Chapter 7

Speech for Content Creation

Joseph Polifroni
Nokia Research Center, USA

Imre Kiss
Nokia Research Center, USA

Stephanie Seneff
MIT CSAIL, USA

ABSTRACT

This paper proposes a paradigm for using speech to interact with computers, one that complements and extends traditional spoken dialogue systems: speech for content creation. The literature in automatic speech recognition (ASR), natural language processing (NLP), sentiment detection, and opinion mining is surveyed to argue that the time has come to use mobile devices to create content on-the-fly. Recent work in user modelling and recommender systems is examined to support the claim that using speech in this way can result in a useful interface to uniquely personalizable data. A data collection effort recently undertaken to help build a prototype system for spoken restaurant reviews is discussed. This vision critically depends on mobile technology, for enabling the creation of the content and for providing ancillary data to make its processing more relevant to individual users. This type of system can be of use where only limited speech processing is possible.

INTRODUCTION

A couple are visiting Toronto and have just finished a meal at a small Chinese restaurant. The wife makes a habit of scouting out Chinese food in any city she visits and this restaurant was particularly good. As she walks out of the restaurant, she pulls out her mobile phone, clicks a button on the side, and speaks her thoughts about the meal she’s just eaten. She then puts her phone away, having recorded her impressions of the restaurant. Her location and the time of day have been recorded as part of the interaction. Our hypothetical user then hails a cab and goes off to the theater. Figure 1 shows what a user might say in this context.

The scenario we describe above is the first-stage interaction with an overall system that uses speech for content creation, social media, and
recommender systems. In subsequent sections, we will enlarge upon this scenario, with further glimpses into the user interaction and the underlying technology required for each step. We argue that these technologies are sufficiently advanced to enable the convenience of recording thoughts and impressions on the go, indexing the results, and extracting enough information to make the interaction useful for others.

One of the most important aspects of this scenario, and the ones that follow, is that the user is in charge of the interaction the entire time. Users do not have to worry about getting involved in an interaction when they’re busy, in a noisy environment, or otherwise unable to devote time to the interface. Users can describe an experience while it is fresh in their memory through an interface that is always available to them. When they have the time and the inclination to make further use of the information, they can examine, review, and, ultimately, share it. The spoken input takes the form of a “note to self,” where the user does not have to plan carefully what to say (Figure 2).

In this initial scenario, the user’s interaction with the system stops after the review is spoken. Either immediately, or when connectivity is re-established, speech is uploaded to a cloud-based system. With a combination of automatic speech recognition (ASR) and natural language processing (NLP) technologies, the system goes to work on indexing and deriving meaning from the dictated review. In the best case scenario, information about individual features, such as food quality or service, are extracted and assigned a scalar value based on user input. These values are used to populate a form, combined with other online sources of information (derived from GPS coordinates associated with the speech at the time of data collection), and made available to the user to review, modify, and share. Various other fallback levels of analysis are always available, so that the information is never completely lost or ineffectual. For example, the system may be able to only assign a single overall polarity to the entire review, or just extract some keywords for indexing. In the worst case, a simple audio file is saved and associated with a time-stamp and GPS location.