G-g-go! Juuump! Online Performance of a Multi-keyword Spotter in a Real-time Game

Jill Fain Lehman[1], *Nikolas Wolfe*[1], *André Pereira*[1]

[1]Disney Research, 4720 Forbes Ave, 15213 Pittsburgh, PA, USA.
Jill.Lehman@disneyresearch.com

Abstract

We report results for an online multi-keyword spotter in a game that contains overlapping speech, off-task side talk, and keyword forms that vary in completeness and duration. The spotter trained on a data set of 62 children, and expectations for online performance were established by 10-fold cross-validation on that corpus. We compare the *post hoc* data to the recognizer's performance online in a study in which 24 new children played with the real-time system. The online system showed a non-significant decline in accuracy which could be traced to trouble understanding the *jump* keyword and the predominance of younger children in the new cohort. However, children adjusted their behavior to compensate, and the overall performance and responsiveness of the online system resulted in engaging and enjoyable gameplay.

Index Terms: automatic speech recognition, speech analysis, human-computer interaction

1. Introduction

Existing speech technology tends to be poorly suited for young children at play, both because of their age-specific vocal characteristics and because they typically play simultaneously. With this population, side banter, turn-taking directives, and emotional outbursts signaling excitement and delight are continuously interwoven into the activity and seem intrinsic to its enjoyment [1, 2]. Because our primary goal is to create language-based experiences that are engaging and fun, we work to develop Automatic Speech Recognition (ASR) based play that will support young children's natural behavior. In previous publications we have described the underlying multi-keyword spotting algorithm [3] and extensions to the basic algorithm required for online performance [4] in detail, as well as the proxy scale for engagement used to code our data [5]. We review those aspects of the system briefly but focus here on the culmination of the work by presenting online performance results for pairs of children playing a fast-paced, voice-controlled game in the presence of challenging phenomena like keyword variation, social side-talk, and overlapping speech.

The online results reported here are contrasted with the offline performance of the same word spotter on an earlier data set [4]. In the games that produced the earlier data, keyword recognition was performed by a human wizard, and the spotter was evaluated using a *post hoc* 10-fold cross-validation on the children's audio. The system's overall performance was good, with mean accuracy, precision, and recall of 83%, 83% and 93% respectively.

In this paper, we are concerned with how children experience and play the game with the word spotter as ASR. Note that what children experienced before was an interaction with the wizard's recognition performance. Their actions were a response to the qualities and characteristics of that performance. To the extent that the online spotter behaves differently than the wizard, children may change their behavior, exposing problems that were hidden in the offline analysis and/or compensating for weaknesses in the technology by their own adaptation. Thus, to evaluate the system's performance in autonomous, real-time play, we ask and answer three questions: Is performance during actual gameplay significantly different from the expectations set by the *post hoc* analysis? If so, do the differences create corresponding changes in the children's behavior? And if that is also true, do those changes impact the children's enjoyment?

1.1. The Training Corpus

Mole Madness (hereafter, MM) is a speech-controlled interactive video game similar to the Super Mario Bros®. Gameplay consists of two players moving an animated mole through its environment, acquiring rewards and avoiding obstacles (see Figure 1). In a simple design that is easy for even very young children, one player creates horizontal movement using the word *go* and the other creates vertical movement with *jump*. Successful movement requires coordinated, sometimes simultaneous turn-taking, leading to typical behavior for multi-child interaction: overlapping speech and the use of task vocabulary in side-talk [6]. The limited expressiveness of the two-word vocabulary also gives rise to creative keyword variations using elision ("G- g-go") and elongation ("Juuuuump"). Although the meaning of the alternate forms is game-specific, variability in vowel duration is a general problem in understanding children's speech [7].

Sixty-two children between the ages of five and ten played MM as part of a multi-activity data collection in July and August of 2015 that we call "The Summer Games" (hereafter, SG). Participants' mean age was 7.45 years (standard deviation 1.44 years), and 48% were female.[1] Children were compensated for their participation.

