Integrating Verbal and Nonverbal Input into a Dynamic Response Spoken Dialogue System

Ting-Yao Hu, Chirag Raman, Salvador Medina Maza, Liangke Gui,
Tadas Baltrusaitis, Robert Frederking, Louis-Philippe Morency, Alan W Black, Maxine Eskenazi
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, Pennsylvania 15213

Abstract

In this work, we present a dynamic response spoken dialogue system (DRSDS). It is capable of understanding the verbal/nonverbal language of users and making instant, situation-aware response. Incorporating with two external systems, MultiSense and email summarization, we built an email reading agent on mobile device to show the functionality of DRSDS.

Introduction

Current task-oriented spoken dialogue systems provide satisfactory performance in many scenarios. However, they still lag behind humans because of several reasons. Two examples are that people rely on both verbal and nonverbal language to communicate, and we perform dialogues incrementally, not in a turn-by-turn manner. This suggests several ways to improve current mainstream dialogue systems.

Some previous works explored these directions by developing nonverbal language detection and incremental processing components for spoken dialogue system. Perception makeup language (Scherer et al. 2012) was designed to describe the nonverbal behavior of users. On the other hand, Incremental language understanding (DeVault, Sagae, and Traum 2011) and incremental speech synthesis (Baumann and Schlangen 2012) were proposed to increase the responsivity of systems. However, to the best of our knowledge, there is no practical spoken dialogue system incorporating these two types of functions together.

We propose a dynamic response spoken dialogue system (DRSDS) framework. Comparing to conventional dialogue systems, DRSDS framework has two additional modules: nonverbal language understanding (NVLU) and dynamic natural language generation (dNLG). NVLU accepts a stream of nonverbal language detection results (e.g. users emotion, gesture...), and interprets it in real time, and dNLG module estimates the speech synthesis progress. Incorporating with NVLU and dNLG, applications based on our DRSDS framework can understand verbal and nonverbal language and give out dynamic response according to the inline expression of users.

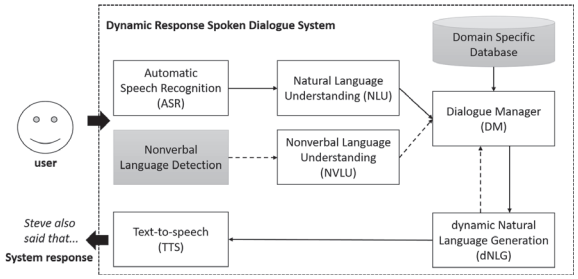

Figure 1: Architecture of DRSDS. The arrow with dot line represents the real time, incremental interaction between modules. The gray blocks are required external modules

In this demo, we implement an email reading agent using our DRSDS framework. This agent receives the users vocal request and reads the summarized content of an email. While the agent reading, the granularity can be changed dynamically based on user's reaction. To accomplish this application, we apply two external modules: Multisense and email summarization. The former detects users' facial expression, and the latter provides the content of summarization.

DRSDS Architecture and Functionality

Fig 1. illustrates the flow chart of DRSDS architecture. The system preserves the basic components of conventional turn-based spoken dialogue system: automatic speech recognition (ASR), natural language understanding (NLU), dialogue manager (DM), natural language generation (NLG) and text-to-speech (TTS). ASR takes the audio signal of user utterance as input and outputs the text. NLU analyzes the user's intention and extracts key information pieces from the raw text. DM keeps tract of the current dialogue state and decides which action to take. NLG and TTS organizes the content of system output utterance and plays it to the users.

Besides the basic functionality, DRSDS possesses two key additional features. It is capable of understanding users verbal and nonverbal language, and providing instant, situation-aware response according to user's inline expression and the current content system speaking. To achieve these two features, we developed nonverbal language understanding (NVLU) and dynamic natural language generation (dNLG), which will be introduced in the following. On the

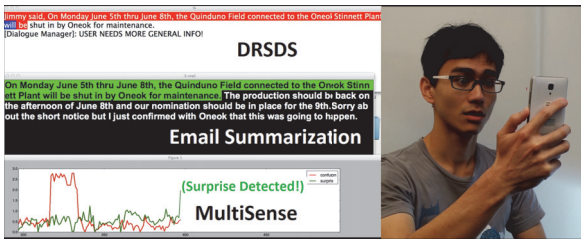

Figure 2: Screenshot of our demo video.

other hand, DRSDS framework has to adopt two external modules, nonverbal language detection and domain specific database, to build a practical application.

