Adaptive Speech Recognition and Dialogue Management for Users with Speech Disorders

I. Casanueva, H. Christensen, T. Hain, P. Green

Department of Computer Science, University of Sheffield, United Kingdom
i.casanueva@sheffield.ac.uk, {h.christensen,t.hain,p.green}@dcs.shef.ac.uk

Abstract

Spoken control interfaces are very attractive to people with severe physical disabilities who often also have a type of speech disorder known as dysarthria. This condition is known to decrease the accuracy of automatic speech recognisers (ASRs) especially for users with moderate to severe dysathria. In this paper we investigate how applying probabilistic dialogue management (DM) techniques can improve interaction performance of an environmental control system for such users. The effect of having access to different amounts of adaptation data, as well as using different vocabulary size for speakers of different intelligibilities is investigated. We explore the effect of adapting the DM models as the ASR performance increases, such as is the case in systems where more adaptation data is collected through system use. Improvements compared to a non-probabilistic DM baseline are seen both in terms of dialogue length and success rate, 9% and 25% mean relative improvement respectively. Looking at just the more severe dysarthric speakers these numbers rise 25% and 75% mean relative improvement. These improvements are higher when the ASR data adaptation amount is small. Further results show that a DM trained on data from multiple speakers outperform a DM trained on data from a single speaker.

Index Terms: dysarthric speech, dialogue management, environmental control system

1. Introduction

Automatic speech recognisers (ASRs) have poor performance for dysarthric users, but often these users have physical disabilities that would make speech enabled environmental control (EC) very attractive. A previous study [1] has shown that maximum a posteriori (MAP) adaptation can significantly increase the accuracy of the ASR for mild dysarthric speakers, but the performance is still very low for speakers with lower intelligibility. During the last decade, probabilistic dialogue management (DM) for spoken dialogue systems has shown promising performance in terms of robustness when used in conjunction with systems with high word error rate. These techniques showed a higher relative improvement as the ASR performance decreases than other DM techniques [2]. Another advantage of using probabilistic DM is that its dialogue policy can be optimised regarding to a specific reward function using reinforcement learning [3], meaning that the dialogue manager will automatically adapt its behaviour to the ASR characteristics. Probabilistic DM was previously used with dysarthric speakers with promising results [4].

In [5] we presented an EC system (homeService) for dysarthric speakers, where the ASR acoustic models are updated as more adaptation data is collected. One of the key points of this system is the adaptation to a specific user, meaning that the performance of the system will improve as the user interacts more with it. Probabilistic DM naturally fits into this system configuration, since it will be able to adapt its dialogue policy both to the user specific characteristics, as well as to the ASR accuracy changing over time. A slightly different DM configuration has been implemented, where the dialogue models tracking the current state of the dialogue (user intention) change over time as the ASR accuracy changes. The main improvement we expect to obtain from this framework is to have a system which has a more "conservative" dialogue policy, meaning that it will ask more confirmation questions when the performance of the ASR is poorer, and is able to automatically adapt this policy to a more straightforward one when the ASR performance improves.

2. Dysarthric data

UASpeech [6] is by far the largest database of dysarthric speech suitable for training acoustic models for ASR. It includes speech from 15 speakers with a range of impairment levels with a total of about 18 hours of speech. Each recording is a single word, and the database includes 10 numbers, the 26 NATO alphabet letters, 19 command words, 100 common words, and 300 uncommon words.

All speakers have dysarthric speech with different severity. The database comes with some information about the speakers' characteristics as well as their intelligibility measure, which is clustered in 4 groups; *very low*, (2% to 15%, 4 speakers); *low*, (28% to 43%, 3 speakers); *mid*, (58% to 62%, 3 speakers) and *high*, (86% to 95%, 5 speakers).

3. Effect of amount of adaptation data and vocabulary size on ASR accuracy

Maximum a posteriori (MAP) adaptation [7] has been shown to be a successful way of establishing acoustic models when faced with limited amounts of data from a given speaker. In [1], accuracy results on the UASpeech task are presented, using about 40 mins of data for each speaker employing the whole 455 word vocabulary. Here we investigate the effect on accuracy of using less data for adaptation, as would be the case when initially setting up e.g. a new EC system. The effect reducing the vocabulary and hence the decoder confusability is also investigated.

