Speech-Based Situational Awareness for Crisis Response

Dmitri V. Kalashnikov[1], Dilek Hakkani-Tür[2], Gokhan Tur[3], Nalini Venkatasubramanian[2]

1. Introduction

The goal of our work is to explore research in the framework of an end-to-end speech processing system that can automatically process human conversations to create situational awareness during crisis response. Situational awareness refers to knowledge about the unfolding crisis event, the needs, the resources, and the context. Accurate assessment of the situation is vital to enable first responders (and the public) to take appropriate actions that can have significant impact on life and property. Consider, for instance, a situation of a large structural fire wherein teams of fire fighters enter into a burning building for search and rescue. Knowledge of the location of fire fighters, their physiological status, the ambient conditions and environment are critical for the success and safety of both the victims and fire fighters. Appropriate situational awareness is critical not just at incident level, but at all levels of response. For instance, knowledge of occupancy levels, the special needs of the populace, the road-closures, the geographical scope of the disaster (e.g., the fire perimeter), etc. play a vital role in evacuation and shelter planning and in organizing medical triage.

The importance of accurate and actionable situational awareness in crisis response is now well recognized [1] and has led to significant research on appropriate sensing, networking, sensor processing, information sharing, data management, and decision support tools. The research team at the UC Irvine, Center for Emergency Response Technologies (CERT) has been extensively involved in working with a variety of first responder organizations to build a variety of such tools in the context of the NSF funded RESCUE (Responding to Crises and Unexpected Events) project [2] and the DHS funded SAFIRE project [3]. Our experience working with rescue personnel has clearly established that while sensors (including motes, video, physiological, location, environmental) are important, **speech** is undoubtedly a single most important source of situational information. The very first point of contact of citizens with the responders during an emergency is through a telephone call to the 911 dispatch system. In case of large disasters that involve a larger team of responders (such as a fire-fighting team), the primary mechanism used for communication and coordination among response teams is through radios carried by the first responders. Such conversations contain perhaps what constitutes the most important situational information that has direct implications on the efficacy of the response. Despite importance of speech, today, assimilation of situational information from speech is almost entirely done manually.

2. Observational Speech System (OSS)

In our work, we envision a speech-based situation awareness system, that we refer to as *Observational Speech System* (OSS), that **observes** human-to-human communications, **understands** the context and content of such communications, **models** and **represents** the extracted content, and **enhances** and/or **augments** the corresponding human processes and interactions. Such enhancements could be in a variety of forms: e.g., as a log/record of dynamically evolving human communications that can be browsed and searched; as a mechanism to triage important information to others who may be potentially interested in the content in speech; and as a mechanism to create real-time situational awareness from dynamically captured speech.

We conjecture that OSS has the potential to dramatically influence crisis response domain bringing in fundamental changes, improvements, and new efficiencies many of which we cannot even conceive of today. We highlight the importance of such a system through two examples. First consider the example above of structural fire. Speech technologies coupled with emotion detection could lead to determination of

[1] University of California, Irvine; Computer Science Department.
[2] International Computer Science Institute (ICSI), Berkeley, CA.
[3] SRI International, Menlo Park, CA.

physiological health/status of the fire fighter (stress related heart attacks are known to be the leading cause of on-site deaths amongst fire fighters), it could lead to determination of location of fire fighters (enabling rapid intervention teams to reach injured fire fighters faster), improved coordination and new efficiencies in fire fighting practice (e.g., a fire fighter reporting that 'a corridor is now clear' if captured and triaged correctly may allow incident commanders to redirect time critical resources accordingly). As another example, consider OSS being used as the underlying technology for a next-generation emergency dispatch service. Information extracted from calls in queue (by allowing callers to interact with the system while waiting) can be used to synthesize key information such as emergency type and location to support informed decision-making by dispatchers, to enable appropriate call routing and call prioritization. It could also be used as an effective tool for caller initiated call resolution (e.g., in case of duplicate calls about incidents already reported and phantom calls). Such technologies can help alleviate vulnerability of chronically understaffed dispatch services by improving dispatcher productivity and reducing call volume. Indeed, such an enhanced 911 service was the motivation for our seed funding from NSF that enabled us to explore emotion and intent determination techniques for 911 calls – techniques we build upon in our work. There is a substantial opportunity for speech and spoken language processing technology to improve crisis response.

The past two decades of research on automated speech recognition and spoken language processing provides a solid foundation to build the envisioned observational speech system. The OSS system, extends beyond automated speech recognition (that simply transcribes "what is said") into a more contextual understanding and interpretation of speech input requiring techniques from natural language understanding, information extraction, and multi-modal reasoning -- each of which our team has significant expertise and prior work on. Building an envisioned end-to-end system that can serve as a powerful decision support tool in the crisis response context is not simply a matter of integrating such component technologies together. The challenges posed by the application/domain context as well as the end-to-end perspective require significant advances on each of the component technologies, as well as a deeper understanding of interdependencies and relationships between components that can lead to significantly novel research directions.

