Sorting Pauses in Speech Based Dialog Systems: Effects of Knowing Whether Pauses in Speech are Mid-Turn Or At The End of The Turn

Mohamed Jemmali
Northeastern Illinois University
Chicago, IL, U.S.A.
mjemmali@neiu.edu

Ginger Dragon
Northeastern Illinois University
Chicago, IL, U.S.A.
G-dragon@neiu.edu

Raelyn Mendoza
Northeastern Illinois University
Chicago, IL, U.S.A.
rmmendo8@neiu.edu

Daniel Stille
Northeastern Illinois University
Chicago, IL, U.S.A.
dstille@neiu.edu

Patrick Poinar
Northeastern Illinois University
Chicago, IL, U.S.A.
papoinar@neiu.edu

Aditya Ojha
Purdue University
West Lafayette, IN, U.S.A.
ojha0@purdue.edu

Kevin Baez
Northeastern Illinois University
Chicago, IL, U.S.A.
kabaez@neiu.edu

Sooyeon Jeong
Purdue University
West Lafayette, IN, USA
sooyeonj@purdue.edu

Renu Balyan
SUNY Old Westbury
Long Island, NY, USA
balyanr@oldwestbury.edu

Francisco Iacobelli
Northeastern Illinois University
Chicago, IL, U.S.A.
f-iacobelli@neiu.edu

ABSTRACT

Intelligent tutoring systems (ITS) can provide scalable and accessible education at scale. Generally, speech-based tutoring systems interpret the users' long pauses as a signal to take the floor. However, such behavior create barriers for minority and low-literacy individuals who pause mid-sentence just to finish articulating their thoughts. For them, these systems inappropriately interrupt them, resulting in frustration. In this study, we present a bigram pause detection model that determines whether a pause is occurring in the middle or at the end of a sentence and implemented this model on an ITS (MODIFIED). We also compared the effect of this modification with a baseline ITS that interprets long pauses as the end of a turn (TRADITIONAL). Our results show that participants interacted longer with MODIFIED, and perceive it more human-like than the TRADITIONAL system. We argue that conversational systems need to be designed considering unique needs of their intended interlocutors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; *Empirical studies in HCI*; • **Applied computing** → **Interactive learning environments**.

KEYWORDS

Speech Based Tutoring Systems, Linguistic Patterns, Human-Computer Interaction

ACM Reference Format:
Mohamed Jemmali, Ginger Dragon, Raelyn Mendoza, Daniel Stille, Patrick Poinar, Aditya Ojha, Kevin Baez, Sooyeon Jeong, Renu Balyan, and Francisco Iacobelli. 2018. Sorting Pauses in Speech Based Dialog Systems: Effects of Knowing Whether Pauses in Speech are Mid-Turn Or At The End of The Turn. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Intelligent Tutoring Systems (ITS) emulate human tutors by providing real-time feedback, guidance, and personalized learning experiences. They can often result in enhanced student engagement, academic achievement, and retention[12]. ITSs also have a great potential for providing accessible health-related support and improving their health literacy for marginalized populations [5]. For people with low literacy, conversational ITS can especially be beneficial since they can leverage spoken language to communicate with users.

However, research shows that there are cultural variations in how people interact with voice user interfaces (VUI). In particular, lower literacy Hispanics may exhibit long pauses in the middle of

sentences that may serve to finish articulating ideas and not as a signal to yield the floor [?]. However, traditional speech-based dialogue systems interpret long pauses as the end of a turn and take the floor whenever a pause occurs. This is detrimental when interacting with individuals who often pause mid-sentence because the system would interrupt the user inappropriately and can cause frustration and confusion, resulting in negative interaction experience.

To close this gap, we developed a bi-gram pause detector model that classifies whether a sentence is complete or incomplete when a pause occurs. This model was implemented on a health education tutoring system designed to interact with Hispanic adults. We conducted a between-subject study to compare this tutoring system with a baseline tutoring system that considers every pause as an end-of-speech. The first variation is traditional in that it waits for a long pause and interprets it as the end of the user's turn (TRADITIONAL). The second variation classifies whether the pause occurred in an unfinished sentence and acts accordingly (MODIFIED). If the pause occurred in mid-sentence, the system prompted the user to complete the rest of the utterance with encouraging utterances (e.g., "go on"). If the pause occurred at the end of a sentence the system interprets it as the end of the user's turn.