The ASR accuracy results are shown in fig. 1; each line shows the mean and standard deviation for each intelligibility group, as a function of the amount of data used for MAP adaptation. Especially in the 36 word case, the accuracy improvement converges after a certain amount of data. Reducing the vocabulary size has the effect of increasing the accuracy and decreasing the amount of data needed until this convergence point, which is a key point for a system like [5]. Collecting large amounts of enrolment data from a dysarthric user is infeasible but 36 com-

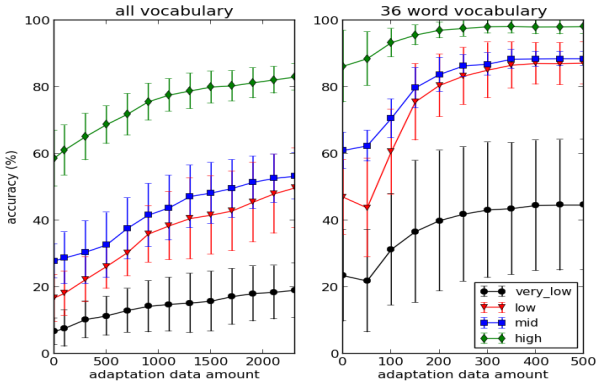

Figure 1: *Effect of increasing the amount of data for MAP adaptation in a 455 word and a 36 word vocabulary. X axis represents the amount of word recordings used to adapt the acoustic models*

mands would be enough to control a simple system. The relative accuracy increase is higher for very low intelligibility speakers, which is promising as it might permit them to use spoken control systems with better performance.

4. Dialogue Management background

In a spoken dialogue system, a DM is the module in charge of controlling the flow of the dialogue. It has two main functions: firstly, the DM tracks the *dialogue state* which maps the history (observations from the user and actions taken by the DM) into an internal state representing the user's goal [8]. Secondly, given the dialogue state, the DM will choose the (optimal) action in each turn to satisfy the user's goal, known as *policy optimization* [3]. Probabilistic dialogue management has proven to increase robustness against poor ASR performance and permits less constrained interactions (mixed initiative dialogues) [2]. A popular approach during the last decade has been casting the dialogue as a Partially Observable Markov Decision Process (POMDP) [9], where the DM works over a distribution over dialogue states (known as belief state) instead of a single state. This gives the system the advantage of having parallel dialogue state hypotheses and using local confidence scores to update the belief state. This enables the use of n-best recognitions instead of a single ASR output to infer the user's goal. After each user interaction, the belief state, $b(s)$, is updated as:

$$b(s_t) = P(s_t|\mathbf{H}) = \\ k \cdot P(\mathbf{o_t}|s_t, a_{t-1}) \sum_{s_{t-1}} P(s_t|s_{t-1}, a_{t-1}) b(s_{t-1}), \quad (1)$$

where $\mathbf{H} = (\mathbf{o_0}, \mathbf{o_1}, ...\mathbf{o_t}, a_0, a_1, ..., a_{t-1})$ is the dialogue history until turn t, with $\mathbf{o_t}$ representing an observation (typically a n-best list of recognised user acts) and a_t representing a system action. k is a normalization constant and $P(\mathbf{o_t}|s_t, a_{t-1})$ and $P(s_t|s_{t-1}, a_{t-1})$ are called the *observation model* and *transition model*, respectively.

Another advantage of POMDPs is that, given a local reward function $R(a_t, s_t)$ and using reinforcement learning algorithms, it is possible to optimize the dialogue policy $P(a_t|b_t)$ so it maximizes the total dialogue reward. The problem with this approach is that the POMDP solving algorithms scale poorly with the number of states and observations. Approximated methods exist for solving POMDPs [10], but they are also intractable for large scale dialogue systems.

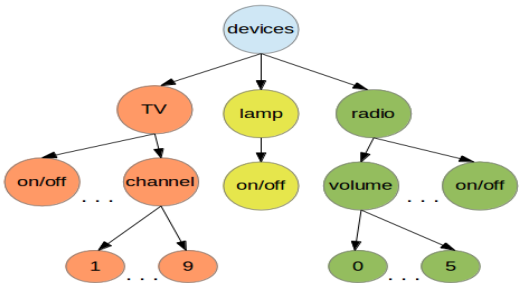

Figure 2: *Part of the hierarchical architecture of the environmental control system*

Research in the last couple of years has shown that discriminative approaches for state tracking methods [11] improve the generative model in eq. 1, and Temporal Difference Reinforcement Learning algorithms [12] do not need to define a dialogue model and can optimize the policy in fewer interactions.

