Discovering Tutorial Dialogue Strategies with Hidden Markov Models

Kristy Elizabeth BOYER[a,1], Eun Young HA[a], Michael D. WALLIS[a,b], Robert PHILLIPS[a,b], Mladen A. VOUK[a], and James C. LESTER[a]
[a] *Department of Computer Science, North Carolina State University Raleigh, North Carolina, USA*
[b] *Applied Research Associates, Inc. Raleigh, North Carolina, USA*

Abstract. Identifying effective tutorial strategies is a key problem for tutorial dialogue systems research. Ongoing work in human-human tutorial dialogue continues to reveal the complex phenomena that characterize these interactions, but we have not yet seen the emergence of an automated approach to discovering tutorial dialogue strategies. This paper presents a first step toward establishing a methodology for such an approach. In this methodology, a corpus is first annotated with dialogue acts that are grounded in theories of tutoring and natural language dialogue. Hidden Markov modeling is then applied to discover tutorial strategies inherent in the structure of the sequenced dialogue acts. The methodology is illustrated by demonstrating how hidden Markov models can be learned from a corpus of human-human tutoring in the domain of introductory computer science.

Keywords. Tutorial dialogue, tutorial strategies, machine learning, hidden Markov modeling.

1. Introduction

Tutorial dialogue is a rich form of communication in which a tutor and a learner interact through natural language in support of a learning task. To date, a number of successful tutorial dialogue systems (*e.g.*, AUTOTUTOR [1], BEETLE [2], CIRCSIM [3], GEOMETRY EXPLANATION TUTOR [4], ITSPOKE [5], PROPL [6], RESEARCH METHODS TUTOR [7], and WHY2/ATLAS [8]) have been developed. Current tutorial dialogue research aims to identify and implement tutoring strategies that maximize targeted cognitive and affective outcomes for each learner. Toward that end, rigorous studies of human-human tutorial dialogue have paved the way for a deeper understanding of the conversational phenomena occurring in natural language tutoring (*e.g.*, [9-11]). In addition, there is growing recognition that the most effective strategies in certain learning contexts may not simply be those that occur most frequently with human tutors (*e.g.*, [12, 13]) and that careful evaluation of strategies must be conducted to establish relative effectiveness [14]. These lines of investigation would be well served by a scalable, corpus-based approach to automatically discovering tutorial strategies; however, the field has not yet seen the emergence of such an approach.

[1] Corresponding Author: keboyer@ncsu.edu

Tutorial dialogue studies frequently share a common methodological element: beginning with a set of tutoring strategies, a corpus of tutorial dialogue is examined in a "top-down" fashion for patterns that fit the set of strategies. In contrast to the historical top-down approach, a complementary approach is to examine a corpus of tutorial dialogue in a "bottom-up" fashion. In this methodology, theories from tutoring and natural language dialogue are used to inform corpus annotation at the dialogue act level, and then statistical models are built that induce the strategies inherent in the structure of the sequenced dialogue acts. This direction has been pursued, for example, in work using pairs of dialogue acts to determine the tutoring strategies used to respond to student uncertainty [15]. This research and similar work that endeavored to discover strategies from dialogue acts [16] have made great strides toward assessing the differential impact of localized tutoring strategies on student outcomes. However, these studies are limited to a small window in the tutoring session. Hidden Markov models (HMMs) can move beyond this limitation by inducing a model of tutorial strategies that is based on the entire input sequence of dialogue acts.

This paper presents a first step toward a machine learning-based methodology that automatically discovers tutorial strategies from a corpus of human-human tutorial dialogue. In this approach, a corpus is manually annotated with dialogue acts that are grounded in theories of tutoring and general dialogue. Hidden Markov modeling is then applied to group the dialogue acts into aggregate states that are interpreted as tutorial strategies or dialogue *modes*.[2]