Nonverbal language understanding (NVLU)

NVLU module aims to interpret human's nonverbal language, e.g. facial expression and gesture. Considering that most types of nonverbal language are conveyed through visual information, our system captures the video stream of user and send it to external nonverbal language detection component. Then a stream of detection results are accepted by NVLU module, and NVLU reports the happening of events to DM. Currently, those events are defined manually.

Dynamic natural language generation (dNLG)

While regular NLGs usually send the system output utterance to the text-to-speech (TTS) engine, our dNLG module reports the current progress of the TTS engine back to the dialogue manager (DM). Although existed incremental TTS engine (Baumann and Schlangen 2012) has the ability to retrieve the timing of each word in the system utterance, we choose Google TTS because of its good quality and accessibility from mobile device. Since Google TTS generates the audio, we adopt another open source TTS engine, Flite (Black and Lenzo 2001), to estimate the current progress of system speaking.

With NVLU and dNLG, our system can predict what the user is perceiving, and align the content with the inline expression of the user, which is detected from NVLU.

Technical Details of Email Reading Agent

We present an email reading agent based on our DRSDS framework. This agent accepts the users vocal request and reads the summarized content of an email. The granularity of summarization can be dynamically changed, and the change is dependent on users nonverbal language, and the current content users perceiving. To accomplish this, DRSDS interfaces with two external systems: MultiSense and email summarization.

MultiSense

MultiSense is a state of the art framework of tools for automatic nonverbal behaviour analysis. One of these libraries is OpenFace (Baltrušaitis et al. 2016), capable of providing facial behaviour understanding in real-time. We use OpenFace for detecting two user affects: confusion and surprise. It achieves this by detecting and estimating the intensity of Facial Action Units (Ekman and Friesen 1977) using appearance and geometry features (Baltrušaitis, Mahmoud, and Robinson 2015). Confusion and surprise intensities are then computed as a linear combination of these Facial Action Unit intensities per frame.

Email Summarization

The text is summarized using TextRank (Mihalcea and Tarau 2004), a document graph-based ranking method where an undirected fully-connected graph is built from the text, with each sentence as a node, and with edge weights set using pointwise mutual information between word stems. Pagerank is then used to rank the sentences according to their centrality, and the ranked list is delivered to DRSDS to be shown as needed to the user.

Implementation and Demo

Most components of DRSDS (NLU, NVLU, DM and dNLG) on this email reading agent are implemented within an Android application. We adopt Microsoft bing API and Google TTS for ASR and TTS, respectively. MultiSense and email summarization are server side programs, and deployed as APIs for mobile clients. Fig. 2 illustrates the screen shot of our demo video. The video shows an user operating the email reading agent, along with the real time visualization of DRSDS, email summarization and MultiSense.

References

Baltrušaitis, T.; Robinson, P.; Morency, L.-P.; et al. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10.

Baltrušaitis, T.; Mahmoud, M.; and Robinson, P. 2015. Cross-dataset learning and person-specic normalisation for automatic action unit detection. In *Facial Expression Recognition and Analysis Challenge, in conjunction with FG*.

Baumann, T., and Schlangen, D. 2012. Inpro_iss: A component for just-in-time incremental speech synthesis. In *Proceedings of the ACL 2012 System Demonstrations*, 103–108.

Black, A. W., and Lenzo, K. A. 2001. Flite: a small fast runtime synthesis engine. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*.

DeVault, D.; Sagae, K.; and Traum, D. 2011. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue and Discourse* 2(1):143–70.

Ekman, P., and Friesen, W. V. 1977. *Manual for the Facial Action Coding System*. Palo Alto: Consulting Psychologists Press.

Mihalcea, R., and Tarau, P. 2004. Textrank: Bringing order into texts.

Scherer, S.; Marsella, S. C.; Stratou, G.; Xu, Y.; Morbini, F.; Egan, A.; Rizzo, A.; and Morency, L.-P. 2012. Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors. In *IVA*.