5. An environmental control style experimental system

We are interested in testing the performance of different dialogue managers when interacting with a dysarthric speaker, in a spoken control system that adapts its ASR to improve performance over time. For this purpose a simulated dialogue environment was designed. A spoken control system architecture for assistive technology, similar to the one described in [5], was constructed using a total vocabulary of 36 commands. The system is designed to control several devices in a home (e.g. TV, lights...). The user can navigate through the system in a hierarchical fashion. For instance, if the user wants to change the TV channel to "five", the sequence of words to utter will be *"TV", "channel", "five"*. A set of speaker specific *simulated users* (SUs) have been built with the purpose of testing the system. In the experiments, each SU was tuned to a specific ASR accuracy, arising from a specific amount of adaptation data available.

5.1. Environmental control system architecture

The control architecture of the system is organised in a tree setup as shown in fig.2. In such a finite state automata, each node represents either a device (e.g. "TV"), a functionality of that device (e.g. "channel"), or actions that trigger some change in one of the devices (e.g. "1", child of "channel", will change the TV to the first channel). When the system transitions to one of these terminal nodes, the action associated with this node is performed, and subsequently the system returns to the root node. In the following experiments a dialogue will be considered finished when one of the actions was carried out. In the remaining nodes the user may either say one of the commands available in that node (defined by its children nodes) to transition to them, or say the meta-command "back" to return to its parent node.

5.2. Simulated user

For each speaker in UASpeech and for different amounts of adaptation data (same step-size as in section 3) a different SU has been built. In the following, $SU_{F02,50}$ is used to indicate the SU trained with data of speaker F02 and and using 50 words for the MAP adaptation. We use a similar approach as in [13] where the simulated user is factorised into two parts, the behaviour simulator, which models the behaviour of the user when interacting with the system, and the ASR simulator, which mod-

els the effect of the ASR channel in the actual word spoken by the user. For each interaction and given a user goal G, the observation $\mathbf{o}_t$ seen by the system is given by:

$$P(\mathbf{o}|G) = P(A_u|G)P(\mathbf{o}|A_u), \quad (2)$$

where A_u is the user command, $P(A_u|G)$ corresponds to the user behaviour, $P(\mathbf{o}|A_u)$ to the ASR channel and $\mathbf{o} = (\tilde{\mathbf{A}}_\mathbf{u}, \mathbf{c}) = (\tilde{A}_{u1}, \tilde{A}_{u2}, ..., \tilde{A}_{un}, c_1, c_2, ..., c_n)$ is an n-best list of noisy command recognitions with normalised confidence scores of length n. The behaviour model is independent of the ASR so it is common for all speakers and amounts.

To build the behaviour model, $P(A_u|G)$, a knowledge based approach has been followed, similar to the one presented in [14]. At the beginning of each simulated dialogue, a user goal G is randomly generated (e.g. [*TV,channel,five*]). The behaviour of the SU is then defined by an agenda which is updated each turn. The SU will say one command of the goal at each interaction, repeat his last command if asked so, and say the meta-command *"back"* if the system has transitioned to a wrong state [1].

A different ASR simulation model, $P(\mathbf{o}|A_u)$, is trained for each speaker and adaptation amount combination. The data used to train this models is obtained from the experiments presented in section 3. A generative model was trained with the following assumptions to reduce its complexity, due to the data sparsity. The ASR simulation model is approximated as:

$$P(\tilde{\mathbf{A}}_\mathbf{u}, \mathbf{c}|A_u) \simeq \prod_{i=1}^{n} P(\tilde{A}_{ui}|A_u) \prod_{i=1}^{n} P(c_i|i) \quad (3)$$

where $P(c_i|i) = P_{cor}(c_i|i)$ if $\tilde{A}_{ui} = A_u$ and $P_{inc}(c_i|i)$ otherwise. The generation of $\mathbf{o}$ must satisfy three constraints, $\tilde{A}_{ui} \neq A_{uj} \, \forall \, j \neq i$, $c_i \geq c_j \, \forall \, i > j$ and $\sum_{i=1}^{n} c_i \leq 1$.