2. Background

2.1. Tutorial Dialogue Strategies in Human-Human Tutoring

Human-human tutorial dialogue has been an active area of research in the Artificial Intelligence in Education community for several decades, motivated in part by the hypothesis that the behavior of tutoring systems should be informed by an empirical understanding of human tutoring. Early work included exploring the cognitive and motivational strategies used by human tutors in a variety of domains (*e.g.,* [17, 18]). Subsequent research has revealed regularities in the structure of natural language tutorial dialogues [9] and suggested hypotheses for why one-on-one tutoring is so effective [10]. Comparative studies of tutoring strategies (*e.g.*, [19]) and contrasting expert and novice tutorial behaviors (*e.g.*, [3]) have shed light on the importance of considering the differential impact of various approaches. Recently, rigorous and relatively large-scale studies of expert human tutors have expanded the field's understanding of tutoring modes [11].[3] In recent corpus-based work, reinforcement learning has been used to compare effectiveness of local tutorial tactics [20] and to determine which dialogue features impact the choice of tutorial dialogue policy [21]; however, this work has not focused on automatically extracting tutorial dialogue structure from unrestricted human-human corpora.

---

[2] The empirically-identified aggregate states reported here do not perfectly correspond to traditional tutoring strategies from the literature; the term *mode* will be used as in [11].

[3] A comprehensive review of human-human tutorial dialogue studies over the past three decades can be found in VanLehn *et al.* [12].

2.2. Hidden Markov Modeling of Tutoring Phenomena

Tutorial interactions in both human-human and human-computer learning environments are often studied by collecting a record of the student-tutor interaction (*e.g.*, dialogue transcripts, student action traces). This record constitutes the observable behavior that results from the tutorial interaction, but in many instances the phenomenon of interest (*e.g.,* tutorial dialogue strategy, student engagement) cannot be directly observed in the data. Hidden Markov modeling is well-suited to such cases because it represents higher-level "hidden states" as probabilistic distributions over the observed values. Beal *et al.* [22] applied hidden Markov modeling to sequences of student actions that had been recorded as the students interacted with a mathematics tutoring system. The model used patterns in the observed student actions to fit the hidden states, which were interpreted as student engagement level. Jeong *et al.* [23] applied HMMs to observed student action traces to model a hidden variable interpreted as activity type, giving insight into the sources of effectiveness between treatment groups in a learning-by-teaching environment. Soller *et al.* [24] used HMMs and clustering on student-student object oriented analysis and design problem-solving session logs to discover the structure of effective and ineffective peer knowledge sharing interactions. Clustering the hidden Markov models based on their likelihood vectors yielded groups of similar models with each group capturing a generalized knowledge-sharing interaction. The application of HMMs to tutorial dialogue reported in this paper is analogous to these prior applications: dialogue acts are treated as the observed values, and dialogue mode is modeled as a hidden state.

3. Tutorial Dialogue Modeling with HMMs

Hidden Markov modeling constitutes a stochastic approach to characterizing the observed signals emitted from a source [25]. The premise of HMMs is that some aspect of the signal source is *hidden* (*i.e.*, not directly observable), and that the values of this hidden variable (*i.e.*, the *hidden states*) are important for modeling the system as a whole. The model is said to be "in" one of the N hidden states at each step in the observed sequence. Each hidden state is characterized by a probability distribution over the observed symbols called the *emission probability distribution* while the transitions among hidden states are governed by the *transition probability distribution*. In a first-order HMM such as the one used here, the transition probability to the "next" hidden state depends only on the current state and not on a longer state history. When training an HMM to model a particular phenomenon, the goal is to select the model that maximizes the probability of the observed input.

3.1. Corpus and Dialogue Act Annotation

The corpus used to train the HMM consists of tutorial dialogue from keyboard-to-keyboard tutoring sessions between human tutors and forty-three novice computer science students. The corpus contains 4,864 dialogue moves which were manually annotated with dialogue act tags (Table 1). Details of the study procedure used to collect the corpus along with learning gains and inter-rater reliability for the tagging scheme are presented in [16].