Speech recognition during SG was performed by a wizard who listened via headphones in a separate room, trying to map each command to a button press on a game controller. The study produced ~7 hours of gameplay which was then hand-segmented and transcribed to create the corpus of ~11.8K non-overlapping instances of *go*, ~9.4K non-overlapping instances of *jump*, ~10.1k instances of overlapping keywords (*mixed*), ~2.1K *social* utterances, and ~12.9k *background* segments.

[1] Children played in pairs and 1-on-1 with a robot. We focus here on the child-child play because it is more variable, produces higher engagement scores, and is more difficult for the recognizer, and so offers a more rigorous test of online performance.

http://dx.doi.org/10.21437/WOCCI.2016-9

Figure 1: Two participants playing *Mole Madness* (left) and a screenshot from the game (right).

2. The Multi-keyword Spotter

Applications that use speech recognition are becoming more common due to the amount of data available and the maturation of machine learning approaches. However, existing systems tend to derive their acoustic models largely from adult speech [8, 9] and/or their language models from single-user search and scheduling tasks [10, 11]. As a result, current ASR tends to be poorly suited for young children, whose language is acoustically, lexically, syntactically, semantically, and pragmatically distinct [12, 13].

The ASR we've developed is an example-based keyword spotter that, at its core, handles an audio stream as the random output of a generative probabilistic model in which observed overlapping and non-overlapping keywords are the products of a latent mixture of Student's t-distributions (hereafter, TMM)[3].[2] The current spotter extends our earlier work in three ways [4]. First, we add separate models for *social* speech and *background* noise segments to the original *go*, *jump*, and combination (*mixed*) models. Second, we use a 150ms time slice rather than the original 300ms during training and at run time for better responsiveness.

The third extension is to replace the original threshold-based heuristic that was used to assign a final keyword label to the output posterior distribution of the TMM over the labels. Instead, we assign a label using a multi-class classifier (SVM) that takes as input the current time slice posteriors as well as the posteriors of four previous time slices. The SVM achieves greater overall accuracy than the spotter alone by exploiting localized temporal patterns in the training data. The solution as a whole - using the full context of the relative likelihood of the five different classes of speech over time - both regains context lost by switching to a shorter time slice and takes advantage of information present during social speech (allowing the system to differentiate between the imperative "go" and the "go" in "No, don't say go yet," for example).

At run time, a segment label of *go* or *jump* sends the corresponding command to the game; a segment label of *mixed* sends both commands.

3. Evaluation Metrics

Because our goal is to understand the spotter's performance in terms of the child's experience, we use a method of evaluation here, as in the original *post hoc* analysis, that focuses on the causal connection between words and action. Put concretely: if one or both keywords are spoken to the mole, the child should perceive the corresponding action(s) as the mole's response; if a keyword is spoken incidentally in a social context, or if no keyword has been spoken, then no corresponding change in the mole's behavior should be seen. Duration of the keyword is also considered meaningful. If a typical *go* or *jump* intends the corresponding action within a causally-meaningful period of saying it, then fast speech via elided forms ("g- g- g- go") is defined with respect to that expectation (i.e., faster speech should produce faster movement). Slow speech, on the other hand, appears to have two distinct meanings. Emphatic elongation ("gooo!") seems to ask for a single bigger movement (or possibly, a movement right away), while prolonged elongation ("gooooooooooooo") seems to ask for steady or on-going movement. It often has a sing-song quality and may simply be a way to vary speech when the vocal chords become tired.[3]

To evaluate the above phenomena we use the human-segmented and labeled data as ground truth. The human annotations give single-unit status to whole or partially formed keywords, entire utterances of social speech, and (by default) uninterrupted segments of silence, independent of their length.