5.3. Baseline dialogue manager

To control the flow of the dialogue, a local confidence score based dialogue manager was designed. This dialogue manager will perform the command with higher confidence score of the observed n-best list if its confidence score is higher than a threshold, else it will ask to repeat the command. This threshold is optimized for each speaker and each amount of ASR adaptation data. If the recognised command in the top of the n-best list is not one of the available commands in the node the EC system is in, it will ask to repeat it again.

5.4. Probabilistic Dialogue Manager

One of the main concerns when interacting with dysarthric speakers is the performance of the ASR, which is usually poorer than the average, with high variation between speakers. Probabilistic DM has shown to improve dialogue interaction robustness when dealing with high error rate ASR systems, so it naturally fits into interaction with dysarthric speakers. It is also a data driven method, meaning that the models can be learnt with user specific data to deal with the variability between speakers.

To overcome the problem of scalability (and data sparsity) when training a full POMDP model for the dialogue task, the dialogue management is done in a hierarchical way. Following the same tree configuration as illustrated in fig. 2, a POMDP is defined for each non-terminal node. The states of each of these POMDPs will be the available commands (children) in the node, plus a state for the meta-command "back". The corresponding actions are: the action associated with each child node, a "back" action, which performs transition to those nodes and to the parent node respectively; and an "ask" action to ask the user to repeat his last command. The local reward function is defined as -1 for the ask action and for the rest of actions a reward of +10 is given if the action is the user goal and 0 otherwise.

The advantage of this DM design, is that is enables the study the robustness of the DM against ASR errors. This way, the flexibility given by probabilistic DM is sacrificed to make it computationally tractable and less data to train the dialogue models is needed. POMDPs also factorize the models into an observation model and a transition model. As it is assumed that the user won't change his goal during a single dialogue and will collaborate with the system, the transition model is trivial and can be hand-crafted. Only the observation model needs to be trained from data. Deriving from eq. 1, and assuming that the user doesnt change his goal, $P(s_t|s_{t-1}, a_{t-1})$ is equal to 1 if $s_t = s_{t-1}$ and 0 otherwise. Assuming that the observation is independent of the previous system action and commands in the n-best list are independent between them, $P(\mathbf{o_t}|s_t, a_{t-1})$ can be approximated as a sum of the probability of each noisy recognition of user command given the state (actual user command), $P(\tilde{A}_u|s)$, weighted by the confidence score of that command in the observed n-best list $c(\tilde{A}_u)$. The belief state can now be written as:

$$b(s_t) = k \cdot b(s_{t-1}) \sum_{\tilde{A}_u} P(\tilde{A}_u|s) c(\tilde{A}_u) \quad (4)$$

As the transition model, the observation model and the reward function are defined, the optimal policy can be found by point-based value iteration reinforcement learning methods [10]. To solve the POMDPs the zmdp toolkit [15] was used.

5.4.1. Probabilistic Dialogue Manager variations

In the following we analyse the effects of training the DM using data from just a single user or from multiple users, and analysing if the dialogue model should change as the performance of the ASR changes. To test this, 4 variants of DM were trained; two speaker dependant (SD), trained on the tested speaker data, and two speaker independent (SI), trained in all speakers data. Of these four, two are trained only on data from the user interacting with a ASR with no MAP adaptation (0A), and two vary their dialogue models as the data used for ASR MAP adaptation increases (IA). This case could be seen as a set of dialogue managers, each trained on data originating from an ASR adapted with a specific amount of data. Four probabilistic DM have been tested, *SD 0A, SD IA, SI 0A* and *SI IA*. The data used to train these models is independent from the data used to train the SU.

6. Results

To compare the performance of the different DMs, a set of experiments were carried out using the dialogue environment described in section 5. For each speaker in UASpeech, the dialogue managers were tested interacting with simulated users of that speaker for each amount of adaptation data. 5000 dialogues were simulated for each combination of speaker and amount and two metrics were measured; the dialogue success rate, whereby a dialogue is considered successful if the terminal action carried out by the system matches the user goal, and the dialogue reward, which is -1 for each interaction in the dialogue +20 if the dialogue is successful.

