Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 68 (2015) 85–96

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A prosody-based vector-space model of dialog activity for information retrieval ☆

Nigel G. Ward [a,*], Steven D. Werner [a], Fernando Garcia [b], Emilio Sanchis [b]

[a] Computer Science, University of Texas at El Paso, 500 West University Avenue, El Paso, TX 79968, USA
[b] Departament de Sistemes Informatics i Computacio, Universitat Politecnica de Valencia, Cami de Vera s/n, 46020 Valencia, Spain

Received 16 January 2014; received in revised form 21 October 2014; accepted 14 January 2015
Available online 28 January 2015

Abstract

Search in audio archives is a challenging problem. Using prosodic information to help find relevant content has been proposed as a complement to word-based retrieval, but its utility has been an open question. We propose a new way to use prosodic information in search, based on a vector-space model, where each point in time maps to a point in a vector space whose dimensions are derived from numerous prosodic features of the local context. Point pairs that are close in this vector space are frequently similar, not only in terms of the dialog activities, but also in topic. Using proximity in this space as an indicator of similarity, we built support for a query-by-example function. Searchers were happy to use this function, and it provided value on a large testset. Prosody-based retrieval did not perform as well as word-based retrieval, but the two sources of information were often non-redundant and in combination they sometimes performed better than either separately.

Keywords: Search; Speech; Audio; Similarity judgments; Similarity metrics; Principal components analysis

1. Introduction

Searching for desired content in recordings of speech is today difficult and uncommon. This is despite the clear demand for tools to support search through recordings of lectures, meetings and dialogs, and despite substantial research on technologies to support audio search (Larson and Jones, 2012). Compared to text, speech is disadvantaged as a medium for search in some ways, notably the difficulty of automatically identifying the words spoken, but it also has a potential advantage in the presence of prosody, which often encodes information that may not be expressed in words. The potential value of prosodic information for search has long been noted (Hakkani-Tur et al., 1999), however demonstrated utility has been lacking.

In this paper we develop a new way to use prosodic information for search. The contributions include:

- developing a way to represent prosodic-context information with a vector space model (Section 3.1),
- showing that proximity in this space relates to dialog-activity similarity, to topic similarity, and to human judgments of similarity (Sections 3.2 and 4.2),

☆ We thank Martha Larson, Alejandro Vega, Steve Renals, Khiet Truong, Olac Fuentes, David Novick, Shreyas Karkhedkar, Luis F. Ramirez, Elizabeth E. Shriberg, Catharine Oertel, Louis-Philippe Morency, Tatsuya Kawahara, Mary Harper, and the anonymous reviewers. This work was supported in part by the National Science Foundation under Grants IIS-0914868 and IIS-1241434 and by the Spanish MEC under contract TIN2011-28169-C05-01.

* Corresponding author.
E-mail addresses: nigelward@acm.org (N.G. Ward), stevenwerner@acm.org (S.D. Werner), fgarcia@dsic.upv.es (F. Garcia), esanchis@dsic.upv.es (E. Sanchis).
URL: http://www.nigelward.com (N.G. Ward).

http://dx.doi.org/10.1016/j.specom.2015.01.004
0167-6393/

– finding that users appreciate a more-like-this feature when searching in dialog archives (Section 4),
– presentation of a corpus of "social speech," dialogs among members of an organization, annotated for similarity (Section 5),
– a new measure for judging the utility of the results of search in unsegmented speech (Section 5.4),
– finding that simple city-block distance outperforms Euclidean distance as a proximity metric, and that weighted distance measures do even better (Section 7), and
– finding that prosodic similarity provides less value for search than lexical similarity measures (Section 8).
– finding that prosodic information can usefully complement word-based search (Section 9).

This article brings together results that have been previously reported only piecemeal,[1] and presents new results.

