ("Input: Manual control"). For example, the user could give the vocal command "Turn on the television" together with pressing the standby button on the television remote control. The manual control is represented using a *semantic frame* ("Frame Description") [7], a data structure that represents the semantic concepts that are relevant to the execution of the action and which end-users are likely to refer to in their commands.

The vocal command is processed into low level and intermediate level acoustic representations processed in the module "Word finding". The Word finding module, described in Section 5, uses the semantic frame descriptions together with the vocal commands to find recurring acoustic units such as words, that constitute the user's commands. Using these audio patterns, again together with the frame descriptions describing the user's actions, the module "Grammar induction" described in Section 6, attempts to build a grammar that describes the relation between parts of complex commands. For example, the learned vocabulary could become "Turn on" and "the television" and the learned grammar could be <turn on> <device>.

Using the knowledge acquired in the training phase, depicted by the boxes "Grammar" and "Vocabulary + Acoustic model", in the usage phase the most likely command is induced from a vocal command ("Input: Audio"). This recognized command is expressed once again as a semantic frame ("Frame Description") which is sent to the target device ("Device"), for example a television.

Although we distinguished a training and usage phase in this description, in fact the training phase is more or less optional since the training of the system does not stop after the training phase: In the usage phase, the system keeps updating itself when a new training item is provided by the user. Any use of a manual control together with a voice command will be treated as a training item. Likewise, the use of a manual control shortly after executing a spoken command, will be regarded as a correction, causing the system to undo its last command and treat the spoken command as a training item.

3. User-centred design

Both VUI design (as a non-standard interface) and designing for users with disabilities require a thorough User-Centred Design (UCD) approach [8, 9]. In this approach, the end-users are the central focus in the design and development of new products or applications. This way, the match between the products or applications under development on the one hand and the user needs on the other hand can be optimized from an early stage onwards. In the ALADIN project, the problems, needs, and tasks faced by the end users in their daily lives are queried using various techniques. To establish the user's needs, users and their caregivers were invited to talk about their daily life in interview sessions, focusing explicitly on activities they regard as being important, difficult or even dangerous. Additionally, participants were asked to guide us through their homes while talking about and performing their daily tasks, e.g. explaining the tools they use, difficulties they encounter, etc. This research showed that voice controlled home automation systems were most in demand. Perhaps unexpected, there was also a large demand for voice control of leisure activities such as playing games.

To further guide the design process, we created a set of personas. Personas can be described as archetypes based on real users. Though fictional, they represent real people, summarizing the detailed richness from user research activities (contextual inquiry, interviews,...) into short archetype descriptions that can be referred to in the remainder of the development process. They are a tool that allows developers and researchers to avoid designing technology for featureless users, but instead allows them to tailor it to distinct user types. Over the first two years of the ALADIN project, these persona descriptions have been updated based on a deepened understanding of the end user population on the one hand, and on an alignment with the technological project focus on the other hand.

To get an understanding of how people address voice-controlled technology, test sessions were held with scenario visualizations in the form of storyboards. The storyboards, as a translation of a conceptual model into a narrative, were used to ask the respondents to formulate relevant voice commands. Participants were told not to worry about system limitations, and the visualisations made sure respondents were not biased towards specific words or sentence structures while formulating their commands. During these sessions, significant variation was found, for instance concerning the addressee of user's commands: some respondents addressed individual devices directly, without addressing the voice-control system ("door, open"), while other respondents addressed the voice-controlled system as a whole, telling it to act on the environment and control other devices ("[SYSTEM], open the door"). For this last group, addressing separate objects such as doors felt very unnatural.

Additionally, respondents used a variety of styles in addressing the voice-control system, ranging from a purely 'technical', command-style interaction to a more anthropomorphized, personal communication with the system [10]. The latter interaction style focuses on the system as a conversation partner, addressing the system in a human or natural way. The former, 'command style' interaction implies a focus on efficiency and reliability characteristics that are very important for this specific target group. Based on these observations, as well as technical considerations, it was decided that to use the system, users will have to address it directly by calling its name (*ALADIN*, or any other name), saying, for instance, "ALADIN, open the door" instead of "door, open up". Addressing the system with a proper name harmonizes and steers user interactions and, from a technical point of view, it has the advantage that the name can be used as a cue for the system to be activated, and actively 'listen' to the command that is being given.