We bridge the divide between recognizer output every 150ms and human judgments of semantically-meaningful units by aggregating across time slices with respect to the annotation. The window over which we aggregate reflects assumptions about how long a lag there can be between voicing the command and seeing the mole's behavior change before the child no longer experiences the two as causally connected. Choosing the right recognition window is non-trivial because it may depend on the particular activity or game [15], but evidence suggests a lag in the 300-700ms range is acceptable [16, 17]. Following the literature, we define the window over which recognizer output is aggregated to extend from the beginning of the annotation through three time slices (450ms) after the end of the annotation. This definition may give credit for detecting a keyword based on evidence that occurs after the annotation ends, but does so under the assumption that children would also attribute the mole's action to their utterance within that period.

[2] A phone-based approach to keyword spotting for children is described in [14], but does not deal with overlapping speech.

[3] We derive the practical definitions of fast, medium and slow speech from the distribution of keyword lengths in the SG data. In particular, we define fast and slow speech to correspond to a keyword with length that is less/more than half a standard deviation from the mean length, respectively.

To compute standard statistics, then, each ground truth keyword annotation counts as either a **false negative** (FN) or a **true positive** (TP). FN/TP is scored when there is no/at least one time slice with the keyword's label in the window. A time slice labeled as *mixed* represents both a *jump* and a *go* command for evaluation purposes.

The TP definition means that a single keyword annotation in the corpus "consumes" all the matching time slice detections within the window's bounds. An average length keyword (~300ms) is likely to cross multiple timeslices and generate one, occasionally two, commands per annotation. For fast (elided) keywords, the system's performance is bounded - even if a child can emit /g/ more quickly than every 150ms, the recognizer can detect and generate at most one command per time slice.

The TP definition also means that an annotation is counted as a true positive even if there is a time slice within the window's bounds that does not match. When the annotation is an elided or typical duration keyword, this makes sense because the existence of an intermixed *social* or *background* label has little effect on the mole's physics or what the child experiences. For a slow/elongated keyword, however, the TP definition biases the statistics in the recognizer's favor, potentially giving full credit to a five second *go* that has only an occasional correctly-labeled time slice in it even though the apparent behavior would not correspond to the steady movement that is intended. To remove this bias, we preprocess single slow speech annotations into separate consecutive 300ms keyword segments and apply the TP/FN definitions to each segment individually.

Social speech and, by default, non-speech segments with silence and/or background noises are the potential sources of **true negative** (TN) and **false positive** (FP) judgments. They are scored as TN if there is no time slice with a keyword label in the window. An FP is scored when a keyword label generated by the recognizer does not fall within the window of any keyword annotation (and so must be assigned to a non-keyword segment). An FP is also scored for any isolated (non-overlapping) keyword recognized as an instance of the other keyword.

4. Expected versus Autonomous Performance

A multi-keyword spotter built using all of the data from SG was integrated into *Mole Madness* and tested with 24 new children between the ages of four and nine as part of a multi-activity data collection in the winter of 2015 (WG). The WG cohort had a mean age of 6.7 years (standard deviation 1.6 years) and 58% were female. Subjects were compensated as in SG.

Gameplay was hand-segmented and labeled to produce ground truth annotations, with 67 minutes of play generating ~1250 non-overlapping instances of *go*, ~1900 non-overlapping instances of *jump*, ~1900 instances of overlapping keywords, ~630 social utterances, and ~2300 background segments. We compute performance statistics using the method described above, but with the actual timestamp of every *go* and *jump* executed by the game rather than *post hoc*, derived labels because the synchronized log data more accurately describes when the children saw the named action.

The purpose of the current study is to answer three questions, the first of which is: Are there significant differences between the performance of the deployed system and the expectations set by the *post hoc* analysis? The top of Table 1 contrasts the statistics in SG and WG. Neither precision nor accuracy is statistically different across the two corpora, but the difference in recall rate is significant (Student-t, p <.01). The remainder of the table entries break recall down with respect to keyword, overlap, and speech rate, showing that the overall difference in recall occurs primarily in the *jump* command and is not due to either overlap or fast speech.

