Yelling at Your TV: An Analysis of Speech Recognition Errors and Subsequent User Behavior on Entertainment Systems

Raphael Tang,[1,2] Ferhan Ture,[1] and Jimmy Lin[2]
[1] Comcast Applied AI Research Lab
[2] David R. Cheriton School of Computer Science, University of Waterloo

ABSTRACT

Millions of consumers issue voice queries through television-based entertainment systems such as the Comcast X1, the Amazon Fire TV, and Roku TV. Automatic speech recognition (ASR) systems are responsible for transcribing these voice queries into text to feed downstream natural language understanding modules. However, ASR is far from perfect, often producing incorrect transcriptions and forcing users to take corrective action. To better understand their impact on sessions, this paper characterizes speech recognition errors as well as subsequent user responses. We provide both quantitative and qualitative analyses, examining the acoustic as well as lexical attributes of the utterances. This work represents, to our knowledge, the first analysis of speech recognition errors from real users on a widely-deployed entertainment system.

ACM Reference Format:
Raphael Tang, Ferhan Ture, and Jimmy Lin. 2019. Yelling at Your TV: An Analysis of Speech Recognition Errors and Subsequent User Behavior on Entertainment Systems. In *42nd Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331271

1 INTRODUCTION

Increasingly popular are TV-based entertainment systems such as the Comcast X1, the Amazon Fire TV, and Roku TV—three platforms whose collective subscriber counts exceed 80 million, based on conservative figures derived from reports on the companies' websites. A speech-enabled remote controller is an important component of these intelligent systems, allowing viewers to conveniently perform voice search over shows, channels, and live events. For example, in response to the spoken query "The Sopranos", the TV immediately switches to the desired show. Some entertainment platforms also support more complex queries, such as "Show me all Scorsese movies with Joe Pesci". Speech input enables long, freeform queries, which are too cumbersome to type on a keypad.

Automatic speech recognition (ASR) systems transcribe these voice queries into text, over which rule- and deep learning-based models [8] can be applied. Unfortunately, ASR systems frequently produce incorrect transcriptions and corrupt the original queries, forcing users to reformulate their requests or surrender entirely. This is especially harmful in our domain, since voice input is the only tractable method for entering longer queries. In the mobile web search literature, there exists a large body of work that characterizes these recognition errors and subsequent user behavior [5, 6, 12]. However, as far as we are aware, this topic remains unexplored for voice queries to entertainment systems.

Why is it important to specifically study this vertical? First, our domain is unique and specific in content, revolving around the needs of TV viewers, instead of the broad information needs of general web users. Second, entertainment systems differ greatly from mobile devices in input modality, with users typically sitting down and issuing queries far from the feedback source. We believe that voice interactions in our context differ in key ways from personal assistants, smart speakers, and other voice-enabled devices.

This paper presents an analysis of speech recognition errors and subsequent behavior on the Comcast Xfinity X1 entertainment platform based on voice queries from real users. We conduct a quantitative analysis of speech recognition errors and subsequent user reformulations, examining both acoustic and lexical features. To shine light on the more nuanced verbal, non-verbal, and social behavior of viewers, we report qualitative observations as well. We are, to the best of our knowledge, the first to conduct an analysis of this kind for voice-enabled entertainment systems.

2 BACKGROUND AND RELATED WORK

Previous studies on mobile web search suggest that voice input represents a paradigm shift from text input, concluding that it is not merely text input enabled by speech. Depending on the platform, voice queries are either shorter or longer than text queries [3, 11]; they appear more natural than text queries [3]; voice query users tend to stick to voice input when reformulating queries (and not switch to text input) [12]. Furthermore, voice input technology is dogged by speaker variability and environment noise [13], issues not present with text input.

Commercial voice-enabled entertainment systems are backed by ASR systems which, despite significant advances in deep learning-based acoustic and language models, remain imperfect. Xiong et al. [14] are the first to report achieving human-level accuracy on the NIST 2000 speech recognition task, with a word-error rate (WER) of 5.9%. Recently, Chiu et al. [2] successfully apply sequence-to-sequence neural models to their 12,500-hour Google voice search task, achieving a state-of-the-art WER of 5.6%. While certainly impressive, these WERs are still much greater than zero, resulting in higher whole query-error rates when compounded across multiple words in a query. Rao et al. [8–10] describe several ASR and query understanding challenges in their work on the Comcast X1 entertainment system.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331271

Thus, characterizing speech recognition errors is an important task, supported by a plethora of studies. In a comprehensive, controlled study, Jiang et al. [6] examine the query reformulation patterns of a group of participants when they encounter recognition and system errors on Google. The researchers fix the information need to selected TREC topics, finding that both speech recognition errors and system interruptions have significant deleterious effects on ranking quality. In contrast, we collect and listen to actual user queries on our platform, since behavior such as shouting may be absent in a controlled, monitored setting. In a follow-up study, Jiang et al. [5] use acoustic and lexical features to automatically evaluate speech recognition and intent classification quality. Users tend to emphasize speech [6] and reduce the speaking rate [5] in their responses to recognition errors; however, a more accurate phonetic-level analysis has not been performed.

