Ward, NG.; Werner, SD.; García-Granada, F.; Sanchís Arnal, E. (2015). A prosody-based vector-space model of dialog activity for information retrieval. Speech Communication. 68:85-96. doi:10.1016/j.specom.2015.01.004.

A Prosody-Based Vector-Space Model of Dialog Activity for Information Retrieval

Nigel G. Ward, Steven D. Werner

Computer Science, University of Texas at El Paso
500 West University Avenue, El Paso, Texas 79968 USA

Fernando Garcia, Emilio Sanchis

Departament de Sistemes Informatics i Computacio, Universitat Politecnica de Valencia
Cami de Vera s/n, 46020, Valencia, Spain

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.

Key words: search, speech, audio, similarity judgments, similarity metrics, principal components analysis, social media, query by example, evaluation by user simulation

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,

☆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.

Email addresses: nigelward@acm.org (Nigel G. Ward), stevenwerner@acm.org (Steven D. Werner), fgarcia@dsic.upv.es (Fernando Garcia), esanchis@dsic.upv.es (Emilio Sanchis)

URL: www.nigelward.com (Nigel G. Ward)

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),

– 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], providing a self-contained explanation and situating the findings with respect to past work and future possibilities.

2. Background: Prosody for Search in Speech

Most current spoken dialogue 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

[1]specifically two conference proceedings, three workshop papers, and two technical reports: The idea of using a vector-space model of prosodic context for information retrieval was presented in (Ward and Werner, 2013b), the qualitative analysis of similarity in this space was presented partly there and partly in (Ward et al., submitted), the initial user study was 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 and combination with lexical measures of similarity was reported in part in (Garcia et al., 2013) and in (Ward et al., submitted).

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 and Levinson, 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 et al., 2011; Freedman et al., 2011). This work has shown 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.

volume: -3150∼-1550, -1550∼-750, -750∼-350, -350∼-150, -150∼-50, -50∼0
pitch height: -4650∼-2250, -2250∼-1050,-1040∼-450, -450∼-150, -150∼0
pitch range: -1840∼-920, -920∼-460, -460∼-230, -230∼0
speaking rate: -2640∼1320, -1320∼-660, -660∼-330, -330∼0

Table 1: The window sizes used to compute the 76 prosodic features input to PCA for the final experiments. Windows starts and ends are in milliseconds before from the point of interest. The same window sizes were used, symmetrically, to compute features after the point of interest, and a complete set of features for the interlocutor track were also computed.

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.

3. Approach: A Vector-Space Representation of Dialog Activity

3.1. Defining the Vector Space

Our representation is derived by applying Principal Component Analysis (PCA) to a wide sampling of local prosodic features. While using the common features pitch height, pitch range, speaking rate, and volume, this feature set is novel in three ways. First, the features are computed over different windows across six-seconds of context, thus capturing significant local context. Second, they are computed for both participants in the dialog, thus capturing both speaker and listener behavior. Third, they are computed over fixed windows, rather than being word-, syllable-, or phrase-aligned, thus better capturing dialog-activity information. The specific feature set was chosen for simplicity of computation, for providing coverage of most of the prosodic aspects known to be most relevant for dialog, and for symmetry, as shown in Table 1.

Surprisingly, the dimensions resulting from PCA turned out to be meaningful: each of the top couple dozen turned out to align with some known aspects of dialog, such as grounding, turn-taking, seeking and expressing sympathy, degrees of novelty and interest, topic shifts and closings, emphasis, explanations, humor versus regret, personal versus impersonal topics, and facts versus opinions (Ward and Vega, 2012; Ward, 2014). Table 2 shows our tentative interpretations of the top twenty dimensions. We therefore can refer to the space defined by these dimensions as "dialog-activity space." In a dialog every point in time maps to a point in this 76-dimensional space.

3.2. Similarity and Proximity in this Space

Proximity in this vector-space model can serve as a measure of similarity. Initially we used simple Euclidian distance, over all 76 dimensions. As a preliminary exploration, we selected a few seeds and examined what positions the model found as most similar. As

1	this speaker talking vs. other speaker talking	32%
2	neither speaking vs. both speaking	9%
3	topic closing vs. topic continuation	8%
4	grounding vs. grounded	6%
5	turn grab vs. turn yield	3%
6	seeking empathy vs. expressing empathy	3%
7	floor conflict vs. floor sharing	3%
8	dragging out a turn vs. ending confidently and crisply	3%
9	topic exhaustion vs. topic interest	2%
10	lexical access or memory retrieval vs. disengaging	2%
11	low content and low confidence vs. quickness	1%
12	claiming the floor vs. releasing the floor	1%
13	starting a contrasting statement vs. starting a restatement	1%
14	rambling vs. placing emphasis	1%
15	speaking before ready vs. presenting held-back information	1%
16	humorous vs. regrettable	1%
17	new perspective vs. elaborating current feeling	1%
18	seeking sympathy vs. expressing sympathy	1%
19	solicitous vs. controlling	1%
20	calm emphasis vs. provocativeness	1%