Table 1 – Dialogue Acts

Act	Description	Tutor and Student Example Utterances	Relative Frequency Across Corpus (%) Student	Tutor
Question (Q)	Questions about goals to pursue, domain concepts, etc.	"Where should we start?" "How do I declare an array?"	5.72	0.53
Evaluative Question (EQ)	Questions that explicitly inquire about student knowledge state or correctness of problem-solving action.	"Do you know how to declare an array?" "Is that right?"	8.55	6.15
Statement (S)	Declarative assertion.	"You need a closing bracket there." "I am looking for where this method is declared.."	4.34	40.85
Grounding (G)	Conversational grounding.	"Alright." or "Okay." "Thanks." or "Hello."	5.12	3.72
Extra Domain (EX)	A statement not related to the computer science discussion.	"The problem description is on your desk." "Can I use my book?"	2.75	3.58
Positive Feedback (PF)	Unmitigated positive feedback regarding problem solving action or student knowledge state.	"Yes, I know how to declare an array." "That is right."	2.38	10.63
Lukewarm Feedback (LF)	Partly positive, partly negative feedback regarding student problem solving action or student knowledge state.	"Sort of." "You're close." or "Well, almost."	0.66	2.01
Negative Feedback (NF)	Negative feedback regarding student problem solving action or student knowledge state.	"No." "Actually, that won't work."	1.89	1.11

3.2. Dialogue Mode Discovery

In this application of HMMs to tutorial dialogue, the observed signal is comprised of tutorial dialogue acts augmented with tags indicating the speaker (*e.g.*, G_S, G_T, Q_S, S_T, S_S, PF_T,…). These observed symbols are provided, without any additional context regarding their meaning, as the input sequence. The hidden variable is interpreted as the dialogue *mode*. Rather than specifying *a priori* the number of dialogue modes, the best-fit number N of hidden states was learned from the observed sequences during model training. For each value of N, seven models were built and each was ten-fold cross-validated on the corpus to obtain an average log-likelihood value.[4] A model containing N=6 hidden states produced the best log-likelihood fit for the current corpus.[5] Figure 1(A) presents the emission probability distribution for each hidden state in the best-fit model.

We interpreted each state as a dialogue mode, and assign intuitive state names, by examining the emission probability distribution of dialogue acts that occur in that state. Because State 0 is dominated by student evaluation questions, statements, and feedback, this state is interpreted as *Student Reflection* mode. State 1 is dominated by extra-domain talk and conversational grounding by both the student and tutor, so this state is interpreted as *Conversational Grounding/Extra-Domain* mode. State 2 consists

[4] Model parameters were learned with the Baum-Welch algorithm, beginning with randomly-initialized parameters and then iterating until convergence [25]. Training between five and ten models is in keeping with standard practice when this random initialization approach is used.

[5] Log-likelihood fit is a measure of how likely the observed sequences would be under a proposed model. The number of hidden states, N, was allowed to range from 2 to 20, with the best fit produced by N=6.