3 METHODS

Users of the Comcast Xfinity X1 product interact with the entertainment platform using a voice-enabled remote controller and set-top box, which displays feedback on the television. Comcast has delivered more than 20 million voice-enabled remotes to customers across the United States, processing more than 9 billion voice commands in 2018. To issue a voice query, the user depresses a microphone button, dictates a command, and releases the button, all the while receiving feedback through the television from our streaming ASR system. Most of the queries can be classified as "view" or "browse" intents, where the user desires to watch a certain channel, program, or live event. However, the X1 supports a broad range of additional functionalities—see Rao et al. [10] for a taxonomy of intents we've previously developed.

As expected, viewers watch television for extended periods of time, punctuated by intent switches. Thus, following the same procedure as Rao et al. [10], we use these watch events as delimiters for sessionizing queries, where a set of time-ordered voice queries is defined as a *session* if it satisfies these conditions:

(1) Each query is issued by the same device.
(2) Each non-first query occurs within 45 seconds of the last.
(3) There exists a watch event at the end, of at least 150 seconds in duration, within 30 seconds of the previous query.

To construct our datasets, a team of four annotators listened to and annotated thousands of voice queries from the week of January 2–9, 2019. Note that we collect and use all data in accordance with our Privacy Notice; we do not store customer- and household-identifiable information with voice recordings or transcriptions. One annotator first transcribed all the queries; then, the other three took turns verifying all of the annotations. Upon disagreeing, they discussed the conflict to arrive at a unanimous decision. This annotation process resembles Jiang et al. [6].

Our data consists entirely of sessions, with all other queries discarded (e.g., those with no final watch event), since they are not relevant to answer our session-based research questions. To impose a reasonable limit on the session length, we filter out all sessions with more than five queries, which contribute to less than 0.2% of all sessions. From this pool of sessions, we then sample from sessions of lengths one to five to construct a session-oriented dataset. This dataset contains 1012, 517, 346, 262, and 199 sessions with one to five queries, respectively, for a total of 5127 queries, of which 2250 are unique. We name this dataset DS5Q.

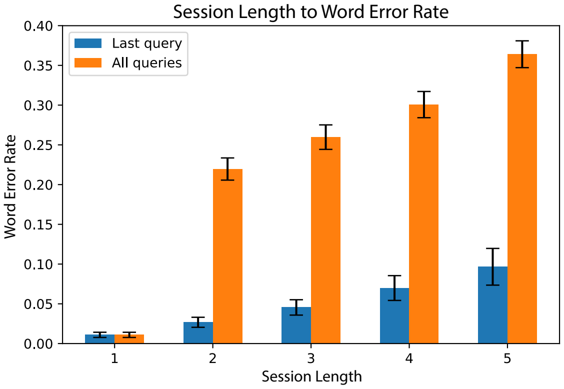

Figure 1: WER along with 95% CIs for each set of sessions in DS5Q, grouped by the number of queries. "Last query" denotes statistics of the last query of each group.

For our second dataset, to conduct a targeted, acoustic analysis of our most popular keywords, we focus on two-query sessions that satisfy these three conditions: First, both voice queries must have the same true transcription. Second, they are one of these eight keywords: "BET", "CNN", "Disney Junior", "Fox News", "Lifetime", "Netflix", "Nickelodeon", and "YouTube". Third, the first query is incorrectly transcribed by the ASR system, while the second is correct. Although limited to eight keywords, this dataset represents more than 10% of our total voice traffic. It should admit lower acoustic sample variance than DS5Q in erroneous queries and their responses, having fixed the view intent and transcription. The dataset, which we call DS8K, contains 54 instances of "BET", 30 of "CNN", 31 of "Disney Junior", 19 of "Fox News", 25 of "Lifetime", 45 of "Netflix", 13 of "Nickelodeon", and 33 of "YouTube", for a total of 250 queries.

4 ANALYSIS

Closely following Jiang et al. [6], we present analyses of speech recognition errors and user responses, guided by four research questions: When do speech recognition errors occur, and what effects do recognition errors have on the transcription? These seek to characterize the nature of recognition errors in our session-oriented dataset, DS5Q. Next, how do users lexically reformulate their queries in response to errors? We conduct this analysis on the sessionized DS5Q. Finally, how do users acoustically reformulate their queries in response to errors? We conduct this analysis on the keyword-oriented DS8K.

4.1 Recognition Error Analysis

RQ1: When do recognition errors occur? In DS5Q, we find that longer sessions are plagued with higher word error rates (WERs); see Figure 1. "Watchthrough", analogous to clickthrough in our domain, is a highly effective form of implicit feedback for the transcription quality, with the last query of each session achieving much lower WER than the overall session average. Single-query sessions in particular have an extremely low WER of 0.01, yielding

Measure	No ASR Errors		With ASR Errors	
	Mean	SD	Mean	SD
# queries	3757	–	1370	–
Session len.	1.65	0.99	1.72	0.97
Query word len.	1.96	1.15	2.30	1.57
Query char. len.	11.1	7.11	12.4	8.57

Table 1: Query and session length statistics on DS5Q.