1.1 Background

Conversational Intelligent Tutoring Systems (CITS) present a promising avenue to engage students via human-like interactions. For example, Auto Tutor [7] used embodied conversational agents to support learning gains for students in many different domains [16]. Users of tutoring systems have different needs and learning styles, and Latham et al. [14] introduced Oscar, an Intelligent Adaptive Conversational Agent Tutoring System designed to dynamically predict and adapt to each students' learning style. Guo et al. [8] provided a multidisciplinary perspective on ITS research, and highlights the role of adapting speech-processing technologies to enhance tutoring systems' reach. However, most existing ITS are designed to support students who are fluent in the language used by the system, and there exists a need to explore how conversational ITS can optimize engagement and learning outcomes for people with diverse speech patterns.

Automatic speech recognition (ASR) has advanced significantly over the past several decades. For example, ASR can detect filler words in pauses, such as "um", "erm", etc., [3], stutter [19], and correction of disfluencies [17]. These capabilities, when incorporated on the ITS system, can lead to a better understanding of users and learning experience. Kang et al. [11] developed a tutoring system in English that is able to understand non-native speakers. However, challenges persist in accurately recognizing speech patterns, particularly pauses in diverse learner populations. Janning et al. [10] addressed these challenges by investigating speech and pause patterns, highlighting the importance of context-aware algorithms for effective tutoring interactions. To this effect, some ASR systems use prosody to detect the end of a query [18]. However, recent research [?] suggests that low-literacy Latinas use long pauses in the middle of sentences before they finish articulating an idea and that it is often the case that the ASR fails to interpret that pause correctly (i.e. not yielding the floor). Failing to understand whether a pause is in the middle vs at the end of an utterance could be detrimental to ITSs aimed at low literacy populations as typical ASR systems commonly interpret long pauses as the end of the user's turn and may take the floor, interrupting the thought process of the user and acting on incomplete information [?].

2 METHODOLOGY

2.1 Bi-gram Pause Detection Model

We developed a simple bigram language model that identifies complete and incomplete sentences. In our model, each token (words, punctuation, numbers, etc.) in a sentence was mapped to its part of speech (POS) tag as returned by spaCy's POS tagger. Therefore, each sentence became a sequence of POS tags, with the last tag marking the end of a sentence. With those, our model computes the probability that the last two tags in the sequence are followed by the end of a sentence tag. That is, $P(EOS|t_k, t_{k-1})$ where EOS indicates the end of a sentence, k is the number of POS tags in the sequence, t_i is the tag at position i in the sequence. The parameters of the model were trained using 20,000 sentences and questions from Training Set #5 of the Experimental Data for Question Classification [15]. To test the model, we used a subset of 800 questions and statements from the training set, 380 sentences taken from the SPAADIA corpus –a set of annotated telephone conversations citeLeech2003GenericDialogue, and 800 sentences taken from a corpus of senate transcriptions [4]. With these sentences, we randomly truncated 50% of them at random places and left the other 50% complete. We then, had the system use our language model to classify each sentence as either a complete or an incomplete sentence. Finally, We used a receiver operating characteristics (ROC) curve to determine the robustness and optimal threshold for classification. Figure 1 shows the ROC curve with an area of 0.82. The curve shows our classification method, with an optimum threshold of 0.15, yielding an accuracy of 75.1% and an F1 score of 79.3. The true positive rate was 0.95 and the false positive rate was 0.45.

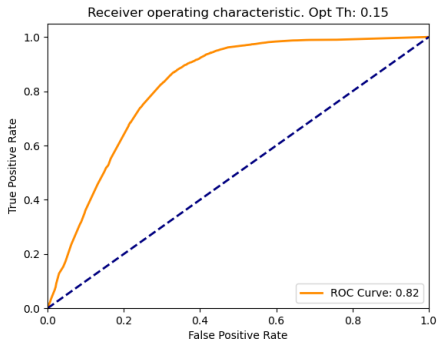


Figure 1: ROC curve of the different thresholds that could be used to classify a sentence as incomplete. The optimum threshold is 0.15

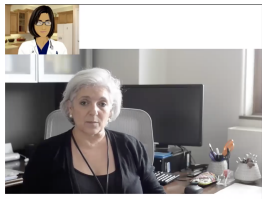
2.2 Experimental Study

We recruited a total of 16 adults for our study, comprising 8 males and 8 females, all of whom are 21 years or older. All participants

stated their ethnicity as Hispanic. One participant was not fluent in English, and two participants were not comfortable speaking Spanish. The educational attainment in the group was mostly some college or above, with only one participant reporting high school education as their highest level of education.