First and foremost, speech in crisis context is expected to be emotionally charged and excited. Moreover, the acoustic conditions of the environment could mismatch the existing acoustic models. Furthermore, the speakers' utterances may not be grammatical and may include large number of disfluencies, such as false starts, breaths, and repetitions. All of these make automated speech recognition and understanding significantly complex requiring new directions / approaches. Another major challenge is that automated speech processing systems are unlikely to achieve perfect performance at any time in the near future.

To address these challenges, in this project, the initial focus of research is analyzing the degree of the dependencies and investigating joint classification methods via multi-task learning for language understanding and emotion detection. Our initial findings from the previous seed NSF project support our expectation of improved capabilities in these difficult circumstances. In this data we have noticed that while emotional speech does not necessarily correspond to more urgent emergencies (as professional caregivers calling 911 tend to be unemotional even in most dire situations and vice versa), there is a correlation between unemotional speech and non-emergency intents.

Furthermore, principled exploitation of semantic information, including temporal or locative contextual constraints, and using them for robust spoken language processing offers substantial promise. To overcome challenges posed by imperfections in the input data, appropriate situational awareness and decision-making tools are needed. Techniques that can overcome the data quality challenge broadly belong to two different (complementary) types. First are robustness techniques that exploit variety of contextual and domain knowledge/semantics to mask errors and improve data quality. Such techniques have been explored in the literature for very different contexts (e.g., data integration, intent determination, entity resolution, and disambiguation, to name a few). Appropriately adapted and extended, such techniques could provide a powerful tool for addressing data quality challenges in the envisioned speech processing system. The second approach is to design applications / data analysis techniques that can tolerate errors in data. Such approaches attempt to optimize end-goals of the application by providing the best answers given underlying

uncertainty in the data, thereby realizing the trade offs between specific quality needs of applications with the degree of uncertainty. As a concrete example, a situational awareness technology for triaging / triggering (at the keyword level) could be designed even with relatively inaccurate recognition quality simply by indexing words offered as alternatives by the speech recognizer. Of course, such a triaging system in an attempt to reduce false negatives may introduce additional false positives but such a trade off might be acceptable if the application semantics dictates that missing triggers is much less desirable compared to the information overload that may result due to false positives.

2. OSS Architecture

Our proposed research is driven by a new vision for observational speech system (OSS) that dynamically captures and processes speech input conversations amongst humans to extract situational information in the context of real-world activities (such as crisis response) thereby creating awareness. Figure 1 illustrates the primary components of the envisioned system. In OSS, dynamically captured speech input is first processed by an automatic speech recognizer (ASR) that converts the audio signal into a sequence of words (or a word lattice) possibly uttered by the speaker. The text output of ASR along with related meta-information such as speaker identity, speaker location, and identity of receiver(s), form a raw level speech database that itself can be useful in supporting a variety of awareness applications. ASR output can further be processed by two tightly coupled spoken language processing components, namely emotion and intent/incident detection to form a more semantically enriched representation that associates semantic annotations in the form of intent and emotions with speech segments. Finally, information extraction techniques applied to both spoken words as well as N-best list output of speech processing, create a situational representation that enables awareness applications. On top of these components, robustness techniques that exploit variety of semantic reasoning to improve data quality are applied on the stored information. Decision support tools such as retrieval and triggering mechanisms are designed to handle errors/uncertainty in situational data.

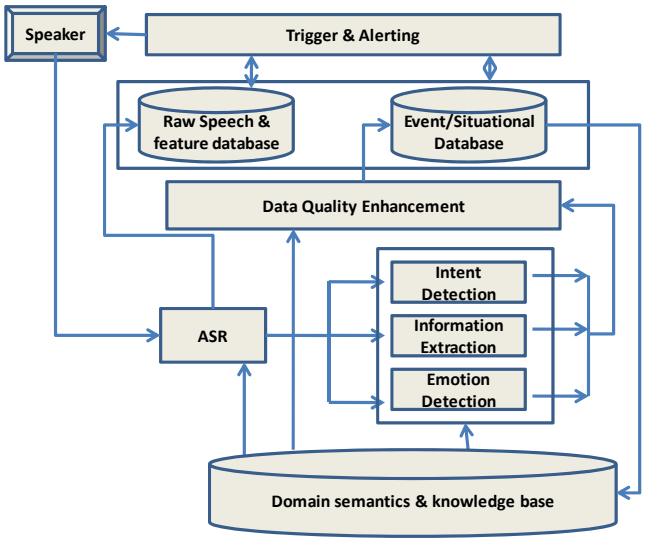

Figure 1. Components of the envisioned Observational Speech System (OSS)