Table 2: Interpretations of the top 20 dimensions, with the variance explained by each, from Ward and Vega (2012)

hoped, proximity correlated with similarity: generally the closer the proximity to the seed, the more similar the regions sounded. These similarities were not just in dialog activity but frequently also in content.

For example, an attempt to find information about family members across two 5-minute dialogs, by unrelated speakers, using as seed the phrase *my brother's a trim carpenter,* using proximity with a certain threshold, gave us 14 matching regions, and of these 7 included information about family members (husbands, children), including some where there was no noun present, only the word *he*. Interestingly, most false positives related to house construction, and might have been prevented by negative relevance feedback. By comparison, textual search on the transcripts of these dialogs using ten family-related terms gave only 4 hits, and these were hidden among false positives including, for example, a generic discussion of moms who work.

A second interesting example was a search for complaints about the government. A little browsing turned up a complaint about a book, using this as a seed led to a complaint about a husband, using this as a seed in turn led to a complaint about the metric system, which finally led to results including complaints about property tax, gun laws, and public schools.

4. Preliminary Human-Subjects Evaluation

We performed two evaluations of this model. In the first we experimented with making results of prosody-based search directly available to users.

For this, in the interest of realism, we developed our own evaluation suite. Existing standard evaluation infrastructures for audio search are mostly of two kinds. The first kind supports evaluation only at the level of component technologies, such as search for words, rather than measuring the utility for people actually performing search tasks. The second kind does involve tasks, but the granularity of retrieval is coarse, only returning "documents," such as news segments or Youtube videos (Hanjalic et al., 2012; Larson et al., 2011; Garofolo et al., 2000; Larson and Jones, 2012; Eskevich et al., 2013), rather than helping searchers locate specific relevant content inside such documents. The sole exception, which both uses task-based evaluation and returns sentence-level results, addresses only search for factoids in monologs (Akiba et al., 2011). Thus a realistic evaluation was something we had to develop ourselves.

As our aim was not to discover user needs (Whittaker et al., 2008), but to support controlled experiments, we prepared a fixed list of 20 search tasks for subjects. These were developed by first considering 14 diverse audio search scenarios and identifying representative tasks (Ward and Werner, 2012), and then selecting 20 tasks that made sense for the dataset we chose, Switchboard (Godfrey et al., 1992). In Switchboard the speakers were given suggested topics, but in practice talked mostly about whatever they wanted to, exhibiting a large variety of topics and dialog styles, making it suitable for evaluating diverse tasks. The tasks we chose included looking for places in dialog where: the participants started to relax with each other, mentioned where they live, revealed or discovered that one participant was older or higher in status, talked about future plans, tried to teach or explain something, used the word *nine*, or discussed something that might suggest a birthday present idea. Each task was given to the searchers as a short paragraph explaining what information was wanted and giving a plausible reason why.

4.1. Interface and Procedure

We hypothesized that providing prosody-based results would be liked by searchers even if they already had access to standard lexical search. To test this we built a comprehensive search interface. Best practice in interfaces for audio search includes the ability to jump to search results in the audio and the ability to listen to and navigate within the audio (Whittaker et al., 1999). Our interface was accordingly built on top of a simple audio listening application that displays the manually transcribed text and the audio signal. This was augmented with a sidekick window to provide the search functions.

Lexical search was provided with an input box for the search terms and a display area for matches. Search was simple substring search, without advanced features. Each match was displayed as a few words of context centered on the matching word. Users could click on a match to jump to that position in the audio display, where they could hear the result and see the full transcript. After determining that a phrase or utterance is relevant, the

searcher could click and sweep over those few seconds to specify the region and save it as a result.

Prosody-based search was also provided with a simple interface. To access prosody-generated results, the user selected a point in the audio display as a seed and invoked the search command. Regions that were prosodically similar were then presented in a list for the user to peruse. Matches were also shown in the audio display, with a horizontal bar over each similar region, with a match-quality score also shown. As the similarity metric provided scores only for individual timepoints, there was an extra step involved in generating the regions: this involved selecting all timepoints within a proximity threshold of the query, and grouping nearby sets of such points into regions, overlooking gaps of up to 50 milliseconds, in order to avoid multiple fragmentary close results.