[1] It is assumed that the user knows how the control system works, (i.e. knows which command to utter in each turn to fulfil his goal) and is fully collaborative with the system (e.g. if the system asks to repeat his last command, the user will repeat it, and the user doesnt change his goal during a single dialogue)

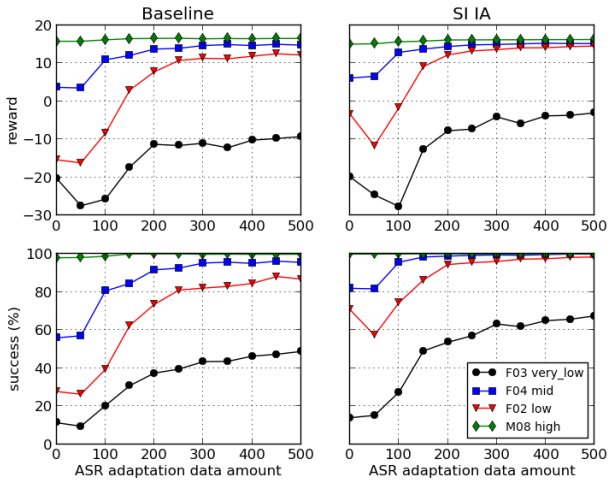

Figure 3: *Reward and success rate for local confidence score based DM (left) and "SI IA" DM (right) for four speakers with different intelligibility. X axis represents the amount of word recordings used to adapt the acoustic models*

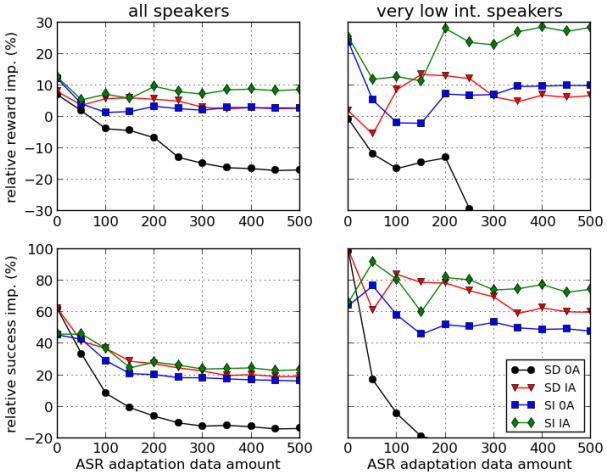

Figure 4: *Mean relative reward and success improvement rate respect to baseline for all speakers (left) and very low intelligibility speakers (right) with different DM. X axis represents the amount of word recordings used to adapt the acoustic models*

Fig. 3 (left) shows the dialogue metrics for the baseline DM (section 5.3), tested with a simulated user that changes its ASR simulation model to simulate MAP adaptation for different amounts of data (X axis represents the amount of data). Four UASpeech speakers are plotted, as examples of the average speaker for each of the intelligibility levels in UASpeech. It is seen that with a 36 word vocabulary size EC system, high intelligibility speakers have an almost perfect performance. Mid and low speakers reach a good performance when some adaptation data has been used in the ASR, but for very low intelligibility speakers the performance is poor, especially when the amount of data for ASR adaptation is small.

Fig. 4 (left) shows the relative improvements with respect to the baseline in reward and success rate for the four probabilistic dialogue managers described in 5.4.1. The lines plotted are the mean over all speakers. First, observe that *SD 0A* behaves similarly to the other DM when no ASR adaptation is done, but its performance falls below the baseline when the ASR changes. This is because of the mismatch of the data used to train the dialogue models (which comes from an ASR with high error rate) and the data produced by the simulated user. This effect does not occur in *SI 0A*, where the DM is fixed regardless of the ASR performance. This DM is trained on an order of magnitude more data than *SD 0A*, so the extra data may compensate the effects of the mismatch. *SD IA* shows slight improvement in reward but high improvement in success rate, especially when the ASR adaptation is small. This shows that probabilistic DM is especially useful in the first period of time when an EC system of this kind is in use, when there is not enough user specific data to adapt the ASR. It also shows that it is necessary to adapt the DM as the ASR is adapted to improve performance. *SI IA* further improves the results of *SD IA*, showing that using more data from different speakers improves the DM performance.