Participants were asked to fill out demographic questionnaires and a health literacy questionnaire (HLiTT) [9]. In HLiTT, larger scores correlate to better health literacy. They were also asked to indicate where they thought they were in the McArthur socioeconomic ladder [1] to understand their perceived status in society. Scores range from 1 (the lowest place in society) to 10 (the highest place in society).

Participants were then randomly assigned to one of two conditions: the **TRADITIONAL** condition interacted with a traditional speech-based tutoring system that interpreted the user's long pauses as yielding a turn. The system used Google Cloud Speech-to-Text[1] with its default settings for speech recognition and turn detection. The **MODIFIED** condition interacted with a modified tutoring system that could classify whether a pause occurred in mid-sentence or at the end of a sentence. If pauses were detected mid-sentence, the system to make sure the user had nothing else to say, or to encourage her to finish the sentence. For this, the system used utterances such as "Go on." or "Is that it?". All other utterances and algorithm of the agent's operation were identical in both conditions. A depiction of the agent while watching a video with a participant can be seen on Figure 2 (a), and the agent in full screen while talking with the participant in Figure 2 (b).



(a) Participants first watched an education video with the Agent. When co-viewing the video, the agent is displayed smaller to give more space to the video.

(b) After co-viewing the video, the Agent is displayed in fullscreen mode and the interaction was recorded via Zoom. The face of the participant has been altered for privacy.

Figure 2: Two interactions modes with the ITS

The interaction with the tutoring system consisted of watching two videos about breast cancer. After each video, the tutoring system asked questions to elicit users' reflection about the topics of the videos. At the end of these two conversations, participants were asked to fill out a questionnaire about the content of the videos. The interactions were timed and transcribed automatically by the speech recognition engine.

Then, they were asked to fill out a usability questionnaire that has been used in previous research[20]. This questionnaire asked users to rate their level of agreement with statements like "The system understood what I was saying", "It is easy to learn to use [the system]", "I found the system unnecessarily complex", "I would recommend this system to a friend", etc. We then inverted the scores of four negative statements so that for all questions, a higher score indicated better usability.

Lastly, they had to fill out a Godspeed questionnaire [2], which assesses the perception of social robots and embodied conversational agents. The questionnaire asks questions regarding anthropomorphisms, animacy, likeability, perceived intelligence, and perceived safety of the computer agent.

Three participants were excluded from this analysis because they interacted with an initial glitchy version of the system. This yielded 6 participants in TRADITIONAL and 7 in MODIFIED.

3 RESULTS

Participants in the MODIFIED condition engaged with the system for 720.2 (SD: 115.3) seconds and uttered 328.9 (SD: 199.2) words on average, while participants in TRADITIONAL condition engaged with the system for 451.9 (SD: 42.3) seconds utilizing 118.3 (SD: 62) words. The Welche's t-test revealed that there were statistically significant differences in the number of words spoken, $t(7.32) = -2.65, p < 0.05$ (Fig 3). We also found a statistically significant effect of the experimental condition on the time participants interacted with each ITS system, $t(7.8) = -5.72, p < 0.05$.

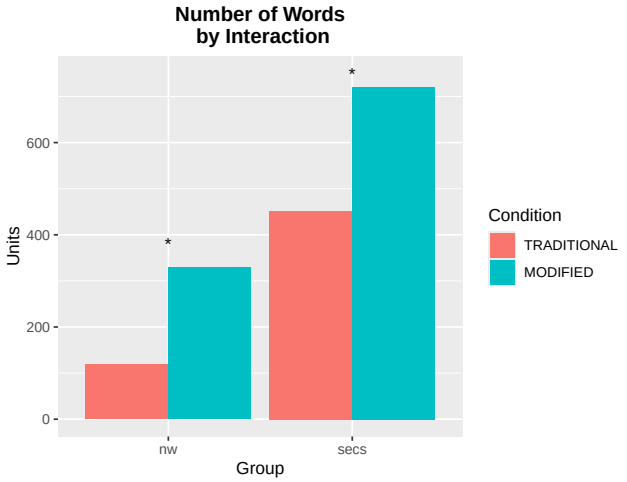


Figure 3: Number of words spoken and duration of the interaction in seconds.

