Exploiting the ASR N-Best by tracking multiple dialog state hypotheses

Jason D. Williams

AT&T Labs – Research, Shannon Laboratory, 180 Park Ave., Florham Park, NJ 07932, USA
jdw@research.att.com

Abstract

When the top ASR hypothesis is incorrect, often the correct hypothesis is listed as an alternative in the ASR N-Best list. Whereas traditional spoken dialog systems have struggled to exploit this information, this paper argues that a dialog model that tracks a distribution over multiple dialog states can improve dialog accuracy by making use of the entire N-Best list. The key element of the approach is a generative model of the N-Best list given the user's true hidden action. An evaluation on real dialog data verifies that dialog accuracy rates are improved by making use of the entire N-Best list.

Index Terms: dialogue modelling, dialogue management, spoken dialogue systems, confidence score, N-Best list

1. Introduction

For spoken dialog systems, speech recognition errors are common, and so identifying and reducing dialog understanding errors is an important problem. One source of potentially useful information is the *N-Best list* output by the ASR engine. The N-Best list contains N ranked hypotheses for the user's speech, where the top entry is the engine's best hypothesis. The hope is that when the top entry is incorrect, the correct entry is contained lower down in the N-Best list.

To illustrate, suppose a dialog system asks the question "Who would you like to call?" twice, receiving 2 N-Best lists. If different names appear in the top positions of the two N-Best lists but the same name appears in the second position on both lists, then intuitively that name ought to be a candidate for the user's intention. What is needed is a dialog model which can detect commonality over multiple N-Best lists and accurately assign a joint likelihood. Unfortunately, traditional dialog systems cannot perform this type of reasoning because they consider only the top entry on each N-Best list and make accept/reject decisions locally.

In this paper we argue that a *probabilistic dialog model* has the properties required to exploit the information on the N-Best list. A probabilistic dialog model tracks a distribution over multiple hypotheses for the current dialog state. Each entry on the N-Best list provides support for its corresponding dialog state hypothesis, and the probability of a dialog state hypothesis aggregates support over all N-Best lists observed in the whole dialog. Procedures for tracking a distribution over multiple dialog states are established in the literature but have been limited to using the top entry on the N-Best list [1, 2, 3, 4]. In this paper we extend a probabilistic dialog model to handle the entire N-Best list, and verify that this results in an improvement in dialog accuracy on real dialog data. Existing work not based on a probabilistic dialog state has suggested combining multiple N-Best lists by summing confidence scores [5, 6] or summing dialog act sequence likelihoods [7]. Because our approach is based on a principled probabilistic model, it is more general and can accommodate expectations about user actions, accurately model the confidence score, account for out-of-grammar speech, and maintain a proper whole-dialog probability for each hidden dialog state.

In this paper, section 2 reviews the mechanics of tracking multiple dialog states; section 3 presents our model of the N-Best list; sections 4 and 5 describe the evaluation data and results; and section 6 concludes.

2. Background

We begin by reviewing the mechanics of tracking multiple dialog, broadly following the SDS-POMDP model [1]. At each turn, the dialog is in some hidden state s, which cannot be directly observed by the dialog system. The state s includes the user's complete goal, such as a travel itinerary like "from London to Boston on June 3".

The dialog system takes a speech action a, such as "Where are you leaving from?". This causes the hidden dialog state s to transition to a new dialog state s' according to a model $P(s'|s,a)$. The user then responds with action u', such as "Boston", according to a model $P(u'|s',a)$. This u' is processed by the speech recognition engine to produce an N-Best list $\tilde{\mathbf{u}}$ and other recognition features $\mathbf{f}$ such as confidence scores, likelihood measures, etc. produced according to a model $p(\tilde{\mathbf{u}}', \mathbf{f}'|u')$. For example, $\tilde{\mathbf{u}}'$ might be an N-Best list with $N = 2$ entries, where the first entry is "AUSTIN" and second entry "BOSTON" and $\mathbf{f}$ might be corresponding confidence scores of 73 and 64.

Since the true state of the dialog s is not directly observed by the dialog system, the dialog systems tracks a distribution over dialogue states $b(s)$ called a *belief state*, with initial belief b_0. At each time-step, b is updated by summing over all possible hidden states and hidden user actions:

$$b'(s') = \eta \cdot \sum_{u'} p(\tilde{\mathbf{u}}', \mathbf{f}'|u') P(u'|s',a) \sum_{s} P(s'|s,a) b(s) \quad (1)$$

where η is a normalizing constant.