Table 1: Mean (standard deviation) for expected (SG) and observed (WG) performance.

	SG	WG
Accuracy	0.83 (.10)	0.77 (.10)
Precision	0.83 (.10)	0.80 (.11)
Recall*	0.93 (.09)	0.83 (.10)
Go	0.94 (.10)	0.89 (.09)
Jump*	0.92 (.09)	0.81 (.13)
Not Overlapping*	0.93 (.08)	0.81 (.11)
Overlapping	0.93 (.10)	0.86 (.11)
Slow*	0.94 (.08)	0.83 (.10)
Medium*	0.94 (.08)	0.83 (.11)
Fast	0.88 (.14)	0.82 (.16)

* statistically significant at 5% level

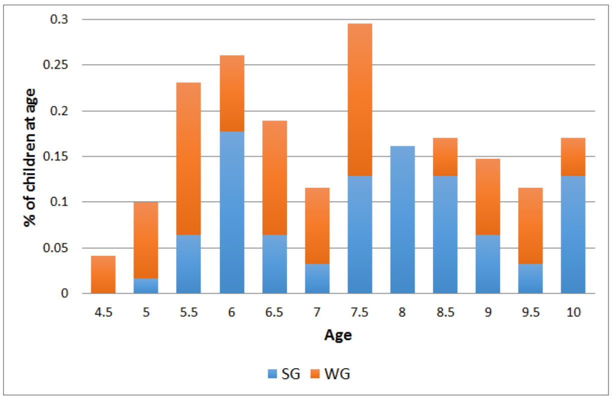

Figure 2: Age distributions of children in the training (SG) and test (WG) sets.

Further analysis suggests that the increase in recall error stems from the difference in the age distributions of children in the training (SG), and test (WG) sets (see Figure 2). The mean age of SG is almost half a year older than the mean age of WG, and only 8% of SG is under five and a half years (and none are under five) while 32% of WG is under 5 and a half years (and three children are under five). Every measure in Table 1 is significantly positively correlated with age in WG, and no measure correlates with age in SG. Concerning *jump* recall, in particular, we note that deaffrication is a type of disarticulation in which an affricate such as /j/ is replaced with a stop like /d/. Although the literature suggests that this phonological process is typically outgrown during the fourth year [18], we find that if we restrict the analysis to pairs of children in which both are older than five, the distinction between *go* and *jump* recall in WG disappears (the overall recall rate rises from 83% to 87% and accuracy rises from 77% to 80%). In short, the models of *jump* trained from the predominantly older children's speech in SG do more poorly with the younger speakers in WG.

Having found significant differences between expected and observed values, we proceed to the second question: Did the children change their behavior as a result? To answer this question, we must consider what the SG children experienced - the wizard's behavior - rather than what they would have experienced if the word spotter had been used. Table 2 repeats the WG statistics and contrasts them with the same statistics computed for the wizard's button presses given a 450ms lag and a 700ms lag (near the outer limit for causal connection suggested by the literature, but fairer to the wizard, who had a mean reaction time of 529ms (standard deviation 419ms)). With either lag, the children's experience with the wizard was consistent with respect to overlap and speech speed. At 700ms the wizard is also consistent with respect to keyword, while at 450ms she does better on *jump* than *go*. The word spotter is also consistent with respect to speed, but does significantly worse for *jump* than *go* and worse for non-overlap than overlap.

Against this backdrop, we examine the prevalence of each keyword in the two corpora and again find significant differences. WG has a higher percentage of *jumps* (59% versus 48%, Student-t, p <.01) as well as higher percentages of slow and medium speech (Student-t, p <.03 and p <.05, respectively). Taken with the discussion above, these patterns suggest that with the autonomous system the younger children had to say *jump* more carefully and more often to get the mole to move. This interpretation also explains why the drop in fast speech recall that was expected, based on the *post hoc* performance of the spotter in SG, did not materialize in WG.