Participants' perceived usability showed a statistically significant difference between the two experimental groups on statement 06 ("It is easy to learn to use [the system]") with the mean scores $M_{TRADITIONAL} = 3.67$ $(SD : 1.21)$ for the TRADITIONAL condition and for $M_{MODIFIED} = 4.7$ $(SD : 0.48)$ for the MODIFIED condition, $t(6.38) = -1.99, p < 0.05$. For statement 2 ("I understood what the system was saying") and statement 5 ("I learned to use the system quickly") the differences were not statistically significant, but suggest a trend ($p = 0.08$ and $p = 0.06$, respectively). All other items did not show statistically significant differences (Table 1).

Analysis of the Godspeed questionnaire reveals that in the category of anthropomorphism (how human-like the character is in

[1]https://cloud.google.com/speech-to-text

Table 1: Mean scores of each usability questionnaire item. (*) denotes a difference significant at p<0.05

Usability Statement	TRADITIONAL	MODIFIED
S01. The system understood what I was saying	2.83	2.85
S02. I understood what the system was saying	3.83	4.71
S03. I think I have learned more about breast cancer	3.67	4.00
S04. I enjoyed my interaction with the system	3.00	3.43
S05. I learned to use the system quickly	3.83	4.71
S06. It is easy to learn to use it	3.67	4.71*
S07. I would recommend this system to a friend	3.83	3.28
S08. It is fun to interact with the system	3.67	3.71
S09. I think that I would like to use this system frequently	3.17	2.86
S10. I found the system unnecessarily complex	2.83	2.42
S11. I think that I would need the support of a technical person to be able to use this system	3.00	2.57
S12. I found the system very cumbersome to use	2.67	1.86
S13. I felt very confident using the system	3.83	4.29
S14. I needed to learn a lot of things before I could get going with this system	3.0	3.14

appearance from 0 Fake to 5 Natural), participants in the MODIFIED condition gave a higher score than those in the TRADITIONAL condition ($M_{TRADITIONAL}$ = 2.83 (SD : 0.75), $M_{MODIFIED}$ = 3.86 (SD : 0.69), $t(10.3) = -2.54$, $p < 0.05$). For the item that asked them to rate from 0 (Moving Rigidly) to 5 (Moving Elegantly), there was a trend of participants in the MODIFIED rating the agent higher, $M_{TRADITIONAL}$ = 2.67 (SD : 1.03), $M_{MODIFIED}$ = 3.7 (SD : 1.11), $t(10.9) = -1.75$, $p < 0.06$ (Figure 4). For the Animacy item in the Godspeed questionnaire, in which users are asked to evaluate how close to a living, dynamic, and lively creature they perceive the system to be, participants in MODIFIED rated the system as being more organic ($M_{TRADITIONAL}$ = 2.8(SD : 0.41), $M_{MODIFIED}$ = 3.7 (SD : 1.11), $t(7.8) = -1.94$, $p < 0.05$) and tended to evaluate it more lively ($M_{TRADITIONAL}$ = 3.17 (SD : 0.76, $M_{MODIFIED}$ = 4.0 (SD : 1.0, $t(10.86) = -1.7$, $p < 0.06$) than the participants in TRADITIONAL (Figure 5). There were no significant differences in the items regarding perceived safety, likeability, and intelligence.

Finally, in the learning tests about the first video, the TRADITIONAL group scored 39.16% (SD = 12.8) while the MODIFIED group scored 47.85% (SD = 18.0). On the second video, the TRADITIONAL group scored 48.3% (SD = 27.1) and the MODIFIED group scored 52.8% (SD = 20.5) These differences, however, were not statistically significant at the p<0.05 level.

4 DISCUSSION

Looking at time spent with the system and words spoken to it, which can be used as a proxy for measuring certain aspects of engagement post-taks [6, 13], we find that users in MODIFIED spent significantly more time and spoke significantly more words with the agent than those participants in TRADITIONAL.