"free", high-quality labeled audio for training voice query recognition systems. Beyond single-query sessions, the WER increases dramatically to 0.22 for two-query sessions, then gradually rises to 0.36 for five-query sessions. We observe that users repeat incorrect queries to elicit a correct transcription, hence increasing the session length. Longer sessions and queries are associated with ASR errors (see Table 1), corroborating the findings of Jiang et al. [6].

RQ2: How do recognition errors affect queries? To characterize how speech recognition errors change the system transcription, we modify the high-precision taxonomy for query reformulation introduced in Huang and Efthimis [4]. Although they originally developed the ruleset for query reformulation, we can similarly view the speech recognition system as an imperfect agent that may perturb the original query, leading to an artificial reformulation. We select the rules AddWords, RemoveWords, Stemming, Substring, Superstring, and New, denoting whole word addition and removal, morphological stemming, substring and superstring operations. All other non-matching queries are presumed to have a new intent. Following Huang and Efthimis, classification is accomplished by a cascading sequence of rules, in the order referenced above. We also add two rules of our own, specifically tailored for speech recognition errors on entertainment systems:

- ChannelAcronymError: the incorrect transcription is a channel acronym, e.g., "CNN", and the true query is any single word. This error is particularly egregious, since the TV may tune to a different channel altogether.
- PhoneticConfusion: the incorrect transcription is within a phonetic edit distance of one to the correct transcription. This error suggests an acoustic and language modeling issue, e.g., "Allen Show" is extremely similar to "Ellen Show".

We apply ChannelAcronymError as the first rule, and PhoneticConfusion as the last rule before New, the catch-all category. For our grapheme-to-phoneme model, we use the CMU Flite [1] synthesis engine. We model PhoneticConfusion after Metaphone edit distance, as in Jiang et al. [5], where it is also used to measure phonetic similarity for evaluating speech recognition quality.

The results of applying these rules to DS5Q are shown in Table 2. We see that most transcription errors (63.9%) result in a New query entirely. The recognition errors strongly tend to remove words (14%) instead of adding words (0.9%); likewise, Substring (2.3%) is more frequently applied than Superstring (0.5%). PhoneticConfusion (7.6%) is the third most common error type, followed by ChannelAcronymError (5.8%), where many viewers erroneously switch away from the current channel. We observe that Stemming (5.0%) is relatively benign, with most errors resulting from incorrect (lack of) pluralization.

Type	%	Example (right ↦ wrong)
New	63.9%	Starz Encore ↦ Start Uncle
RemoveWords	14.0%	Let's go with it ↦ Let's go
PhoneticConfusion	7.6%	The Mag ↦ The Meg
ChannelAcronymError	5.8%	KITE ↦ KITV
Stemming	5.0%	Rockets ↦ Rocket
Substring	2.3%	Total DramaRama ↦ Total Drama
AddWords	0.9%	Bug's Life ↦ A Bug's Life
Superstring	0.5%	Tube ↦ YouTube

Table 2: Distribution of recognition error types on DS5Q.

Type	No ASR Errors	With ASR Errors	
	%	%	Absolute Δ%
New	62.6%	45.4%	−17.2%
Same	17.2%	36.4%	+19.2%
AddWords	11.0%	8.1%	−2.9%
RemoveWords	5.2%	5.2%	0%
WordSubstitute	1.4%	1.5%	+0.1%
Superstring	1.3%	1.6%	+0.3%
Stemming	0.7%	1.1%	+0.4%
Substring	0.6%	0.7%	+0.1%

Table 3: Lexical reformulation statistics on DS5Q.

4.2 Reformulation Analysis

RQ3: How are queries lexically reformulated? We also apply the reformulation rules from Huang and Efthimis [4], excluding abbreviation, to classify query reformulations between each pair of consecutive queries on DS5Q. These results are presented in Table 3, split on the presence or absence of recognition errors. Our analyses are, of course, limited to sessions with at least two queries; single-query sessions contain no reformulations. Faced with an ASR error, users simply repeat the same query 36.4% of the time. On the other hand, if no ASR error is present, the same query is repeated 17.2% of the time. Compared to Jiang et al. [6], we note that our users are much more likely to issue the same query, possibly due to the lack of reformulation freedom for entity names, e.g., movie titles and channel names. The causes of repeated queries in cases without ASR errors include system unresponsiveness and the interleaving of voice and keypad input. We see that users are less likely to add words after a recognition error (11% vs. 8.1%). The other reformulation types see only small changes in distribution.

RQ4: How are queries acoustically reformulated? We analyze phonetic-level loudness and length qualities of the utterances in DS8K. To extract phoneme boundaries, we run the popular Montreal Forced Aligner [7] on the dataset. Then, we compute loudness and length statistics over each phonetic segment, where the loudness heuristic is defined as the standard A-weighted root mean square of the amplitude. We split the queries into two groups: those with recognition errors and those without, corresponding to the first and second query of each session.

We report acoustic reformulation statistics in Table 4. From the construction of DS8K, the first query contains an ASR error, while the second query is correctly transcribed (see Section 3). The

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.*