Table 2: Recognition as experienced in WG and in SG with the wizard (Wiz), assuming 450ms or 700ms lag.

	WG (450)	Wiz (450)	Wiz (700)
Accuracy	0.77 (.10)	0.67 (.19)	0.74 (.17)
Precision	0.80 (.11)	0.74 (.19)	0.80 (.16)
Recall	0.83 (.10)	0.69 (.22)	0.76 (.20)
Go	0.89 (.09)	0.66 (.25)	0.74 (.22)
Jump	0.81 (.13)	0.71 (.22)	0.77 (.19)
Not Overlapping	0.81 (.11)	0.68 (.23)	0.75 (.20)
Overlapping	0.86 (.11)	0.69 (.21)	0.77 (.19)
Slow	0.83 (.10)	0.65 (.24)	0.73 (.20)
Medium	0.83 (.11)	0.68 (.22)	0.75 (.19)
Fast	0.82 (.16)	0.71 (.24)	0.78 (.23)

Having found both differences from predicted values and evidence that at least some of the children changed their behavior, we turn to the final question: Did the changes impact their enjoyment of the interaction? The method used for evaluating enjoyment has been described in detail in [5]. In recap: three mothers of young children coded the video of each player in a pilot study using a seven point proxy scale for engagement, with labeled values at 1 (*ready to do something else*), 3 (*could take it or leave it*), 5 (*very much into the game*) and 7 (*can't drag him/her away*) and unlabeled values at 2, 4, and 6.

The same coders and scale were used to evaluate SG and WG. Because there were systematic differences between the coders in the pilot based on how important they found verbal or visual cues, we compute the mean enjoyment score for each child separately for each coder. We find that mean player enjoyment (or, at least, willingness to remain at play) is uncorrelated with age in both SG and WG, and is significantly higher in WG for all three coders (p <.01). The average mean across all players in the Summer Games was 3.64, 3.88, and 3.67 (coders 1, 2, and 3, respectively), while the average mean across all players for Winter Games was 4.68, 5.06 and 4.91. In other words, on average SG players are judged to be less than halfway between *could take it or leave it* and *very much into the game*, while WG players are judged to be solidly enjoying the gameplay. This result is not surprising given the values in Table 2. We assume that children will be variable in how much lag they tolerate, but will need action to occur within 700ms of voicing a command to connect their words to the mole's movements. At shorter lags (e.g., 450ms) more children should connect the events; such children will experience better performance from the spotter than the wizard. At longer lags (e.g., 700ms), the wizard's responses will seem more accurate than they did at 450ms, but fewer children will experience them as causally connected. Accuracy at 700ms for the wizard is about what it is at 450ms for the keyword spotter, but recall - which is necessary to move the mole - is still significantly poorer in all but fast speech.

5. Conclusions and Future Work

A multi-keyword spotter to be used in a fast-paced game was trained on data from 62 children, and expectations for online performance established by 10-fold cross-validation. When the system was tested online, recall dropped significantly from the level predicted by the *post hoc* analysis, due primarily to differences in the age distributions of the two cohorts. Although the online system had trouble understanding *jump* with younger children, they adjusted their behavior to compensate, and the overall performance was both accurate enough and responsive enough to produce an enjoyable interaction. In future work, we intend to combine the SG and WG corpora and contrast performance under models built from all the data to performance under models built separately for younger and older children in our age range.

Other directions for future work focus on understanding both the limitations and broader applicability of our approach. *Mole Madness* was intentionally designed as a well-constrained point in the space of language-based games for children. Overlapping speech in MM contains at most two voices, has only two keywords, and uses keywords chosen to have no common phonemes and approximately the same mean duration under normal conditions. By relaxing each of these constraints systematically, we can gradually open the design space to new and substantively different interactions. Doing so creates the possibility of studying both more difficult instances of the linguistic issues that motivated this work and new linguistic phenomena that will undoubtedly arise.