Results from the usability questionnaire show that there are important differences in how the MODIFIED system was perceived as more usable than the TRADITIONAL system ("*It is easy to learn to use [the system]*"). Although there was no statistically significant difference When we look at the Godspeed questionnaire, we can see that although there are no significant differences between groups in the categories of likeability, perceived intelligence, and

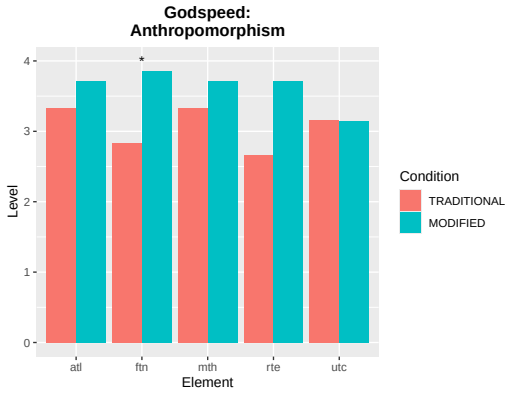


Figure 4: Perceived anthropomorphism: the higher the score, the more anthropomorphic qualities users attribute to the agent. A mark at the top of the bars indicates a difference that is significant at the p<0.05 level.

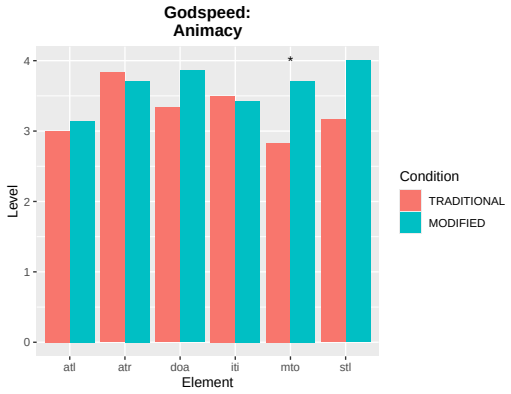


Figure 5: Perceived animacy between the two groups: the higher the score, the more alive and animated the computer agent is perceived to be. A mark at the top of the bars indicate a difference that is significant at the p<0.05 level

perceived safety, participants in MODIFIED generally find more anthropomorphic attributes in the agent. Especially, when asked whether the agent seems fake or natural. Users in MODIFIED also find the agent more organic and less machine-like than those in TRADITIONAL.

Finally, when assessing learning, participants in MODIFIED obtained higher scores than those in TRADITIONAL. However, it is important to note that most participants were highly educated (college or above) and an agent like this may not make a significant difference in their learning.

5 CONCLUSION

When people interact with computer agents via speech, they do not always pause to yield their turn. They sometimes pause mid-thought, before finishing a sentence. However, if the pause is too long, traditional systems will interpret that as a cue to take the floor. In this study, we explored an intelligent tutoring system with an embodied conversational agent that detected whether a long pause occurred in the middle or at the end of a sentence. We assigned some participants to interact with a traditional agent (where a pause yields the turn) and others with this modified agent.

We found that those who interacted with the modified agent (MODIFIED) generally found the agent more human-like in appearance and dynamics and perceived it to be more user-friendly. We also found that they were more engaged with the agent.

We believe that despite the small number of participants on each condition, these results point towards a larger trend and we intend to increase the sample size of this study. In addition, because these users are highly educated, we believe there were no significant differences in learning. However, our results are still promising as better usability, engagement, and perceived usability could potentially lead to larger learning benefit for individuals with low literacy.

6 ACKNOWLEDGEMENTS

This project was funded by the National Science Foundation (Award: 2219586) and by the Northwestern Center for Advancing Machine Intelligence (CASMI), as well as a private grant from the Prasad family.

REFERENCES

[1] Nancy Adler, Judy Stewart, and The Psychosocial Working Group. 2007. *The MacArthur Scale of Subjective Social Status.* Technical Report. University of California, San Francisco. http://www.macses.ucsf.edu/research/psychosocial/subjective.php

[2] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1 (2009), 71–81.

[3] Aggelina Chatziagapi, Dimitris Sgouropoulos, Constantinos Karouzos, Thomas Melistas, Theodoros Giannakopoulos, Athanasios Katsamanis, and Shrikanth Narayanan. 2022. Audio and ASR-based Filled Pause Detection. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 1–7. https://doi.org/10.1109/ACII55700.2022.9953889

[4] Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. 2012. Language and Ideology in Congress. *British Journal of Political Science* 42, 1 (1 2012), 31–55. https://doi.org/10.1017/S0007123411000160