In a second iteration of this research, an interactive 3D-environment was used to allow users to interact with a voice-controlled system in a more immersive manner, using the wizard-of-oz method. This method allows users to use the system as if it is already functional, while in reality a researcher simulates the voice recognition system by operating the various appliances represented in the virtual environment, based on the speech commands uttered by participants. Even though the outcomes of these speech commands were not visible in a real life home automation system, participants commented on this prototype as being highly realistic. This more realistic usage data provided information on the structure of typical utterances, implicit user expectations, and several points of attention for further development of the ALADIN system (incomplete commands, effects of pronunciation difficulties, etc.). Specific challenges include non-specific user commands, for which the system needs additional information about location or context to identify the user's intentions. For instance, "lights on" could be expected to turn on the lights in the room where the user is, without the lamp being verbally specified in the command. Since this requires technology solutions out of the scope of AL-

ADIN, this has implications for the managing of user's expectations.

In the remainder of the ALADIN project, the user-centred design research will continue involving end users in the technology development. As the technical development progresses, the UCD focus will be on evaluating the user's experiences with semi-functional prototypes and project demonstrators.

4. Data collection

Existing databases for ALADIN targets such as home automation and dysarthric speech [11, 12, 13] typically do not offer a sufficient number of repetitions of commands to evaluate the effectiveness of self-taught learning. We have therefore collected three databases: Two of these pertained to home automation systems, the application most desired by the user base, while a third dataset consisted of users playing a card game.

4.1. Home automation

Of the two home automation datasets, the first consists of speakers with non-pathological voices and the second dataset consists partly of speakers with pathological voices. For all speakers, speech intelligibility measurements were taken and automatically processed using the method described in [14].

For the first home automation dataset, non-pathological speech commands were recorded in a realistic setting, i.e. a fully automated room using a wizard-of-oz device control. The commands were prompted using visual cues (a video) on a computer screen. In order to simulate situations with environmental noise, recordings were also made with a concurrent sound source. In addition to a close-talk microphone, multichannel audio recordings were made with multiple microphone-arrays, placed near the user, on walls and near the optional noise source.

The dataset consists of 27 test subjects of which 20 are of the targeted user group. Each person was asked to go repeatedly through a list of 33 different actions, until a recording time of 30 minutes was reached, yielding a dataset of 1888 commands for the target group. In addition to this set, longer recording sessions with 7 non-target users were carried out, yielding 1699 spoken commands.

The second dataset focused on dysarthric speech. However, since collecting a large number of realistic, spontaneous spoken commands is difficult due to the targeted users getting tired quickly, a two-phase data collection method was used. In the first phase 9 users were asked to control 31 different appliances in a 3D environment (c.f Section 3), guided by a scenario in order to ensure an unbiased choice of words or grammar. In the second phase these command lists were read back repeatedly by 21 test users. Of these 21, 8 persons were MS patients, which are known to have an enlarged risk for degeneration of their voice. Future data collection from these patients enable validation of the ongoing user adaptivity of ALADIN.

4.2. Card game

The third dataset was recorded while the test subjects were playing a card game called "patience" using spoken commands. The data was collected from non-target (unimpaired) subjects with non-pathological speech. The user's were free to choose their vocabulary and grammar, although in practice the vocabulary was limited indirectly by the number of cards, card positions and functionality. Utterances in this data set are typically grammatically more complex compared to those in the home automation datasets.

We obtained data for eight participants, each playing two sessions of about 30 minutes (a few weeks apart). To enable evaluation with larger amounts of training material, additionally 210 minutes of speech spanning 7 playing sessions was collected for one participant. All data was manually transcribed, both in canonical form as well as using semantic frame descriptions. Additionally, for each spoken command the entire game-state was recorded to enable evaluation in the context of the possible card moves according to the rules of the game.