Eventually, the dialog system must commit to a particular dialog state to satisfy the user's goal, such as printing a ticket or forwarding a phone call. The dialog state with the highest probability is called s^* and computed as $s^* = \arg\max_s b(s)$ with its corresponding probability being $b^* = \max_s b(s)$. b^* indicates the probability that s^* is correct given all of the system actions and N-Best lists observed so far in the current dialog.

The belief state b is used at run-time to select a system action using some policy $\pi : b \rightarrow a$, and improvements over traditional methods have been reported by constructing the policy using a wide variety of methods, including POMDPs [2] or decision-theory [3]. However, past work on estimating $p(\tilde{\mathbf{u}}, \mathbf{f}|u)$ has been limited to a 1-Best list. In this paper, we show how to estimate $p(\tilde{\mathbf{u}}, \mathbf{f}|u)$ for a full N-Best list. Our overall aim is to show that extending $\tilde{\mathbf{u}}$ from 1-Best to N-Best causes s^* to more often correspond to the true, correct dialog state.

3. Model of an N-Best list

We begin by defining terms. First, the N-Best list is defined as $\tilde{\mathbf{u}} = [\tilde{u}_1, \ldots, \tilde{u}_N]$ where each $\tilde{u}_n$ represents an entry on the N-Best list with $\tilde{u}_1$ being the recognizer's top hypothesis. The ASR grammar $\mathbb{U}$ encodes the set of all user speech u which can be recognized. Thus each N-Best entry is a member of the set of utterances recognized by the grammar $\tilde{u}_n \in \mathbb{U}$. The cardinality of the set of utterances recognized by the grammar is denoted $U = |\mathbb{U}|$.

The user's speech u may be in the grammar ("ig", $u \in \mathbb{U}$), out of the grammar ("oog", $u \notin \mathbb{U}$), or may be silent ("sil", $u = \emptyset$). This "type" is indicated by $t(u)$:

$$t(u) = \begin{cases} \text{ig}, & \text{if } u \in \mathbb{U}, u \neq \emptyset; \\ \text{oog}, & \text{if } u \notin \mathbb{U}, u \neq \emptyset; \\ \text{sil}, & \text{if } u = \emptyset. \end{cases}$$

The user's speech may appear on the N-Best list in position n ("cor(n)", $\tilde{u}_n = u$), may not appear on the N-Best list ("inc", $u \notin \tilde{\mathbf{u}}, \tilde{\mathbf{u}} \neq \emptyset$), or the N-Best list may be empty ("empty", $\tilde{\mathbf{u}} = \emptyset$). We formalize this as $c(\tilde{\mathbf{u}}, u)$:

$$c(\tilde{\mathbf{u}}, u) = \begin{cases} \text{cor}(n), & \text{if } \tilde{u}_n = u, \\ \text{inc}, & \text{if } u \notin \tilde{\mathbf{u}}, \tilde{\mathbf{u}} \neq \emptyset, \\ \text{empty}, & \text{if } \tilde{\mathbf{u}} = \emptyset. \end{cases}$$

We next define a handful of core probabilities $P_e(c(\tilde{\mathbf{u}}, u)|t(u))$ which describe how often various types of errors are made *in general*. For example, $P_e(\text{cor}(n)|\text{ig})$ is the probability that entry n on the N-Best list is correct given that the user's speech is in-grammar, and $P_e(\text{empty}|\text{oog})$ is the probability that the N-Best list is empty given that the user's speech is out-of-grammar. Some entries are estimated from data ($<$ given $>$), and others are derived, as follows:

$$\begin{aligned} P_e(\text{empty}|\text{sil}) &= < \text{given} > \\ P_e(\text{inc}|\text{sil}) &= 1 - P_e(\text{empty}|\text{sil}) \\ P_e(\text{empty}|\text{oog}) &= < \text{given} > \\ P_e(\text{inc}|\text{oog}) &= 1 - P_e(\text{empty}|\text{oog}) \\ P_e(\text{empty}|\text{ig}) &= < \text{given} > \\ P_e(\text{cor}(n)|\text{ig}) &= < \text{given} > \\ P_e(\text{inc}|\text{ig}) &= 1 - P_e(\text{empty}|\text{ig}) - \sum_{n=1}^{N} P_e(\text{cor}(n)|\text{ig}). \end{aligned}$$

