Type	Query #1		Query #2		
	Mean	SD	Mean	SD	Δ_p%
Total Length	1.14	0.61	1.45	0.63	+42%†
Total Loudness	0.05	0.07	0.06	0.07	+80%†
Silent Length	0.31	0.36	0.33	0.31	–
Consonant Length	0.10	0.06	0.12	0.06	+46%†
Vowel Length	0.11	0.09	0.14	0.08	+54%†
Consonant Loudness	0.031	0.044	0.035	0.041	+92%
Vowel Loudness	0.07	0.09	0.09	0.11	+124%†

Table 4: Acoustic reformulation statistics on DS8K. Length units are in seconds; loudness is dimensionless. Δ_p% denotes relative changes in the means of differences across paired observations. †Differences are significant ($p < 0.01$) according to the paired t-test.

mean total length and loudness of the audio are longer and louder, respectively, in the correct reformulations following the first set of incorrectly transcribed queries. This confirms previous findings [5, 6] on the acoustic qualities of voice query reformulations. The mean length of silence in the audio is not very different, implying that the increased sample length results from actual speech and not holding the microphone button for longer. Our phonetic-level statistics demonstrate increased length and loudness in reformulations as well: on average, in response to a recognition error, users increase consonant and vowel lengths by 46% and 54%, respectively. The loudness of consonants and vowels also rises by a corresponding 92% and 124%—interestingly, users emphasize and elongate vowels more than consonants in their reformulations.

4.3 Qualitative Observations

Through transcribing thousands of actual user queries, we observe a multitude of acoustic and social phenomena that may explain the recognition errors and responses, many of which are unlikely to arise in a controlled laboratory setting, and a few of which are unique to the entertainment domain. They range from guttural, user-created noises, such as laughing during a television show, to social behavior, such as passing around the microphone. Due to the limited size of the dataset, we are unable to conduct significant quantitative analyses on more nuanced user behavior. Nevertheless, we provide qualitative observations to guide future studies.

Verbal Behavior. First, we observe verbal behavior consistent with our quantitative analyses. Upon seeing incorrect transcription feedback, many users increase the loudness of their voices and decrease the phonation rate, which is similar to real-life human behavior. Several users add pauses to their queries, often worsening the desired transcription, such as "Hallmark" being recognized as "Hall mark" in one query. As an expression of frustration, a few users also shout into the microphone, resulting in amplitude clipping of the audio. We further observe that many users tend to "play" with their voices, wildly varying the loudness, pitch, and length of syllables in a phrase. We note that these users are at least partially self-aware, with several instances of them normalizing their speech after a speech recognition error.

Non-Verbal Behavior. Next, we observe a variety of non-verbal sounds in erroneously transcribed queries. Yawning while speaking, laughing at a show—we posit that these all contribute to recognition errors by distorting the acoustic characteristics of speech. Unsurprisingly, eating while watching television is common, with several clips containing chewing and burping noises. Users end the behavior in their correctly transcribed responses, suggesting that they perceive it as detrimental to recognition accuracy.

Social Phenomena. Finally, since watching television is often an interactive experience, we observe complex multi-user behavior in errors and their responses. In a few query sessions, the initial user struggles with obtaining a correct transcription from the ASR system, before another person intervenes and elicits the correct transcription. In one exemplar session, a user struggles to pronounce the desired voice query, then we hear another person in the background teaching the user.

5 CONCLUSIONS AND FUTURE WORK

We present an analysis of speech recognition errors in emerging TV-based entertainment systems, analyzing phonetic, acoustic, and lexical features of user reformulations. To pave the way for future work, we report qualitative observations from transcribing thousands of voice queries. Potential extensions to this work include building a targeted dataset for quantifying these observations, as well as using them to improve existing ASR and NLP systems. For example, we can use these reformulation patterns as priors for the acoustic and language models—if we know the user speaks louder and slower, the model can appropriately compensate. These are all promising steps toward our ultimate goal of natural speech-based interactions with intelligent agents.

Acknowledgments. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

REFERENCES

[1] A. Black and K. Lenzo. 2001. Flite: A Small Fast Run-Time Synthesis Engine. In *4th ISCA Workshop on Speech Synthesis*.
[2] C.-C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani. 2018. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In *ICASSP*. 4774–4778.
[3] I. Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. In *SIGIR*. 35–44.
[4] J. Huang and E. Efthimiadis. 2009. Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs. In *CIKM*. 77–86.
[5] J. Jiang, A. Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. Kulkarni, and O. Khan. 2015. Automatic Online Evaluation of Intelligent Assistants. In *WWW*. 506–516.
[6] J. Jiang, W. Jeng, and D. He. 2013. How Do Users Respond to Voice Input Errors? Lexical and Phonetic Query Reformulation in Voice Search. In *SIGIR*. 143–152.
[7] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech*. 498–502.
[8] J. Rao, F. Ture, H. He, O. Jojic, and J. Lin. 2017. Talking to Your TV: Context-Aware Voice Search with Hierarchical Recurrent Neural Networks. In *CIKM*. 557–566.
[9] J. Rao, F. Ture, and J. Lin. 2018. Multi-Task Learning with Neural Networks for Voice Query Understanding on an Entertainment Platform. In *KDD*. 636–645.
[10] J. Rao, F. Ture, and J. Lin. 2018. What Do Viewers Say to Their TVs? An Analysis of Voice Queries to Entertainment Systems. In *SIGIR*. 1213–1216.
[11] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope. 2010. "Your Word is My Command": Google Search by Voice: A Case Study. In *Advances in Speech Recognition*. Springer, 61–90.
[12] M. Shokouhi, R. Jones, U. Ozertem, K. Raghunathan, and F. Diaz. 2014. Mobile Query Reformulations. In *SIGIR*. 1011–1014.
[13] Y.-Y. Wang, D. Yu, Y.-C. Ju, and A. Acero. 2008. An Introduction to Voice Search. *IEEE Signal Processing Magazine* 25, 3 (2008).
[14] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. 2016. Achieving Human Parity in Conversational Speech Recognition. *arXiv:1610.05256*.