5. Vocabulary acquisition

Conventional VUIs use speech recognizers trained on carefully annotated speech, together with pronunciation dictionaries, to train the acoustic models. In the ALADIN framework, however, such supervisory information is not available: we do not know in advance which words the user will utter for a command, nor what their pronunciation might be. The supervision is therefore considered *weak*: we only know the associated action. Still, by mining the acoustic commonalities and differences with other spoken commands, which may partly share semantics (for example "TV on" and "TV off"), we can find recurring, meaningful acoustic patterns, such as words.

In ALADIN, research has focused on a vocabulary acquisition framework based on non-negative matrix factorization (NMF) [15]. In short, NMF relies on modelling utterances as a composition of the words (or rather, recurring acoustic units) that make up an utterance, and works by decomposing the utterances of the training data into low-rank representations modelling the acoustics of the recurring patterns, and the activations of these patterns in each utterance [16]. This process is guided by the utterance-level labelling derived from the semantic frame description [17].

One of the goals has been to improve the overall accuracy of vocabulary acquisition. At the lowest level, the conversion of spoken commands to low-level spectrographic representations, improvements were obtained through the use of modern, discriminative feature representations [18]. At the intermediate level, various forms of utterance-based representations were proposed to better capture the temporal and spectral details of a sentence in a single representation [19, 20]. At the highest level, methods for acquiring not only recurring acoustic patterns, but also the exact order in which they occur, were developed [21].

Whereas the earlier investigations mostly focused on the accuracy of vocabulary acquisition given large amounts of training samples [16], in the ALADIN project it is imperative that vocabulary acquisition is also *fast*, i.e., enabling learning from only a few examples. This aspect was investigated in [22], where methods were proposed to train acoustic models that scale with the available amount of training data to speed up learning. Currently research focusses on algorithmic approaches to prevent over-fitting, as well as exploiting the use of out-of-domain knowledge (such as a pre-trained phone recognizer from a different language) to speed up learning. Recently, investigations using the pathological home automation database (c.f. Section 4.1) have been initiated and preliminary results seem encouraging.

Another aspect that influences the speed of vocabulary acquisition is the robustness of NMF with respect to uncertain or incorrect semantic frame descriptions. Such incorrect input can occur for example if the user erroneously presses a wrong button or omits a device description from his vocal command. It was shown in [22] that the NMF framework is in fact inherently robust against such errors, which greatly improves the practical

usability of the envisioned technique. This was corroborated by the research in [17], where vocabulary acquisition on the patience dataset (c.f. Section 4.2) was shown to be robust against *ambiguous* semantic frame descriptions.

The subject of environmental noise robustness has not yet been evaluated within the context of the ALADIN project, but will rely on multi-microphone techniques [23]. Additionally, front-end single-microphone denoising scenarios will be considered through the use of novel exemplar-based techniques [24], which are expected to match well with the speaker-dependent setting of ALADIN.

6. Grammar induction

While the recognition of simple commands can be treated as a multi-class classification problem, this quickly leads to data scarcity for more complex commands, for example when the compositionality of the utterance and the order of the words in the sentence differentiate its meaning (e.g., in a card game, "put the four of hearts on the five of clubs" vs "put the five of clubs on the four of hearts"). Just as recurring acoustic patterns can be extracted from the weakly supervised spoken commands, *grammar induction* can automatically find structure by mining the commonalities and differences with other commands.

To accomplish the grammatical mapping between a command and the semantic frame that triggers the correct control, we opted for a *shallow grammar* approach which does not attempt to induce a full grammatical structure for a sentence, but rather defines the problem as a sequence tagging task, in which each word of a command is associated with a *concept tag*, i.e. a label that represents to which semantic frame slot (if any) the word needs to be mapped. Accurate concept tagging then effectively establishes a mapping between the words of the command and the frame description needed to trigger the intended control. To study the learnability of this task, we opted to first study the problem as a supervised classification task, trained on manually transcribed and annotated data.