We hired four students as searchers. Each did twenty search tasks, spending about 30 minutes on each. For half of the tasks the dialog-activity search function was available; for the others it was turned off.

4.2. Results

While the results were not entirely clear-cut (Ward and Werner, 2013b), there was good evidence that the searchers liked the query-by-example feature. First, in the free comments section of the final questionnaire all four searchers noted that it was beneficial, although with different nuances. For example, one said, "Although it did take me a while to really learn how to use the lexical plus prosodic system, in the end, I was able to generate about twice the results (as compared with just using the lexical system)." Second, we looked at what the searchers had done just before identifying a result that they chose to save, based on the logfiles. This indicated that 110 of the results were attributable to dialog-activity search. These were 40% of the 272 found when dialog-activity search was available, indicating that it was frequently used and useful. Third, although there was no requirement to use dialog-activity search after the initial training, all searchers used it at least occasionally throughout the experiment; no one abandoned it to revert to purely lexical search. These results show that query-by-example search, using a prosody-based vector-space representation of dialog acts, can be a useful and user-appreciated addition to standard word-based search.

This was despite several methodological limitations. For one thing, the search interface was not as good as it could have been: our implementation was slow, using linear search without indexing rather than a nearest-neighbors algorithm (Slaney et al., 2012); the search results were presented in straight temporal order, rather than ordered by match quality; and there was no way to combine or cross-filter the prosody-based results and the lexical results. These factors suggest that the benefit seen here may understate the true value. Another issue was that the baseline search used accurate hand-labeled transcriptions, although in practice only errorful speech-recognition output is available. Finally, the weights for the various dimensions were not tuned. We therefore began to plan a follow-on study, to examine in particular these last two issues.

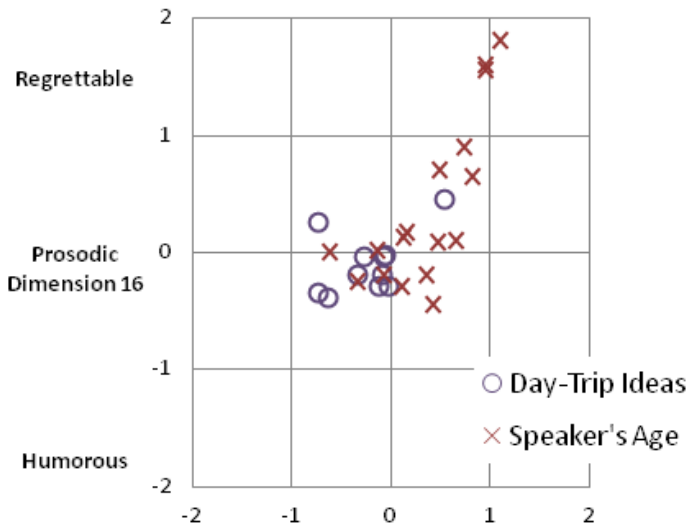

Figure 1: Locations in dialog-activity space of all results saved by searchers for the tasks of finding "day-trip ideas" and "information about the speaker's age," projected to two dimensions.

4.3. Analysis

We also explored how proximity was working, by examining the distributions of the saved results. For each task, we found the average location in dialog-activity space, for each user, of all the results he or she saved. We did this for all four searchers, and noted the dimensions for which the averages of all four fell in the same halfspace.

For the task of finding information about a speaker's family structure, the result averages fell on the high side of dimension 2, the low side of dimension 4, and the high side of dimension 8, among others. Referring to the interpretations by Ward and Vega (2012a), this means that talk about the family occurred: when both speakers are involved, rather than one producing a monolog; when new information is being introduced and being grounded, rather than being elaborated upon; and when the speaker is confident and ending crisply, rather than dragging out the turn. Considering that talk about family members with strangers tends to be brief and unelaborated, courteously acknowledged, and generally incidental to other topics, this distribution makes sense.

In comparison, for two different tasks the distributions were different. For the task of finding complaints about the government, the result averages fell in the halfspaces of swift topic shifts, seeking empathy and agreement, floor conflict, topic involvement, floor claim attempts, using contrast as a rhetorical structure, presenting new perspectives, and being provocative. For the task of finding where the speaker is currently living, the averages were in the halfspaces of turn grabs, floor sharing, rambling on, restating previously-mentioned information, and speaking off the cuff rather than after preparatory thought. Figure 1 also illustrates topics occupying different regions in dialog-activity space.