For non-empty recognitions, we regard the *length* of the N-Best list N as given and do not model it explicitly; i.e., we will assume that accuracy is conditionally independent of N for non-empty recognitions given the recognition features $\mathbf{f}$. In addition, we will assume that all pairwise confusions at a given N-Best entry n are equally likely. [1]

We can now develop the probability of generating a *specific* N-Best list $\tilde{\mathbf{u}}$ given u, $P_{\tilde{\mathbf{u}}}(\tilde{\mathbf{u}}, c(\tilde{\mathbf{u}}, u)|u, t(u))$. We begin by considering the case where the user says something out of grammar. With probability $P_e(\text{inc}|\text{oog})$ some N-Best list is generated. Since all confusions are equally likely, the probability of all N-Best lists are equal. The probability of the first entry on the N-Best list is $1/U$ (since it is chosen at random from U possible items); the probability of the second entry is $1/(U-1)$ (since 1 entry is no longer available); the probability of the third entry is $1/(U-2)$, and so on. Each generation is independent and so in general (for any N), we have:

$$\begin{aligned} & P_{\tilde{\mathbf{u}}}([\tilde{u}_1, \ldots, \tilde{u}_N], \text{inc}|u, \text{oog}) \\ &\quad = P_e(\text{inc}|\text{oog}) \frac{1}{U} \cdot \frac{1}{U-1} \cdots \frac{1}{U-(N-1)} \\ &\quad = P_e(\text{inc}|\text{oog}) \prod_{i=0}^{N-1} \frac{1}{U-i} \end{aligned}$$

For convenience we define P_b^N:

$$P_b^N := \prod_{i=0}^{N-1} \frac{1}{U-i}$$

and so we can write

$$P_{\tilde{\mathbf{u}}}([\tilde{u}_1, \ldots, \tilde{u}_N], \text{inc}|u, \text{oog}) = P_e(\text{inc}|\text{oog}) P_b^N$$

With probability $P_e(\text{empty}|\text{oog})$, no recognition result is generated and so we have simply:

$$P_{\tilde{\mathbf{u}}}([], \text{empty}|u, \text{oog}) = P_e(\text{empty}|\text{oog})$$

The cases where the user says nothing ($t(u) = \text{sil}$) can be derived in the same way.

Now consider the case where the user says something in grammar. As above, with probability $P_e(\text{empty}|\text{ig})$ no N-Best list is generated. With probability $P_e(\text{inc}|\text{ig})$ the N-Best list is generated and does not contain the correct entry. In this case, the probability of the top entry on the list is $1/(U-1)$ (since the entry the user said is not available); the probability of the second entry is $1/(U-2)$ (because the item the user said and the first N-Best entry aren't available), and so on. So in general we have:

$$\begin{aligned} & P_{\tilde{\mathbf{u}}}([\tilde{u}_1, \ldots, \tilde{u}_N], \text{inc}|u, \text{ig}) \\ &\quad = P_e(\text{inc}|\text{ig}) \cdot \frac{1}{U-1} \cdot \frac{1}{U-2} \cdots \frac{1}{U-N} \\ &\quad = P_e(\text{inc}|\text{ig}) \frac{1}{U-N} \frac{1}{\frac{1}{U}} \prod_{i=0}^{N-1} \frac{1}{U-i} \end{aligned}$$

For convenience we define P_u^i:

$$P_u^i := \frac{1}{U-i}.$$

and so we can write

$$P_{\tilde{\mathbf{u}}}([\tilde{u}_1, \ldots, \tilde{u}_N], \text{inc}|u, \text{ig}) = P_e(\text{inc}|\text{ig}) \frac{P_u^N}{P_u^0} P_b^N$$

With probability $P_e(\text{cor}(1)|\text{ig})$ the N-Best list contains the correct entry in the first position. Since all confusions are equally likely, the probability of the *second* entry is $1/(U-1)$ (since 1 entry is no longer available); the probability of the third

[1] The hope is that phonetically confusable entries will appear on the N-Best list, and so the model will implicitly capture confusability. However, it is an open question whether performance would be improved by explicitly modeling phonetic confusability.