6. Acknowledgments

The authors would like to thank Dr. Alex Rudnicky for his keen observations during the preparation of this manuscript.

7. References

[1] J. F. Lehman, "Robo fashion world: A multimodal corpus of multi-child human-computer interaction," in *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, ser. UM3I '14. New York, NY, USA: ACM, 2014, pp. 15–20.

[2] W. Y. Wang, S. Finkelstein, A. Ogan, A. W. Black, and J. Cassell, ""love ya, jerkface": Using sparse log-linear models to build positive (and impolite) relationships with teens," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, ser. SIGDIAL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 20–29.

[3] H. Sundar, J. F. Lehman, and R. Singh, “Keyword spotting in Multi-Player voice driven games for children,” in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, 2015.

[4] J. F. Lehman, N. Wolfe, and A. Pereira, “Multi-party language interaction in a fast-paced game using multi-keyword spotting,” in *Proceedings of the 16th International Conference on Intelligent Virtual Agents (IVA)*. Springer, 2016.

[5] S. Al Moubayed and J. F. Lehman, “Toward better understanding of engagement in multiparty spoken interaction with children,” in *Proceedings of the 17th ACM International Conference on Multimodal Interaction*, ser. ICMI. New York, NY, USA: ACM, 2015, pp. 211–218.

[6] J. F. Lehman and S. Al Moubayed, “Mole madness–a multi-child, fast-paced, speech-controlled game,” in *Proceedings of AAAI Symposium on Turn-taking and Coordination in Human-Machine Interaction. Stanford, CA*, 2015.

[7] M. Gerosa, D. Giuliani, and F. Brugnara, “Acoustic variability and automatic recognition of childrens speech,” *Speech Communication*, vol. 49, no. 10, pp. 847–860, 2007.

[8] S. Fernández, A. Graves, and J. Schmidhuber, “An application of recurrent neural networks to discriminative keyword spotting,” in *Proceedings of the International Conference on Artificial Neural Networks*. Springer, 2007, pp. 220–229.

[9] Y. Zhang and J. R. Glass, “Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams,” in *IEEE Workshop on Automatic Speech Recognition & Understanding. ASRU 2009.* IEEE, 2009, pp. 398–403.

[10] J. R. Bellegarda, “Spoken language understanding for natural interaction: The siri experience,” in *Natural Interaction with Robots, Knowbots and Smartphones*. Springer New York, 2014, pp. 3–14.

[11] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, “Large vocabulary automatic speech recognition for children,” in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, 2015.

[12] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, “A review of ASR technologies for children’s speech,” in *Proceedings of the 2nd Workshop on Child, Computer and Interaction (WOCCI ’09).* New York, NY, USA: ACM, 2009, pp. 7:1–7:8.

[13] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, “Improving Speech Recognition for Children using Acoustic Adaptation and Pronunciation Modeling,” *Proceedings of Workshop on Child Computer Interaction*, Sep. 2014.

[14] M. Wöllmer, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Tandem decoding of children’s speech for keyword detection in a child-robot interaction scenario,” *ACM Trans. Speech Lang. Process.*, vol. 7, no. 4, pp. 12:1–12:22, Aug. 2011.

[15] M. Claypool and K. Claypool, “Latency and player actions in online games,” *Communications of the ACM*, vol. 49, no. 11, pp. 40–45, Nov. 2006.

[16] J. Edlund, F. Edelstam, and J. Gustafson, “Human pause and resume behaviours for unobtrusive humanlike in-car spoken dialogue systems,” *EACL 2014*, p. 73, 2014.

[17] J. A. Dewey and T. H. Carr, “When dyads act in parallel, a sense of agency for the auditory consequences depends on the order of the actions,” *Conscious. Cogn.*, vol. 22, no. 1, pp. 155–166, Mar. 2013.

[18] S. McLeod and K. Bleile, “Neurological and developmental foundations of speech acquisition,” in *Proceedings of the American Speech-Language-Hearing Association Convention*. ASHA, 2003.