2. Background: prosody for search in speech

Most current spoken dialog retrieval systems are based on the view that speech is essentially just noise-corrupted text (Chelba et al., 2008). They use speech recognition techniques to infer the words said, and then use text-based search techniques on the resulting transcript. However the performance of such systems is generally weak, and today audio search is not widely used. While progress is ongoing, some fundamental assumptions – that speech recognition is mostly accurate, that all words are in the recognizer's vocabulary, that ambiguity, anaphor and ellipsis are rare, and that searchers can specify all words and synonyms relevant to their intent – fundamentally limit the performance of search using only this approach.

An alternative view focuses on the fact that spoken dialog is a rich information resource. One way to appreciate this is to think about why people speak to each other at all, especially today when there are more ways to communicate, with texting increasingly popular. Special properties of spoken dialog include its utility for establishing rapport, for allowing self-expression, for conveying and appreciating personality, for talking about personal matters, and for dialog activities that involve emotion or interpersonal interactions, such as persuading, apologizing, justifying, explaining preferences, and reaching decisions. This perspective suggests that we try to exploit such aspects of speech for audio search.

Doing so aligns with the growing understanding that the needs of searchers involve more than just finding content that matches a query. What searchers want may also be characterized in part by an intent (Rose et al., 2004; Hanjalic et al., 2012), and this may in particular relate to an interest in certain dialog processes (Pallotta et al., 2007) or activities, for example recommending, answering a question, agreeing, forming a decision, telling life stories, making plans, hearing surprising statements, giving advice, explaining, and so on.

The use of prosodic information can address these needs, potentially overcoming the shortcomings of lexical search alone by leveraging the pragmatic richness of dialog. Various approaches have been tried. Hakkani-Tur et al. (1999) noted that important words and phrases can be prosodically distinctive and that this can be used to focus search. Most research relating to using prosody for audio search has focused on detecting dialog activities that people might like to search for. Prosody-based classifiers can, for example, spot interactional "hotspots" where the speakers are unusually involved (Wrede and Shriberg, 2003; Oertel et al., 2011), conflicts (Kim et al., 2012), agreements on action items (Purver et al., 2007), various emotional and attitudinal states and stances (Toivanen and Seppänen, 2002; Wollmer et al., 2013), and dialog acts such as question, apology, promise, and persuasion attempt (Larson and Eskevich, 2011; Freedman et al., 2011). This work has shown that many dialog activities are indeed associated with characteristic prosodic features and patterns. In addition there are bottom-up observations that support this approach, for example that "some conversation topics tend to have . . .slower speaking rates" (Yuan et al., 2006).

However the underlying value proposition is not clear. Being able to retrieve all regions that match a specified dialog-act tag is not obviously something that real users want to do, and in fact the benefits of such a function have never been demonstrated. There are, moreover, reasons to think this unlikely to be of great value. In most dialog genres, simple dialog-act tags fail to really capture what is going on in any specific utterance, especially since many utterances have multiple functions (Bunt, 2011). In general, it seems unlikely that such *a priori* tags, or indeed any finite taxonomy of actions, will be adequate for describing the space of human activities (Lukowicz et al., 2012). Even if there were, searchers are unlikely willing to learn them well enough to comfortably use them when formulating queries.

We avoid these problems by adopting more a realistic expectation of users: that they can recognize the sorts of things that they are interested in when they hear them, and thus can benefit from a "more like this" function (Liu and Huang, 2000; Mizuno et al., 2008; Oard, 2012) that returns results similar to a "seed," that is, an audio snippet used as a query. This avoids the finite taxonomy problem. To support this, we employ an empirically-derived representation of dialog activities.

[1] specifically a conference paper, three workshop papers, and two technical reports: The idea of using a vector-space model of prosodic context for information retrieval, the qualitative analysis of similarity in this space, and the initial user study were reported in Ward and Werner (2013b), the need for a corpus of social speech was explained in Ward and Werner (2012) and described partly in Ward et al. (2013) and Ward and Werner (2013a), the comparison of the training schemes and distance metrics was reported in Werner and Ward (2013), and the comparison to lexical measures of similarity was reported in part in Garcia et al. (2013).