We use the patience dataset (c.f. Section 4.2), which contains non-trivial sentences and covers the highest complexity that the ALADIN system is expected to run into. We performed experiments using an off-the-shelf data-driven tagger, typically used to perform morphosyntactic part-of-speech tagging [25]. Training on one portion of the data set and evaluating on held-out data allowed us to study the accuracy with which an automatically induced concept tagger can be expected to process previously unseen sentences. The experiments yielded encouraging results, with accuracy scores of more than 95% on the phrase level [17, 26, 27]. While these experiments were performed on reference transcriptions and the results are not necessarily indicative of the accuracy of the final system, the impact of these preliminary experiments is nevertheless significant: current results indicate that we are able to restrict the task of unsupervised grammar induction to the task of discovering the concept tags that map the words in an utterance to the frame slots of the intended control.

Both the vocabulary acquisition experiments, as well as those with concept tagging, showed a need for a tighter interaction between these two aspects of spoken command learning: the induction of (pseudo-)word units, as well as the mapping of these units onto semantic frame slots, should be done in one processing step, as neither of them can be considered a proper precursor to the other. We therefore decided to investigate whether the automatic finding of recurring acoustic patterns and the induction of the mapping of these patterns onto semantic frame slots can be performed simultaneously. This has the significant additional advantage that grammar induction can then be performed directly on the basis of acoustic patterns, rather than having to assume an intermediate layer of representation.

Current, as of yet unpublished, research establishes this combined approach by using a discrete Hidden Markov Model (HMM) approach, in which slot-values of frames are considered to be the hidden states which generate the observations (occurrence of recurring acoustic patterns). The learning of the ergodic HMM consists of learning the slot-value to observation mapping (vocabulary acquisition), as well as the transitional relation between states (slot-values) over time (the grammar induction). A first set of proof-of-the-principle experiments yielded more than encouraging results. For example, a closer inspection of the learned models revealed that the HMM-model correctly learns the typical word order of patience commands, such as for example the order <from> <to> and <suit> <value>, even though such order information was never made explicit in the supervision.

7. Valorisation

The ALADIN project does not just aim at fundamental research, but also has the clear goal of valorisation: transferring the technology developed in ALADIN to industry during and after the project. The project partners are building relevant *demonstrators* that should allow parties involved with bringing technology to the end-user, such as the application builders and the governmental bodies that legislate reimbursement of assistive devices, to judge the quality of the technology as well as the effort involved in deploying it. They are also meant to start the creative process of product design by giving concrete examples of what is possible and what the limitations are.

The demonstrator currently under development is a virtual assistant that allows access to multiple tasks such as the patience card game also used for the research described in Sections 5 and 6, as well as home automation control. The home automation demonstrator will not only control the 3D environment used for user evaluation and data collection (c.f. Sections 3 and 4.1), for portability, but also enable interaction with existing interfaces such as infrared.

8. Conclusions and future work

In this paper we presented an overview of the ALADIN project, which aims to develop an assistive vocal interface for people with a physical impairment. In contrast to existing approaches, the vocal interface is trained by the end-user himself, which means it can be used with any vocabulary and grammar, and that it is maximally adapted to the — possibly dysarthric — speech of the user.

While the project is ambitious, early results are encouraging and success would mean the possibility of bringing a vocal interface to a group of users that was previously unable to use speech technology in a satisfactory manner. In the remainder of the project, we plan on collecting more pathological speech data, evaluate and improve recognition of dysarthric speech, investigate the impact of environmental noise and build a set of demonstrators to showcase the technology.

9. Acknowledgements

This research is part of the ALADIN project [28] and is funded by IWT-SBO grant 100049.

10. References

[1] J. Noyes and C. Frankish, "Speech recognition technology for individuals with disabilities," *Augmentative and Alternative Communication*, vol. 8, no. 4, pp. 297–303, 1992.

[2] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proc Interspeech 2012*, Portland, Oregon, US, Sep 2012.

[3] K. T. Mengistu and F. Rudzicz, "Comparing humans and automatic speech recognition systems in recognizing dysarthric speech," in *Proceedings of the Canadian Conference on Artificial Intelligence*, 2011.

