Harnessing Imperfections in Continuous Speech Recognition for Creative Design Sessions

Salvatore Andolina
University of Helsinki
Helsinki, Finland
salvatore.andolina@helsinki.fi

Antti Jylhä
University of Helsinki
Helsinki, Finland
antti.jylha@helsinki.fi

Khalil Klouche
University of Helsinki
Helsinki, Finland
khalil.klouche@helsinki.fi

Giulio Jacucci
University of Helsinki
Helsinki, Finland
giulio.jacucci@helsinki.fi

CHI'16, May 7–8, 2016, San Jose, CA, USA. Workshop on Designing Speech and Language Interaction for Mobile and Wearable Computing (DSLI 2016),

Abstract
Continuous speech stream has recently been proposed as a rich source of information as a user input. However, errors in automatic speech recognition (ASR) are often considered an obstacle in broadly adopting these techniques for practical use. In this position paper, we argue that continuous speech stream can be used as a rich source of input, for example, in idea generation meetings. As a case, we present an unobtrusive interactive display called InspirationWall, which tracks the ongoing discussion and proactively proposes search terms related to the discussion topics. Through the case, we show that even imperfectly interpreted speech can enrichen and facilitate idea generation and bring new insights to the conversation. Based on this result, we discuss the possibility of using the principle of randomness in design for enrichening the interaction with ASR-based systems.

Author Keywords
Automatic speech recognition; creativity; serendipity; interaction design.

ACM Classification Keywords
H.5.2 [Information interfaces and presentation (e.g., HCI)]: User Interfaces—*Voice I/O*

Introduction

Collaborative group activities such as idea generation sessions usually comprise vivid discussion, encompassing themes and phrases relevant for the goal of the activity. In such settings, it is typical that the participants also momentarily engage in collaboratively searching for related information via Web-mediated resources [4]. A way to facilitate these collaborative search efforts would be to utilize an automatic speech recognition (ASR) engine to proactively mine search keywords from the group conversation.

In this position paper, we address the use of continuous speech stream as a resource in the social setting of creative idea generation and collaborative search. Through a case of InspirationWall, an unobtrusive display showing keywords tracked from the ongoing conversation, we discuss how even incorrectly recognized phrases can spice up and facilitate the group discussion and foster idea generation.

Fit with the Workshop Topic

The position paper fits perfectly within the scope of the workshop. The paper addresses a relatively new point of research, the continous speech stream as an input resource. The paper dwells in the intersection of speech recognition and interaction design, and, through an exemplary design case, envisions a new way of using ASR imperfections for interactive systems to foster social creativity. In the workshop, we hope to invoke discussion on the topic of randomness in design of ASR-based interactive systems.

Continuous Speech Stream as Input Resource

McMillan, Loriette, and Brown [7] recently proposed continuous speech stream (CSS) as a "a rich resource for identifying users' next actions, along with the interests and dispositions of those being recorded". They envisioned CSS as an input to a rich set of personal applications [7], such as pre-searching for information, memory aid, and tracking the activities of the user. This can be expanded also beyond the single user to groups having a conversation, for example, to proactively track and suggest search keywords and recurring topics. When the group conversation is monitored by ASR, it can be used to track potential keywords, which then can be displayed to the group to facilitate the search, or simply provide a visual record of spoken topics. It is of course likely that the ASR will make recognition errors, but it has previously been argued that imperfect accuracy should not stop designing speech-based interfaces [8] and that 100% accuracy is even unnecessary [9, 5, 3].

When CSS is considered as an input resource in a creative group setting, for example to track the conversation for proactively suggesting search keywords, the output of the recognition system can fall into three categories:

1. Correctly recognized words or phrases that are relevant for the ongoing discussion. This is when the ASR is working as expected and provides meaningful output.
2. Correctly recognized words or phrases that are irrelevant for the ongoing discussion; for example, a joke or unrelated comment by one of the participants. This is when the ASR is working correctly, but the tracked content is not necessarily relevant nor meaningful considering the discussion.
3. Incorrectly recognized words or phrases. This is when the ASR makes an error and provides unexpected output.

Category 1 keywords can clearly be directly useful in the group setting for suggesting relevant search keywords or

themes for discussion. Category 2 keywords are likely not useful as such and could even be filtered out by the system. Category 3 keywords are seemingly noise in the output; however, we argue that they can also be useful for creative collaboration to facilitate out-of-the box thinking and creation of new ideas through the principle of randomness in design [6]. In the following section, we demonstrate this through a case of a system designed to facilitate collaborative idea generation.

Harnessing Imperfections of ASR for Serendipitous Design: Case InspirationWall

Creative ideas are often triggered by unexpected associations. In collaborative settings these associations are facilitated by social interaction and knowledge sharing. However, in such situations certain social processes such as evaluation apprehension [2] may actually discourage people to propose the most divergent and unexpected associations. The solution to that problem may come from a technology that automatically generates those divergent ideas.

InspirationWall [1] is an unobtrusive display that listens to the conversation through continuous speech recognition and expands the currently discussed topic with keywords that relate to the conversation. Recognized sentences are processed by an entity-based keyword suggestion system. The returned keywords are then displayed as slowly crossing the screen from top to bottom to progressively refesh the displayed keywords and to provide a glanceable history of recent keywords.

The system was evaluated with six pairs of non-native English speakers from different countries and cultures, with a similar proficiency in oral English. Participants were given the task to generate ideas for possible student projects on a certain topic. Following a within-subjects design, we compared a traditional idea generation session with no technology involved with a session augmented by InspirationWall. Results showed that InspirationWall contrasted the decay of idea productivity over time, which is a typical phenomenon in idea generation sessions [1].