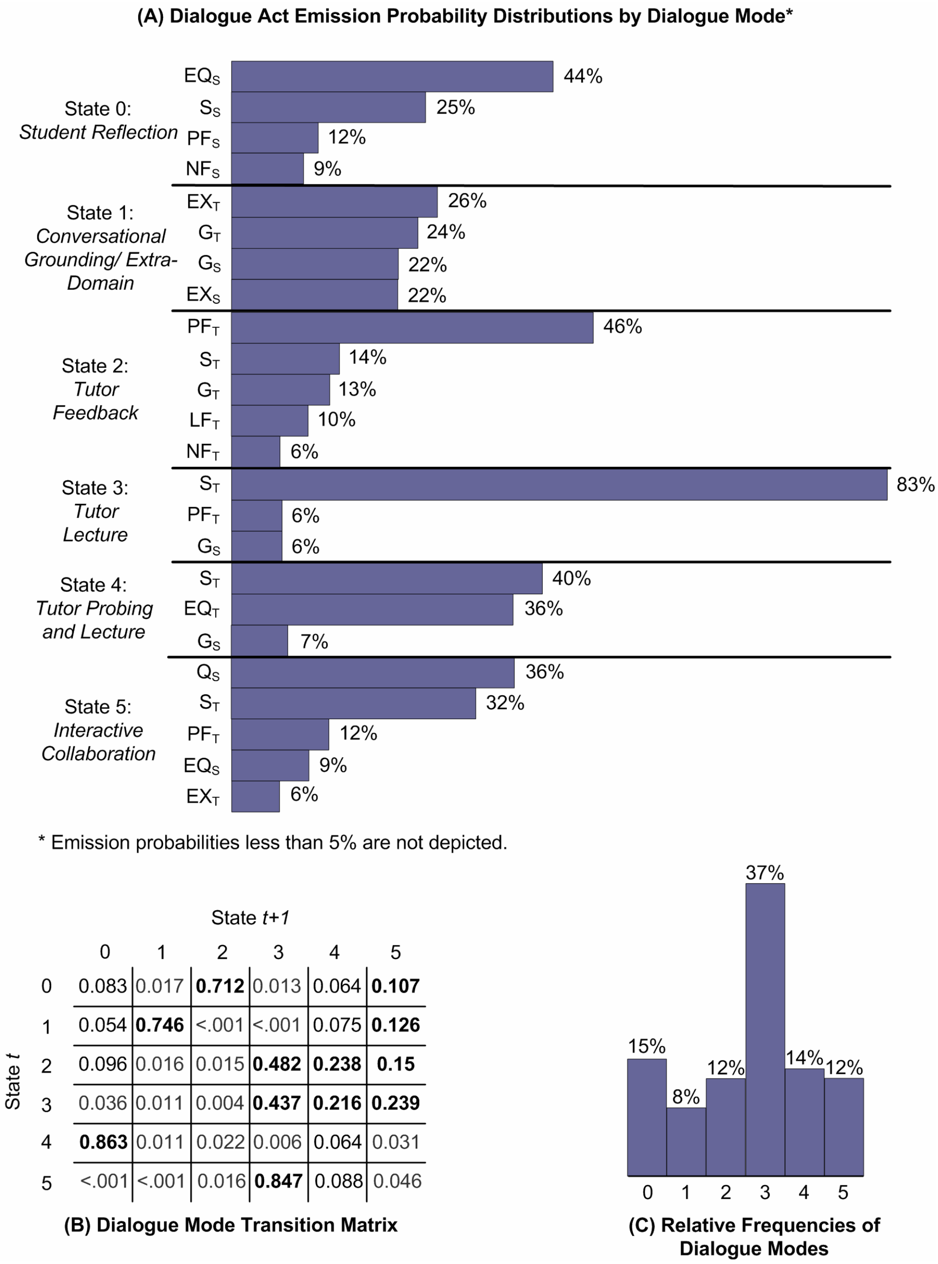

Figure 1 – Learned Tutorial Dialogue Model with *N*=6 Hidden States

primarily of feedback from the tutor, with some statements and tutor grounding, so this state is interpreted as *Tutor Feedback* mode. State 3 is strongly dominated by tutorial statements, so this state is interpreted as *Tutor Lecture* mode. State 4 emits primarily tutor statements and tutor evaluation questions, so this state is interpreted as *Tutor Probing and Lecture*. Finally, State 5 is dominated by a mixture of student questions with tutor statements and feedback, so this state is interpreted as *Interactive Collaboration* mode.

3.3. Dialogue Mode Transitions and Relative Frequencies

The transition matrix in Figure 1(B) depicts the probability of transitioning from one mode to the next. This transition matrix represents the higher-level flow of dialogue. For example, from State 0 (*Student Reflection*), the dialogue transitions with probability 0.712 to State 2 (*Tutor Feedback*) and with probability 0.107 to State 5 (*Interactive Collaboration*). From State 2 (*Tutor Feedback*), the dialogue is most likely to transition to State 3 (*Tutor Lecture*), with State 4 (*Tutor Probing and Lecture*) or State 5 (*Interactive Collaboration*) also likely candidates for the next mode.

Because the learned HMM implies a best-fit sequence of hidden states for each observed sequence of dialogue acts,[6] it is possible to summarize the frequency of each dialogue mode across the corpus as depicted in Figure 1(C). Not surprisingly, State 3 (*Tutor Lecture* mode) occurs most frequently. This result is expected because in the current corpus, tutor statements account for 40% of all dialogue acts (Table 1).