In general these distributions are plausible reflections of the nature of these topics and how people tend to talk about them with strangers. Thus, as expected, the results for different tasks tend to fall in different regions of dialog-activity space, and using proximity

is well founded.

5. The Social Speech Corpus and Task

We needed a new evaluation corpus and performance measure, for three reasons. First, we wanted to be able to train our models. Although simple Euclidean distance worked, we wanted to develop a better metric. Second, we wanted to be able to evaluate our method and our models more objectively. Third, we wanted to support the wider community of researchers working on retrieval in spoken dialog.

While good collections of similarity judgments exist for other tasks, such resources have been lacking for speech data (Jung and Na, 2013). Previous evaluation resources all use only single-facet approximations to similarity. Many evaluations, for example, are based on topic match, which is suitable in domains with clear topics, but not so much for spontaneous-speech collections. Another common approximation is to frame the search problem as that of finding lexical matches to a query term, although in most cases people searching in audio are actually not really interested in finding specific words. Dialog-act match is another approximation to similarity, focusing again on only one facet, and incorporating some limiting assumptions, as noted earlier.

Thus we set out to develop a set of "pure" similarity judgments, not based on any theory or model, but directly based on annotators' unbiased opinions. To enable meaningful judgments, we wanted them to be made by people with good knowledge of the content and context, and with an intrinsic interest in the topics represented. Given that our judges would be students, we accordingly developed a "social speech" corpus of casual dialogs between students.

Wanting this collection to be suitable for evaluating any kind of method for search in spoken dialogs, we took care to avoid anything potentially unfair to one or another approach. This we did with the help of an international advisory board and the organizers of the MediaEval Benchmark program. After several iterations, this collection, together with the performance measure described below, was approved by MediaEval and incorporated as a Brave New Task in their 2013 program. (In the end teams from several institutions participated on this task, and four succeeded in completing systems, in part in a spirit of competition, but mostly to advance their own research agendas, including issues such as segmentation and query expansion (Galuščáková and Pecina, 2014; Levow, 2013).)

5.1. The Social Speech Scenario

With users' growing willingness to share personal activity information (Mairesse et al., 2012; Buckel and Thiesse, 2013), it seems inevitable that social media will eventually include social *multi*media, such as video and audio recordings of casual interactions.

Thus we can imagine scenarios like the following: A new member has joined an organization or social group that has a small archive of conversations among members. He starts to listen, looking for any information that can help him better understand, participate in, enjoy, find friends in, and succeed in this group. As he listens to the archive (perhaps at random, perhaps based on some social tags, perhaps based on an initial keyword search), he

finds something of interest. He marks this region of interest and requests "more like this." The system returns a set of "jump-in" points, places in the archive to which he could jump and start listening/watching with the expectation of finding something similar.

While the scenario is for us primarily a vehicle for obtaining similarity judgments, it may have value in itself; perhaps a future social media platform will support audio recording and provide recommendations or search results over this data. This scenario can also serve as a proxy for other search needs, such as search of workplace recordings, of surveillance recordings, of personal recordings, and so on (Ward and Werner, 2012).

5.2. Data Collection

We recorded two-person dialogs among members of the computer science community at an American university. They talked about whatever they wanted, for about 10 minutes each. They were told that that their dialogs would be annotated for later searching, and many of the conversations turned out to be rich in information likely to be of interest to fellow computer science students.

Subjects were audio- and video-recorded talking through a glass window, with head-mounted microphones to obtain stereo audio with minimal bleeding across tracks (Ward and Werner, 2013a).

We gathered a training set of 20 dialogs, 241 minutes in total, and a test set of 6 dialogs, 68 minutes total. The most common topics related to classes and class assignments, interesting new technologies, career ambitions, games, and movies. The training and test sets were the same in most respects, however the test set had mostly new speakers (9 new, 3 old), was recorded in the summer rather than during a semester, involved only research-active students, and included more students from outside the local area. The topics were largely similar, although the test set had less talk about classes and more about research.

5.3. Annotation

Annotators were then employed to produce similarity judgments. The annotators did their work in two steps. First they watched and listened to a few dialogs and developed a set of tags to use, each tag associated with some set of regions they considered somehow similar and that they thought some future searcher might potentially be interested in. They then did a second pass, this time over all the dialogs, and for every region that they felt relevant to some tag, assigned that tag to that region. Thus every pair of regions that share a tag is an example of a perceived similarity. Below we refer to a set of such regions as a tagset. Finally, each annotator gave each of their tags a quality label, based on the perceived utility of that tag for search, on a scale from 0-3, where the higher values were generally for more cohesive tagsets and more generally-interesting content.