Qualitative analyses of the video data examined the effect of category 1 and category 3 keywords in the ideation process. Those keywords were not directly shown to participants, but were elaborated by the suggestion system to generate related concepts. For example, the category 1 keyword “sleep” led to concepts such as “health”, “time”, and “circadian rhythm”.

Category 1 keywords were generally appreciated by participants, as they expanded the current topic of the conversation with topics that diverged while still being related to the conversation. These keywords helped generating ideas while also provided users with confidence on the system and its capability to understand the conversation. However, the most unexpected associations seem to have been generated by category 3 keywords. The following transcript, from a brainstorming session on possible projects involving wearable computing, provides an example of interaction with a word generated from imperfect recognition:

Fragment 1	*Transcript 1*
P02:	Suicide <reading a keyword from the screen> (2.5)
P01:	Water generator:: <reading> Random sensory:: <reading>
P02:	A sensor that informs the police or something... <starts proposing an idea> ...I was just reading the word suicide <explains where the idea comes from>
P01:	I don’t think you can do much about it=

P02: (My idea it is that) it just informs the police or your relatives that you are going to do something crazy

P01: Maybe it could be useful for people with depression
(Writes down the idea)

The fragment shows how the word “suicide”, which was shown on the InspirationWall due to a recognition error, triggered some unexpected associations among participants. The idea generated then triggered another idea of a similar wearable technology, but used in a different context:

Fragment 2 *Transcript 2*

P01: Ok, this one is a bit crazy <introducing the idea>

P02: Ok.

P01: So: Every convicted person in a jail has some kind of sensor, and by checking the physiology of the person, the ambient noise, and so on, aggregates the data of everybody and compares it with the history, and it is able to predict when there is gonna be a riot, and allows the security people to know (for example when) there is 80% chance, or there is only 30% chance...

P02: I guess this can be related to this sensor

P01: <referring to the suicide prevention sensor just ideated>

P01: Yeah!
(Writes down the idea)

Transcript 2, together with transcript 1, exemplifies the case in which a single category 3 keyword generated a train of thoughts that allowed participants to become more fluent and to generate non-obvious ideas that were not considered in any of the other sessions of the experiment. Informal after-study interviews confirmed how both category 1 and category 3 keywords were considered important for the experience, with divided opinions about which category was the most useful.

Conclusion

Continuous speech stream can be used to infer context and facilitate creative activities. Recent technological advancements on ASR permit to obtain a level of accuracy never experienced before, enabling a range of applications limited only by imagination. However, the problem of recognition accuracy will not totally disappear due to several reasons including noise, mispronunciation, accent, etc. We propose a model in which a certain amount of recognition errors do not represent an issue, but instead contribute to provide a richer user experience. It is clear that the presented case alone is not sufficient for exhaustively presenting the potential of using ASR imperfections for positive results. Nevertheless, we argue that the case provides encouraging evidence for future studies on utilizing the principle of randomness in design [6] when designing ASR-based interactive systems.

In the presented case, tracked keywords were simply displayed back to the participants. However, the system could also easily be integrated as a front-end for a search engine. In this setting, it would be interesting to examine how ASR imperfections might prompt unexpected search attempts and results to further facilitate out-of-the-box idea generation.

REFERENCES

1. Salvatore Andolina, Khalil Klouche, Diogo Cabral, Tuukka Ruotsalo, and Giulio Jacucci. 2015. InspirationWall: Supporting Idea Generation Through Automatic Information Exploration. In *Proceedings of*

the 2015 ACM SIGCHI Conference on Creativity and Cognition (C&C '15). 103–106. DOI: http://dx.doi.org/10.1145/2757226.2757252

2. Michael Diehl and Wolfgang Strpebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology* (1987), 497–509.

3. Yashesh Gaur. 2015. The Effects of Automatic Speech Recognition Quality on Human Transcription Latency. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15)*. 367–368. DOI: http://dx.doi.org/10.1145/2700648.2811331

4. Marti A. Hearst. 2011. 'Natural' Search User Interfaces. *Commun. ACM* 54, 11 (Nov. 2011), 60–67. DOI: http://dx.doi.org/10.1145/2018396.2018414

5. Anuj Kumar, Tim Paek, and Bongshin Lee. 2012. Voice Typing: A New Speech Interaction Model for Dictation on Touchscreen Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. 2277–2286. DOI: http://dx.doi.org/10.1145/2207676.2208386

6. Tuck Wah Leong, Frank Vetere, and Steve Howard. 2006. Randomness As a Resource for Design. In *Proceedings of the 6th Conference on Designing Interactive Systems (DIS '06)*. 132–139. DOI: http://dx.doi.org/10.1145/1142405.1142428

7. Donald McMillan, Antoine Loriette, and Barry Brown. 2015. Repurposing Conversation: Experiments with the Continuous Speech Stream. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. 3953–3962. DOI: http://dx.doi.org/10.1145/2702123.2702532

8. Cosmin Munteanu, Matt Jones, Sharon Oviatt, Stephen Brewster, Gerald Penn, Steve Whittaker, Nitendra Rajput, and Amit Nanavati. 2013. We Need to Talk: HCI and the Delicate Topic of Spoken Language Interaction. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. 2459–2464. DOI: http://dx.doi.org/10.1145/2468356.2468803

9. Gerald Penn and Xiaodan Zhu. 2008. A Critical Reassessment of Evaluation Baselines for Speech Summarization.. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*. 470–478.