4. Discussion and Limitations

The hidden Markov modeling process groups observed dialogue acts into higher-level hidden states which can then be interpreted as dialogue modes. The dialogue modes discovered in the current corpus are *Student Reflection*, *Conversational Grounding/Extra-Domain*, *Tutor Feedback*, *Tutor Lecture*, *Tutor Probing and Lecture*, and *Interactive Collaboration*. The primary benefit of using HMMs to discover these modes lies with the feasibility offered by a bottom-up, data-driven approach in which the theoretical framework is used to devise a set of dialogue act tags that are applied at a low level—usually within a window of approximately one dialogue turn—and then machine learning techniques aggregate the individual dialogue act tags into higher-level modes. This methodology addresses an important limitation of the contrasting top-down approach, namely, that sophisticated tutoring strategies rarely occur with novice tutors; for example, recent findings suggest that some widely-recognized strategies (*e.g.*, *Model-Scaffold-Fade*) may not occur fully intact even with highly skilled human tutors [11]. Identifying dialogue modes with HMMs circumvents the need to manually "design" tutorial strategies and offers an opportunity to automatically discover which strategies are in fact used in practice.

The discovered dialogue modes presented here do not perfectly map to other sets of handcrafted tutoring modes (*e.g.*, [11]). However, such a perfect mapping rarely exists even between sets of handcrafted labels. It is important to note, though, that the more frequently the learned HMM transitions between different hidden states, the farther from our intuitive notion of "tutoring strategies" the aggregate dialogue modes become. This difficulty is due to the fact that when labeling dialogues by hand, consecutive utterances are considered cohesive, while in hidden Markov modeling, the hidden state may change at every step in the input signal sequence. Therefore, as with other applications of HMMs to tutoring phenomena, the best-fit hidden sequences should be examined carefully to facilitate sound interpretation of the model.

[6] The Viterbi algorithm [25] was used to fit the best sequence of hidden states to each observation sequence.

5. Conclusion and Future Work

The methodology introduced here represents a first step toward empirically identifying tutorial dialogue strategies, or modes, from corpora. In this approach, a theoretical framework drawing from both the natural language dialogue and the tutoring literatures is incorporated through the choice of dialogue act tags, and a hidden Markov model is learned from the sequences of dialogue acts. The methodology of using HMMs to model tutorial dialogue modes can be scaled and applied across corpora, given a mapping between sets of dialogue act tags.

Future work includes expanding the input sequences from dialogue acts alone to include the surface-level utterance content. In addition, knowledge of the task state within the tutoring session can be used to segment the dialogue in meaningful ways to further refine the structure of the HMM. It is also possible that dialogue tagging at different granularities could reveal varying and useful models. Moreover, the HMM approach can be used to compare tutorial strategies for effectiveness by correlating hidden state usage with outcomes of interest, and by training models separately for students in different groups (as in [23]). Using clustering (as in [24]) and a finer-grained knowledge model could also reveal more detailed tutoring strategies. Finally, combining the bottom-up approach presented here with top-down approaches offers promising synergies. It is hoped that the methodology presented here and its variants could be used to identify a set of empirically-derived tutoring modes, and that knowledge of their impact on student learning and affect could inform the development of next-generation natural language tutorial dialogue systems.

6. Acknowledgments

This work is supported in part by the NC State University Department of Computer Science along with the National Science Foundation through Grants REC-0632450 and IIS-0812291, a Graduate Research Fellowship, and the STARS Alliance Grant CNS-0540523. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

References

[1] A. Graesser, G. Jackson, E. Mathews, H. Mitchell, A. Olney, M. Ventura and P. Chipman, Why/AutoTutor: A Test of Learning Gains from a Physics Tutor with Natural Language Dialog, *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society* (2003), 1-6.

[2] C. Zinn, J. D. Moore and M. G. Core, A 3-tier Planning Architecture for Managing Tutorial Dialogue, in *Proceedings of the 6th International Conference on Intelligent Tutoring Systems* (2002), 574-584.

[3] M. Evens and J. Michael, *One-on-One Tutoring by Humans and Computers*. Mahwah, New Jersey: Lawrence Erlbaum Associates (2006).

[4] V. Aleven, K. Koedinger and O. Popescu, A Tutorial Dialog System to Support Self-explanation: Evaluation and Open Questions, *Proceedings of the 11th International Conference on Artificial Intelligence in Education* (2003), 39-46.

[5] D. J. Litman, C. P. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe and S. Silliman, Spoken Versus Typed Human and Computer Dialogue Tutoring, *International Journal of Artificial Intelligence in Education,* **16** (2006), 145-170.