[5] Ying Fang, Anne Lippert, Zhiqiang Cai, Su Chen, Jan C. Frijters, Daphne Greenberg, and Arthur C. Graesser. 2021. Patterns of Adults with Low Literacy Skills Interacting with an Intelligent Tutoring System. *International Journal of Artificial Intelligence in Education* (8 2021). https://doi.org/10.1007/s40593-021-00266-y

[6] Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive Engagement: An Efficient Metric for Automatic Evaluation of Open-Domain Dialogue Systems. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (4 2020), 7789–7796. https://doi.org/10.1609/aaai.v34i05.6283

[7] A C Graesser, P Chipman, B C Haynes, and A Olney. 2005. AutoTutor: an intelligent tutoring system with mixed-initiative dialogue. *Education, IEEE Transactions on* 48, 4 (2005), 612–618. https://doi.org/10.1109/te.2005.856149

[8] Lu Guo, Dong Wang, Fei Gu, Yazheng Li, Yezhu Wang, and Rongting Zhou. 2021. Evolution and trends in intelligent tutoring systems research: a multidisciplinary and scientometric view. *Asia Pacific Education Review* 22, 3 (2021), 441–461.

[9] Elizabeth A Hahn, Seung W Choi, James W Griffith, Kathleen J Yost, and David W Baker. 2011. Health literacy assessment using talking touchscreen technology (Health LiTT): a new item response theory-based measure of health literacy. *Journal of health communication* 16 Suppl 3, Suppl 3 (2011), 150–62. https://doi.org/10.1080/10810730.2011.605434

[10] Ruth Janning, Carlotta Schatten, and Lars Schmidt-Thieme. 2015. Recognising perceived task difficulty from speech and pause histograms. In *International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2015).* 14.

[11] Byung Ok Kang, Hyung-Bae Jeon, and Yun Kyung Lee. 2024. AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation. *ETRI Journal* (2024), e12646.

[12] Abdulkadir Karaci, Halil Ibrahim, Goksal Bilgici, and Nursal Arici. 2018. Effects of Web-based Intelligent Tutoring Systems on Academic Achievement and Retention. *International Journal of Computer Applications* 181, 16 (9 2018), 35–41. https://doi.org/10.5120/ijca2018917806

[13] Chandra Khatri, Anu Venkatesh, Behnam Hedayatnia, Ashwin Ram, Raefer Gabriel, and Rohit Prasad. 2018. Alexa Prize — State of the Art in Conversational AI. *AI Magazine* 39, 3 (9 2018), 40–55. https://doi.org/10.1609/aimag.v39i3.2810

[14] Annabel Latham, Keeley Crockett, David McLean, and Bruce Edmonds. 2011. Oscar: an intelligent adaptive conversational agent tutoring system. In *Agent and Multi-Agent Systems: Technologies and Applications: 5th KES International Conference, KES-AMSTA 2011, Manchester, UK, June 29–July 1, 2011. Proceedings 5.* 563–572.

[15] Xin Li and Dan Roth. 2002. Learning Question Classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics.*

[16] Benjamin D. Nye, Xiangen. Hu, Arthur C. Graesser, and Zhiqiang Cai. 2014. AutoTutor in the cloud: a service-oriented paradigm for an interoperable natural-language ITS. *Journal of Advanced Distributed Learning Technology* 2, 6 (2014), 49–63.

[17] Johann C. Rocholl, Vicky Zayats, Daniel D. Walker, Noah B. Murad, Aaron Schneider, and Daniel J. Liebling. 2021. Disfluency Detection with Unlabeled Data and Small BERT Models. (4 2021).

[18] Matt Shannon, Gabor Simko, Shuo Yiin Chang, and Carolina Parada. 2017. Improved end-of-query detection for streaming speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 2017-August. https://doi.org/10.21437/Interspeech.2017-496

[19] Olabanji Shonibare, Xiaosu Tong, and Venkatesh Ravichandran. 2022. Enhancing ASR for Stuttered Speech with Limited Data Using Detect and Pass. (2 2022).

[20] Betina R. Yanez, Diana Buitrago, Joanna Buscemi, Francisco Iacobelli, Rachel F. Adler, Marya E. Corden, Alejandra Perez-Tamayo, Judy Guitelman, and Frank J. Penedo. 2018. Study design and protocol for My Guide : An e-health intervention to improve patient-centered outcomes among Hispanic breast cancer survivors. *Contemporary Clinical Trials* 65 (2 2018), 61–68. https://doi.org/10.1016/j.cct.2017.11.018

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009