Conclusion

To the best of our knowledge, this is a pioneering effort focusing on speech-based situational awareness, which, as envisioned, poses a significant research challenge, especially for the crisis response domain. If successful, this has the potential to very significantly improve the crisis response process. This type of semantically constrained joint processing aiming to be robust to ASR errors can open the way towards significantly improved emergency capabilities, and can also suggest methods for pursuing scientific investigation of speech and language in other challenging circumstances. Several communities should benefit from this research: 1) First-responder agencies (such as the Orange County Fire Authority (OCFA)) will get first-hand evidence and experience on using high fidelity human information in pursuing their mission. 2) Speech and language processing researchers will benefit from the speech corpus with corresponding meta information, test beds, and the prototype framework built in our effort. 3) Students at both UC-Irvine and UC-Berkeley would benefit from the enhanced educational experience and the excitement of participating in research that has potential for such direct impact on the real world.

References

[1] Sharad Mehrotra, Taieb Znatil, and Craig W. Thompson, "Crisis Management.," IEEE Internet Computing, vol. 12, no. 1, pp. 14-17, 2008.

[2] Sharad Mehrotra. Project RESCUE. [Online]. HYPERLINK "http://www.itr-rescue.org" http://www.itr-rescue.org

[3] Sharad Mehrotra. SAFIRE Project. [Online]. HYPERLINK "http://www.ics.uci.edu/~cert/safire" http://www.ics.uci.edu/~cert/safire

[4] M. Pucher, Semantic Similarity in Automatic Speech Recognition for Meetings. Wien, Austria: PhD Dissertation, Graz Technical University, 2007.

[5] Naveen Ashish and Sharad Mehrotra, "XAR: An Integrated Framework for Semantic Extraction and Annotation Cases in Semantic Interoperability for Information Systems Integration," in IGI Global., 2009.

[6] Rabia Nuray-Turan, Zhaoqi Chen, Dmitri V. Kalashnikov, and Sharad Mehrotra, "Exploiting Web querying for Web People Search in WePS2.," in In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, Madrid, 2009.

[7] N. Gupta et al., "The AT&T Spoken Language Understanding System," IEEE Transactions on Speech and Audio Processing, vol. 14, no. 1, 2006.

[8] Z. J. Chuang and C. H. Wu, "Emotion recognition from textual input using an emotional semantic network," in Proceedings of the ICSLP, 2002, 2002.

[9] Dmitri V. Kalashnikov, Zhaoqi Chen, Rabia Nuray-Turan, and Sharad Mehrotra, "Web people search via connection analysis.," IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE), vol. 20, no. 11, 2008.

[10] R. Vaisenberg, S. Mehrotra, and D. Ramanan, "Exploiting Semantics For Scheduling Data Collection From Sensors On Real-Time To Maximize Event Detection," in Multimedia and Computer Networks (MMCN), 2009.

[11] R. Vaisenberg, S. Ji, B. Hore, S. Mehrotra, and N. Venkatasubramanian, "Exploiting Semantics for Sensor Recalibration in Event Detection System," in Multimedia and Computer Networks, 2008.

[12] Daniel Massaguer, Bijit Hore, Mamadou H. Diallo, Sharad Mehrotra, and Nalini Venkatasubramanian, "Middleware for Pervasive Spaces: Balancing Privacy and Utility," in ACM/IFIP/USENIX Middleware 2009, 2009.

[13] Iosif Lazaridis and Sharad Mehrotra, "Optimization of multi-version expensive predicates," in SIGMOD, 2007.

[14] Iosif Lazaridis and Sharad Mehrotra, "Approximate Selection Queries over Imprecise Data," in ICDE, 2004.

[15] Dmitri V. Kalashnikov, Yiming Ma, Sharad Mehrotra, and Ram Hariharan, "Index for Fast Retrieval of Uncertain Spatial Point Data," in Proc. of Int'l Symposium on Advances in Geographic Information Systems (ACM GIS 2006), 2006.

[16] Nilesh Dalvi and Dan Suciu, "Efficient query evaluation on probabilistic databases," in VLDB, 2004.

[17] Lyublena Antova, Thomas Jansen, Christoph Koch, and Dan Olteanu, "Fast and Simple Relational Processing of Uncertain Data," in ICDE, 2008.

[18] Jennifer Widom, "Trio: A System for Integrated Management of Data, Accuracy, an Lineage," in CIDR, 2005.

[19] Sarvjeet Singh et al., "The Orion Uncertain Data Management System," in COMAD, 2008.

[20] Bijit Hore, Jehan Wickramasuriya, Sharad Mehrotra, Nalini Venkatasubramanian, and Daniel Massaguer, "Privacy-Preserving Event Detection for Pervasive Spaces," in PERCOM, 2009.

[21] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Journal of Speech Communication, pp. 1162-1181, 2006.