Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 21 (2007) 296–324
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Accessing speech data using strategic fixation

Steve Whittaker [a,*], Julia Hirschberg [b,1]

[a] Department of Information Studies, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK
[b] Department of Computer Science, Columbia University, 1214 Amsterdam Avenue, M/C 0401, 450 CS Building, New York, NY 10027, USA

Received 28 June 2005; received in revised form 9 June 2006; accepted 9 June 2006
Available online 26 July 2006

Abstract

When users access information from *text*, they engage in *strategic fixation*, visually scanning the text to focus on regions of interest. However, because speech is both serial and ephemeral, it does not readily support strategic fixation. This paper describes two design principles, *indexing* and *transcript-centric access* that address the problem of speech access by supporting strategic fixation. *Indexing* involves users constructing external visual indices into speech. Users visually scan these indices to find information-rich regions of speech for more detailed processing and playback. *Transcription* involves transcribing speech using automatic speech recognition (ASR) and enriching that transcription with visual cues. The resulting enriched transcript is time-aligned to the original speech, allowing users to scan the transcript as a whole or the additional visual cues present in the transcript, to fixate and play regions of interest.

We tested the effectiveness of these two approaches on a set of reference tasks derived from observations of current voicemail practice. A field trial evaluation of JotMail, an indexed-based interface similar to commercial unified messaging clients, showed that our approaches were effective in supporting speech scanning, information extraction and status tracking, but not archive management. However, users found it onerous to take manual notes with JotMail to provide effective retrieval indices. We therefore built SCANMail, a transcript-based interface that constructs indices automatically, using ASR to generate a transcript of the speech data. SCANMail also uses information extraction techniques to identify regions of potential interest, e.g. telephone numbers, within the transcript. Laboratory and field trials showed that SCANMail overcame most of the problems users reported with JotMail, supporting scanning, information extraction and archiving. Importantly, our evaluations showed that, despite errors, ASR transcripts provide a highly effective tool for browsing. Users exploited the enriched transcript to determine the gist of the underlying speech, and as a guide to identifying areas of speech that it was critical for them to play. Long-term field trials also showed the utility of transcripts to support notification and mobile access.

1. The problem of accessing speech archives

Most research on speech interfaces has focused on using speech as a *medium* for interacting with computers either in speech-only dialogue systems (Walker, 2000; Young, 2002; Zue and Glass, 2000) or multimodal ones

* Corresponding author. Tel.: +44 114 222 6340; fax: +44 114 278 0300.
E-mail addresses: s.whittaker@shef.ac.uk (S. Whittaker), julia@cs.columbia.edu (J. Hirschberg).
[1] Tel.: +1 212 939 7114.

0885-2308/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2006.06.004

(Oviatt, 2002; Walker et al., 2004). In contrast, we focus here on interfaces to speech *content*, and the development of interfaces allowing users to *browse and extract information from speech archives*.

Spoken information is ubiquitous (Kraut et al., 1990; Panko, 1993; Whittaker et al., 1994a). Increasingly large amounts of spoken data are being archived, such as meetings, phone calls, broadcast news and talk shows, and impromptu conversations (Galley et al., 2004; Emnett and Schmandt, 2000; Garofolo et al., 2000; Hindus et al., 1993; Janin et al., 2003; M4; Wellner et al., 2004). Currently, however, it is hard to exploit these archives because we lack effective end user tools. Our goal in this study is to devise and test principles for designing effective interfaces for browsing and searching speech corpora.

One obvious strategy for designing speech tools is to capitalize on successful techniques developed for accessing text. But there are crucial differences between speech and text. Although speech has advantages over text in being both expressive and easy to produce, it is serial and ephemeral, giving rise to significant access problems (Arons, 1997; Chalfonte et al., 1991; Hindus et al., 1993; Whittaker et al., 1998a,b). Simple processing studies also show that people extract information more quickly from text: average reading rates for text are 350 words/min compared with listening rates for speech of 180 words/min (Arons, 1992a,b; Beasley and Maki, 1976; Monk, 1984).

These differences arise from the different affordances of speech and text. Text can be processed more quickly because it is a permanent medium that affords *strategic fixation*, allowing readers to focus on important parts of a document while ignoring less significant regions. Studies of eye gaze confirm this; during reading, readers generally fixate upon less common (and hence more information bearing) words. They also fixate on longer words, and content words as opposed to function words (Rayner and Well, 1996; Schilling et al., 1998). They exploit formatting information (Askwall, 1985), suggesting that, overall, users strategically focus on those aspects of the document that provide the most information.

Speech, in contrast, does not readily support strategic fixation, making speech browsing a data-driven process. Indeed, in one study of access from a voicemail archive, we found that users' attempts to increase efficiency by strategic fixation and sampling had negative effects. Users forgot which parts of the archive they had already sampled, leading them to re-access the same information multiple times. On other occasions, sampling caused them to miss important material altogether, resulting in failure of their retrieval task. Overall, their sampling strategy ended up being less efficient than simply playing the speech from beginning to end (Whittaker et al., 1998a).

The goal of this paper is to determine whether we can improve speech browsing by designing interfaces that directly support strategic fixation, converting speech access from a *data-driven* into a *self-paced* activity. We explore two different strategic fixation techniques:

(a) *Indexing*. This involves constructing external visual indices into speech. Users visually scan these indices to find information-rich regions of speech for more detailed processing and playback. This technique clearly depends on the quality of the indices that we can construct.
(b) *Transcription*. This involves transcribing the speech using automatic speech recognition (ASR). By time-aligning the transcript with the speech, users can scan the resulting transcript, fixating and playing regions of interest. One obvious problem with this technique is that ASR transcription is inaccurate with only 50–90% (depending on genre) of words being accurately transcribed (Garofolo, 2000; Huang et al., 2001).

Prior research on speech browsing has generally adopted an *indexing* approach, exploring various different types of indices. *Speaker-based* indexing allows users to choose particular speakers they want to listen to (Degen et al., 1992; Hindus et al., 1993; Kazman et al., 1996; Wilcox et al., 1994). *Intonation* analysis can identify parts of speech that are emphasized so users can focus on these (Arons, 1997; Stifelman, 1996; Stifelman et al., 2001; Wilcox and Bush, 1991). We can also index significant *participant activities* or *external events*, such as when participants take notes during a meeting (Abowd et al., 1996; Moran et al., 1997; Stifelman et al., 2001; Whittaker et al., 1994b; Wilcox et al., 1997), or when a speaker changes slides (He et al., 1999). Others have identified 'hotspots' where multiple participants are highly involved in the conversation (Kennedy and Ellis, 2004; Wrede and Shriberg, 2003). These activities can then be used as landmarks to identify important parts of the meeting. Indices can also be extracted from *significant visual events* such as those detected in video