Fig. 4 (right) shows the relative reward and success improvement only for very low intelligibility speakers. Encouragingly, the relative improvements for these speakers are higher than for the rest, even reaching a 80% success rate improvement with *SI IA*. This is very promising because it might make spoken control systems feasible for these kind of speakers, at the cost of longer interactions.

Fig. 3 right shows the results for the same speakers as in the left part but using *SI IA* DM. Improvements both in reward and success are seen for all speakers, especially when the adaptation in the ASR is small for mid and low intelligibility speakers, and increasing the success rate more than 20% for very low speakers.

7. Conclusions

In this work, the improvements in dialogue performance when coupling an adaptive ASR system with a probabilistic DM are shown. Different DMs are compared and improvements both in dialogue length and success rate with respect to a local confidence based DM baseline are shown. The relative improvement is even higher for lower intelligibility speakers, and it could be considered that for some severe dysarthric speakers this approach would make the spoken interaction feasible (at the cost of longer dialogues, which may be tiring for dysarthric speakers). Adapting the dialogue models at the same time as the ASR performance changes (The interaction environment), further improves the dialogue performance. This improvement is higher in the cases of lower amount of data for adapting the ASR (poorer ASR performance). Finally, we found that a system tailored to a single dysarthric speaker, performs better when boosted by using data from different speakers. From a practical point of view, in a spoken interaction system that uses speaker specific data to improve the ASR, it is important to get the user engaged in using it, especially in the first stages of use of the system. This approach makes the interaction with such a system more robust in the first stages of usage, asking more confirmation questions, and as the ASR improves, the dialogue policy will change to a more straightforward one. This strategy obtains improvements for all different severities of dysarthria, so it is an especially useful framework for a system adapted to a specific user, such as the system presented in [5].

8. Acknowledgements

The research leading to these results was supported by the University of Sheffield studentship network PIPIN and EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

9. References

[1] 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.

[2] S. Young, M. Gasic, B. Thomson and J. D. Williams, "POMDP-Based Statistical Spoken Dialog Systems: A Review," in in *Proceedings of the IEEE, Volume:101 , Issue: 5* , 2013.

[3] F. Jurcicek, B. Thomson, and S. Young. "Reinforcement learning for parameter estimation in statistical spoken dialogue systems," *Computer Speech and Language*, 26(3), 168-192. 2012.

[4] W. Li, J. Glass, N. Roy, and S. Teller, "Probabilistic Dialogue Modelling for Speech-Enabled Assistive Technology," in *4th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2013.

[5] H. Christensen, I. Casanueva , S. Cunningham, P. Green, and T. Hain, "homeService: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition," in *4th Workshop on Speech and Language Processing for Assistive Technologies(SLPAT)*, 2013.

[6] 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*, Brisbane, Australia, 2008.

[7] J. Gauvain and C.H. Lee, "MAP estimation of continuous density hmm: theory and applications," in *Proceeding HLT91 Proceedings of the workshop on Speech and Natural Language,* 1992.

[8] B. Thomson, and S. Young. "Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems," *Computer Speech and Language*, 24(4), 562-588. 2010.

[9] J. D. Williams and S. Young, "Partially observable Markov decision processes for spoken dialog systems," *Computer Speech and Language*, 21(2), 393-422. 2007.

[10] T. Smith and R. G. Simmons, "Point-based pomdp algorithms: Improved analysis and implementation," in *UAI*, 2005.

[11] S. Lee and M. Eskenazi, "Recipe For Building Robust Spoken Dialog State Trackers : Dialog State Tracking Challenge System Description," in *SIGDIAL Conference*, 2013.

[12] M. Gasic, M. Henderson, B. Thomson, P. Tsiakoulis, and S. Young, "Policy optimisation of pomdp-based dialogue systems without state space compression," in *Proc. IEEE SLT Workshop*, 2012.

[13] B. Thomson, M. Gasic, M. Henderson, P. Tsiakoulis, and S. Young, "N-best error simulation for training spoken dialogue systems," in *Proc. IEEE SLT Workshop*, 2012.

[14] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young, "Agenda-based user simulation for bootstrapping a pomdp dialogue system," in *NAACL-'07*, Stroudsburg, PA, USA, 2007.

[15] "Zmdp software for pomdp and mdp planning," http://longhorizon.org/trey/zmdp/.