[6] H. C. Lane and K. VanLehn, Teaching the Tacit Knowledge of Programming to Novices with Natural Language Tutoring, *Computer Science Education,* **15** (2005), 183-201.
[7] E. Arnott, P. Hastings and D. Allbritton, Research Methods Tutor: Evaluation of a Dialogue-Based Tutoring System in the Classroom, *Behavior Research Methods,* **40** (2008), 694-698.
[8] K. VanLehn, P. W. Jordan, C. P. Rose, D. Bhembe, M. Bottner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg and A. Roque, The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing, *Proceedings of Intelligent Tutoring Systems Conference* (2002), 158–167.
[9] A. C. Graesser, N. K. Person and J. P. Magliano, Collaborative Dialogue Patterns in Naturalistic One-to-One Tutoring, *Applied Cognitive Psychology,* **9** (1995), 495–522.
[10] M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi and R. G. Hausmann, Learning from Human Tutoring, *Cognitive Science,* **25** (2001), 471-533.
[11] W. Cade, J. Copeland, N. Person and S. D'Mello, Dialog Modes in Expert Tutoring, in *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (2008)*,* 470-479.
[12] K. VanLehn. The Interaction Plateau. Presented at the *9th International Conference on Intelligent Tutoring Systems,* (2008).
[13] M. T. H. Chi, M. Roy and R. G. M. Hausmann, Observing Tutorial Dialogues Collaboratively: Insights About Human Tutoring Effectiveness From Vicarious Learning, *Cognitive Science,* **32** (2008), 301-341.
[14] S. Ohlsson, B. Di Eugenio, B. Chow, D. Fossati, X. Lu and T. C. Kershaw, Beyond the Code-and-Count Analysis of Tutoring Dialogues, in *Proceedings of the 13th International Conference on Artificial Intelligence in Education* (2007).
[15] K. Forbes-Riley and D. Litman, Using Bigrams to Identify Relationships Between Student Certainness States and Tutor Responses in a Spoken Dialogue Corpus, in *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue* (2005).
[16] K. E. Boyer, R. Phillips, M. Wallis, M. Vouk and J. Lester, Balancing Cognitive and Motivational Scaffolding in Tutorial Dialogue, in *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (2008), 239-249.
[17] M. R. Lepper, M. Woolverton, D. L. Mumme and J. L. Gurtner, Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-Based Tutors, in *Computers as Cognitive Tools*, S. P. Lajoie and S. J. Derry, Eds. Hillsdale, New Jersey: Lawrence Erlbaum Associates (1993), 75-105.
[18] B. A. Fox, *The Human Tutorial Dialogue Project.* Hillsdale, New Jersey: Lawrence Erlbaum Associates (1993).
[19] C. P. Rose, J. D. Moore, K. VanLehn and D. Allbritton, A Comparative Evaluation of Socratic Versus Didactic Tutoring, Tech. Rep. #LRDC-BEE-1 (2000).
[20] M. Chi, P. Jordan, K. VanLehn and M. Hall, Reinforcement Learning-Based Feature Selection for Developing Pedagogically Effective Tutorial Dialogue Tactics, in *Proceedings of the 1st International Conference on Educational Data Mining* (2008), 258-265.
[21] J. R. Tetreault and D. J. Litman, A Reinforcement Learning Approach to Evaluating State Representations in Spoken Dialogue Systems, *Speech Commun.,* **50** (2008), 683-696.
[22] C. Beal, S. Mitra and P. R. Cohen, Modeling Learning Patterns of Students with a Tutoring System Using Hidden Markov Models, in *Proceedings of the 13th International Conference on Artificial Intelligence in Education* (2007), 238-245.
[23] H. Jeong, A. Gupta, R. Roscoe, J. Wagster, G. Biswas and D. Schwartz, Using Hidden Markov Models to Characterize Student Behaviors in Learning-by-Teaching Environments, in *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (2008), 614-625.
[24] A. Soller and R. Stevens, Applications of stochastic analyses for collaborative learning and cognitive assessment, in *Advances in Latent Variable Mixture Models* G. R. Hancock and K. M. Samuelsen, Eds. Information Age Publishing (2007), 217-253.
[25] L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proc IEEE,* **77** (1989), 257-286.