[4] H. V. Sharma and M. Hasegawa-Johnson, "State transition interpolation and map adaptation for hmm-based dysarthric speech recognition," in *HLT/NAACL Workshop on Speech and Language Processing for Assistive Technology (SLPAT)*, 2010, pp. 72–79.

[5] F. Rudzicz, "Acoustic transformations to improve the intelligibility of dysarthric speech," in *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT2011)*, 2011.

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

[7] Y. Wang and A. Acero, "Rapid development of spoken language understanding grammars," *Speech Communication*, vol. 48, no. 3-4, pp. 390–416, 2006.

[8] D. J. Mayhew, *The Usability Engineering Lifecycle: A Practitioners Handbook for User Interface Design*. San Francisco, CA: Morgan Kaufmann, 1999.

[9] H. Sharp, Y. Rogers, and J. Preece, *Interaction design: beyond human-computer interaction*, 3rd ed. Chichester, UK: Wiley, 2007.

[10] M. Verstraete, J. Derboven, J. F. Gemmeke, P. Karsmakers, B. Van Den Broeck, and H. Van hamme, "The design of voice controlled assistive technology for people with physical disabilities," in *Language Technology in Pervasive Computing (LTPC)*, 2012.

[11] X. Menéndez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Proceedings of International Conference on Spoken Language Processing*, 1996.

[12] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proceedings of Interspeech*, 2008, pp. 22–26.

[13] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," in *Language Resources and Evaluation*, 2011, pp. 1–19.

[14] C. Middag, "Automatic analysis of pathological speech," Ph.D. dissertation, Ghent University, Belgium, 2012.

[15] D. Lee and H. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[16] H. Van hamme, "Hac-models: a novel approach to continuous speech recognition," in *Proceedings INTERSPEECH*, 2008, pp. 2554–2557.

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

[18] J. Driesen, J. F. Gemmeke, and H. Van hamme, "Data-driven speech representations for nmf-based word learning," in *Proc. SAPA*, Portland, OR, USA, September 2012.

[19] J. Driesen and H. Van hamme, "Fast word acquisition in an NMF-based learning framework," in *Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 2012.

[20] J. Driesen, J. Gemmeke, and H. Van hamme, "Weakly supervised keyword learning using sparse representations of speech," in *Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 2012.

[21] H. Van hamme, "An on-line NMF model for temporal pattern learning. Theory with Application to Automatic Speech Recognition," in *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel Aviv, Israel, March 2012.

[22] B. Ons, J. F. Gemmeke, and H. Van hamme, "Label noise robustness and learning speed in a self-learning vocal user interface," in *IWSDS*, 2012.

[23] B. Van Den Broeck, A. Bertrand, P. Karsmakers, H. Van hamme, B. Vanrumste, and M. Moonen, "Time-domain generalized cross correlation phase transform sound source localization for small microphone arrays," in *proc. EDERC 2012*, 2012, pp. 76–80.

[24] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech and Language processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[25] W. Daelemans, J. Zavrel, A. van den Bosch, and K. Van der Sloot, "MBT: Memory-based tagger, version 3.2, reference guide," University of Tilburg, Tech. Rep. 10-04, 2010.

[26] J. van de Loo, G. De Pauw, J. Gemmeke, P. Karsmakers, B. van den Broeck, W. Daelemans, and H. Van hamme, "Towards shallow grammar induction for an adaptive assistive vocal interface: a concept tagging approach," in *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*. Istanbul, Turkey: European Language Resources Association (ELRA), 2012, pp. 27–34.

[27] J. van de Loo, J. Gemmeke, G. De Pauw, J. Driesen, H. Van hamme, and W. Daelemans, "Towards a self-learning assistive vocal interface: Vocabulary and grammar learning," in *Proceedings of the First Workshop on Speech and Multimodal Interaction in Assistive Environments (SMIAE)*. Jeju, Republic of Korea: Association for Computational Linguistics, 2012, pp. 34–42.

[28] ALADIN, "Adaptation and Learning for Assistive Domestic Vocal INterfaces," Project Page: http://www.esat.kuleuven.be/psi/spraak/projects/ALADIN.