$P_e(c(\tilde{\mathbf{u}},u)\|t(u))$	Train set	Test set
$P_e(\text{cor}(n=1)\|\text{ig})$	82.5%	60.8%
$P_e(\text{cor}(n\geq 2)\|\text{ig})$	8.5%	13.6%
$P_e(\text{inc}\|\text{ig})$	9.0%	25.6%
$P_e(\text{empty}\|\text{ig})$	0.0%	0.0%
$P_e(\text{inc}\|\text{oog})$	100.0%	100.0%
$P_e(\text{empty}\|\text{oog})$	0.0%	0.0%
$P_e(\text{inc}\|\text{sil})$	31.8%	14.3%
$P_e(\text{empty}\|\text{sil})$	68.2%	85.7%

Table 1: *ASR whole-utterance accuracy in the training and testing set.*

entry is $1/(U-2)$, and so on. Each generation is independent and so in general (for any n and N), we have:

$$\begin{aligned}
&P_{\tilde{\mathbf{u}}}([\tilde{u}_1,\ldots,\tilde{u}_N],\text{cor}(n)|u,\text{ig})\\
&\quad= P_e(\text{cor}(n)|\text{ig})\cdot\frac{1}{U-1}\cdot\frac{1}{U-2}\cdots\frac{1}{U-(N-1)}\\
&\quad= P_e(\text{cor}(n)|\text{ig})\prod_{i=1}^{N-1}\frac{1}{U-i}\\
&\quad= P_e(\text{cor}(n)|\text{ig})\frac{1}{\frac{1}{U}}\prod_{i=0}^{N-1}\frac{1}{U-i}\\
&\quad= P_e(\text{cor}(n)|\text{ig})\frac{1}{P_u^0}P_b^N
\end{aligned}$$

In addition to the content of the N-Best list $\tilde{\mathbf{u}}$, the recognition result also includes the set of recognition features $\mathbf{f}$. To model these we define a probability density of the recognition features $\mathbf{f}$ given $c(\tilde{\mathbf{u}},u)$ and $t(u)$ as $p(\mathbf{f}|c(\tilde{\mathbf{u}},u),t(u))$.

We can now state the full model $p(\tilde{\mathbf{u}},\mathbf{f}|u)$:

$$\begin{aligned}
&p(\tilde{\mathbf{u}},\mathbf{f}|u)\\
&\quad= p(\tilde{\mathbf{u}},\mathbf{f},c(\tilde{\mathbf{u}},u)|u,t(u))\\
&\quad= P_{\tilde{\mathbf{u}}}(\tilde{\mathbf{u}},c(\tilde{\mathbf{u}},u)|u,t(u))p(\mathbf{f}|\tilde{\mathbf{u}},c(\tilde{\mathbf{u}},u),u,t(u))\\
&\quad= P_{\tilde{\mathbf{u}}}(\tilde{\mathbf{u}},c(\tilde{\mathbf{u}},u)|u,t(u))p(\mathbf{f}|c(\tilde{\mathbf{u}},u),t(u))
\end{aligned}$$

4. Dialog data and model estimation

To test this model, we applied it to dialogs collected with a voice dialer application in use within the AT&T research laboratory. The dialer's vocabulary consists of 50,000 AT&T employees. The dialer has been operational for several years and during that time has received hundreds of calls. We chose this application because name recognition is difficult, and often calls contain more than one name recognition attempt which is where our approach can improve dialog accuracy.

From the dialer's logs, we gathered all of the name recognition attempts (discarding others such as confirmations). Then we divided the calls into two sets: calls which contained one attempt at name recognition (1224 calls), and calls which contained two or more attempts at name recognition (479 calls). We used the utterances from the one-attempt calls for training the models, and the multi-attempt calls for evaluation.

Table 1 shows the accuracy statistics (P_e) estimated from the training and testing corpora. In the training set, when the caller's speech is in-grammar, the correct answer appears on the top of the N-Best list 82.5% of the time, and further down the N-Best list 8.5% of the time. This provides an indication that there ought to be value in modeling the N-Best list. Also, the overall accuracy in the test set is much worse than the training set. This is a consequence of the dataset partitioning: the dialogs in the test set included two or more turns and this is symptomatic of poorer recognition accuracy.

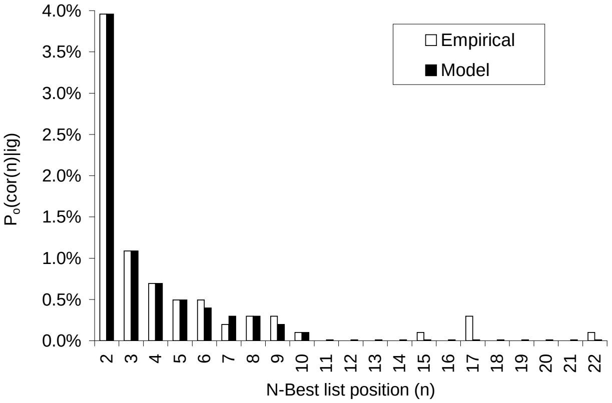

Figure 1: *N-Best accuracy as observed in the training set, and smoothed model $P_e(cor(n)|ig)$ estimated from this data.*

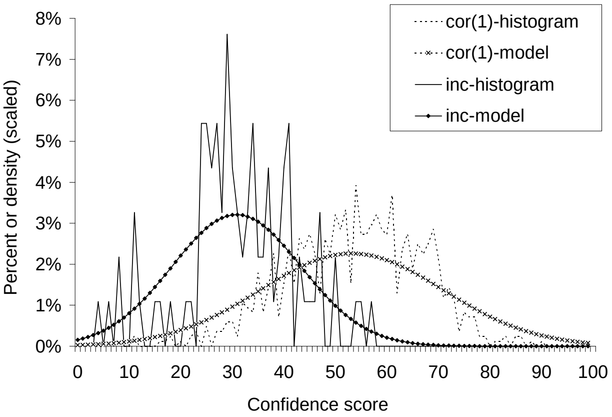

Figure 2: *Histogram and estimated model for the confidence score for 1-best correct* (cor(1)) *and incorrect* (inc) *utterances, given that the user's speech is in-grammar. In general, correct recognitions produce higher confidence scores.*

Figure 1 shows the distribution of correct entries on the N-Best list for the training set. (The graph is truncated at $n=22$ for clarity; the actual model includes entries up to $n=100$.) Most of the mass is concentrated in the first 10 entries, and many correct entries $n>10$ are never observed. As a result, we applied simple smoothing to estimate $P_e(\text{cor}(n)|\text{ig})$ from this data, also shown in Figure 1.

For the recognition features, we used the ASR confidence score. Empirical plots of the confidence score showed that it followed a roughly normal distribution, and so Gaussians were fit to each density of $p(\mathbf{f}|c(\tilde{\mathbf{u}},u),t(u))$. Figure 2 shows two of these densities. Due to data sparsity, we estimated a single density $p(\mathbf{f}|\text{cor}(n),\text{ig})$ for $n\geq 2$.

Finally, a simple user model $P(u'|s',a)$ was also estimated from the training corpus which estimated how often users' speech was in-grammar, out-of-grammar, or silent. Here the dialog state s simply contains the user's goal (the user's desired callee), and so the dialog state was set to be fixed throughout the conversation, $P(s'|s,a)=\delta(s',s)$.

System	Correct dialogs	Percent (of 479)
Baseline	276	57.6%
N-Best	284	59.3%
Oracle (upper bound)	308	64.3%

Table 2: *Dialog accuracy for the 479 calls in the test set.*

5. Experiments and results

Our overall aim is to assess whether our model of $p(\tilde{\mathbf{u}}, \mathbf{f}|u)$ provides an improvement in *dialog accuracy*, defined as the portion of calls which, at the end of the dialog, would have been transferred to the correct callee. This metric attempts to assess whether the caller would achieve their goal and ignores finer-grained measures like word accuracy.

As a baseline, in each call we consider all of the 1-best recognition hypotheses across all turns and chose the one with the highest confidence score. If this hypothesis matches the user's true goal, this call was marked as correct. Then, we estimated an "Oracle" model which forms an upper bound on the effectiveness of our technique. Among the calls which the baseline method yielded the incorrect callee, we observed that our technique could only improve those calls in which the correct name appeared in the N-Best list in two or more dialog turns. We counted the calls on which the baseline method failed and for which this condition was true, and added this count to the baseline performance. Finally, we installed the model $p(\tilde{\mathbf{u}}, \mathbf{f}|u)$ estimated on the training set and ran belief monitoring on the test set. At the end of each call, we looked at the most likely callee in the belief state s^*, and determined if that matched the caller's true goal.

Results are shown in table 2. The baseline yields a dialog accuracy of 57.6%, our method 59.3%, and the oracle 64.3%. The improvement of our method over the baseline is statistically significant at $p < 0.022$ using the two-tail McNemar test. In other words, the maximum improvement of any N-Best processing method over the baseline is 6.7% absolute and our method achieves 1.7% absolute, or 25% of the possible gain. Examining the transcripts, there were 9 dialogs which the N-Best method scored correctly and the Baseline scored incorrectly, and 1 dialog which the Baseline scored correctly and the N-Best method scored incorrectly.

In practice of course, a dialog system does not know whether a hypothesis is correct and relies on a confidence measure to determine whether to accept or reject a recognition hypothesis. In the baseline method, the (highest) confidence score serves as this confidence measure. In our method, the belief in the top dialog state b^* serves this role. As such it is important to evaluate whether b^* has at least the same discriminative power as the confidence score.

Method	System	Accuracy
$\min(R_{FA} + R_{FR})$	Baseline	73.5%
	N-Best	75.2%
$R_{FA} = R_{FR}$	Baseline	71.6%
	N-Best	73.5%

Table 3: *Maximum classification accuracy for the baseline and N-Best methods. Percentages are taken over all 479 calls in the test set.*

To assess this, for a given threshold of the confidence score or b^*, each call was classified as correctly accepted (CA), falsely accepted (FA), correctly rejected (CR), or falsely rejected (FR). Rates R for each call class were computed as fractions of all calls. For example, $R_{CA} = CA/479$. Table 3 reports two evaluations using these classifications. First, the thresholds were varied to search for the minimum sum of FA+FR, $\min(R_{FA} + R_{FR})$. This shows the best possible performance without respect to how FA or FR errors are made. Next, the thresholds were varied to find the equal error rate – i.e., the threshold at which $R_{FA} = R_{FR}$. This shows the performance given the constraint that FA and FR errors are made equally often. The N-Best method outperforms the Baseline method evaluated in either way. An ROC plot, not shown here because of space limitations, also showed a performance improvement. In sum, in addition to more often finding the correct whole-dialog hypothesis, the N-Best method also enables correct and incorrect hypotheses to be better *identified* than does a traditional method.

6. Conclusion

When speech recognition errors occur, the correct hypothesis is often on the ASR N-Best list, yet traditional dialog systems struggle to exploit this information because they rely on local accept/reject decisions. By contrast, given a model of how N-Best lists are generated, maintaining a distribution over many dialog states allows all of the information on the N-Best list to be synthesized. This paper has presented such a method for modelling the N-Best list. This model estimates the probability of generating a particular N-Best list and its features such as confidence score given a true, unobserved user action. Evaluation trained on about a thousand transcribed utterances and evaluated on real dialog data confirms that the method more often finds the correct user goal than the traditional approaches, and that the belief state is a better indication of correctness than local confidence scores. The N-Best model has been incorporated into our voice dialer which is now operational in the AT&T research laboratory [8].

7. References

[1] J. Williams and S. Young, "Partially observable Markov decision processes for spoken dialog systems," *Computer Speech and Language*, vol. 21, no. 2, pp. 393–422, 2007.

[2] S. Young, J. Schatzmann, B. R. M. Thomson, KWeilhammer, and H. Ye, "The hidden information state dialogue manager: A real-world POMDP-based system," in *Proc NAACL-HLT, Rochester, New York, USA*, 2007.

[3] T. Paek and E. Horvitz, "Conversation as action under uncertainty," in *Proc Conf on Uncertainty in Artificial Intelligence (UAI), Stanford, California*, 2000, pp. 455–464.

[4] D. Bohus and A. Rudnicky, "A 'K hypotheses + other' belief updating model," in *Proc AAAI Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems, Boston*, 2006.

[5] A. Kellner, B. Rueber, and H. Schramm, "Strategies for name recognition in automatic directory assistance systems," in *Proc IEEE Interactive Voice Technology for Telecommunications Applications (IVTTA), Torino*, 1998.

[6] N. Kitaoka, H. Yano, and S. Nakagawa, "A spoken dialog system with automatic recovery mechanism from misrecognition," in *Proc SLT, Aruba*, 2006.

[7] H. Higashinaka, M. Nakano, and K. Aikawa, "Corpus-based discourse understanding in spoken dialogue systems," in *Proc ACL, Sapporo*, 2003.

[8] J. D. Williams, "Demonstration of a POMDP voice dialer," in *Proc Demonstration Session ACL-HLT, Ohio*, 2008.