The training set was annotated by six students and the testset by two, one of whom was new. Most of the annotators were members of the same community, including some who had contributed dialogs to the collection. The annotators worked mostly independently. Regions could span any fragment of the dialog, regardless of any notion of topic or utterance boundary. The average durations were 50 seconds in the training set and, after clarifying the instructions to annotators slightly, 31 in the test set. The tagsets were not edited or

coalesced in any way, as we trusted that each tagset was valid for the person who defined it. There were 198 tagsets over the training set, with a total of 1697 tagged regions, and 29 and 189 for the test set.

We observed that most tags related to traditional-style topics, such as #food, #travel, #cars-and-driving, #planning-class-schedules, #TV-shows, #lack-of-money, and #family. but others related mostly to dialog activity, for example #anecdotes, #problems, #advice, #gossip, while some related to both, for example #short-term-future-plans, and #positive-things-about-classes.

5.4. The Searcher-Utility-Ratio Performance Measure

Our task is finding regions similar to a query region. For evaluation purposes, an input to a system is a region from one of the similarity sets of an annotator, and the ideal result is the set of onsets of all the other regions in that set.

We defined and used a large number of measures to evaluate and understand model performance (Ward et al., 2013), however here we limit discussion to one. This is a variant of the standard mGap measure (Liu and Oard, 2006; Eskevich et al., 2012), and is similarly motivated by a weakness of simple precision when the data is not pre-segmented, as here: since jump-in points can be anywhere, those which are closer to the onset of a relevant segment ought to count for more than those farther off.

To quantify this, we devised a rough model of how searchers are likely to use the suggested jump-in points, that is, a sort of user simulation. Based on this we define a "searcher utility ratio," where the numerator is the estimated value to the searcher and the denominator the estimated cost, both measured in seconds. This ratio is our primary performance measure.

The value is modeled as the number of seconds of relevant audio/video a searcher can likely find by using the suggested jump-in points. We assume that she will find a region if a jump-in point is no earlier than 5 seconds before the region start and no later than 3 seconds before the region end. These margins are tighter than others in the literature (Galuscakova et al., 2012), but reasonable for our casual browsing scenario, where searchers are likely to have low persistence.

The cost is modeled as the number of seconds a searcher needs to peruse the suggested jump-in points. There are three cases. 1) If the suggested jump-in point does not correspond to any same-tagset region (a false-positive error), then the cost is 8 seconds, an estimate of the time a searcher needs to recognize a false alarm. 2) If the suggested jump-in point is no more than 5 seconds before the actual region start point, the cost is the time from that jump-in point to the end of the actual region, reflecting the time spent to scan forward to the start of the relevant content and the time to listen to it. 3) If the suggested jump-in point is within the region, then the benefit is the remaining duration of the region, and the cost is the same.

We further assume the searcher devotes a maximum of two minutes to each search. The numerator of the searcher utility ratio is accordingly the total amount of relevant audio she can find and consume in that time, according to the model above. For queries where the corpus has less than two minutes of relevant audio to find, the denominator is the total

duration of the relevant regions, otherwise it is 120 seconds. Thus the denominator is the total time spent, including the unproductive time

This model of search behavior is of course highly simplified. Unrealistically it assumes that all search tasks are equally important, that all similar regions are equally valuable, that every second within a similar region is equally valuable, and that searchers never give up after getting unhelpful search results. It also incorporates the more plausible assumption, for our scenario, that users will be happy with just some of the available information, rather than wanting to find every relevant region.

It is worth noting that this measure, although valid for comparing different systems' relative ability to find similar regions, significantly understates performance. This is because a region similar to a query, but in respects other than those considered by an annotator, will be counted as a false positive. In other words, because each similarity set is generated by a specific annotator, with his or her own perspective and interests, no system can be expected to find the same exact set of results.

6. Three Similarity Models

This data enabled us to train models to better correlate with human similarity perceptions. Many machine-learning algorithms could be suitable, but for this, our first exploration, we tried only simple techniques. This was for three reasons. First, we thought that our features and this task naturally called for distance-based methods, second, we wanted to focus attention on metrics that would work for any language and any data, without training, and third, we thought that simpler models would make it easier to reason back from the observed performance to gain insight into the nature of the problem and better understand the dimensions.

Thus all models were based on the prosodic dimensions, and all judged the similarity of any two points by comparing them across the dimensions, scoring them as more similar to the extent that their values on each of the dimensions were similar.

For these experiments we switched to a new feature set with more volume and rate features, with more speaker features and fewer interlocutor features, and with more narrow-window features close to the point of interest and fewer distant-context features: 78 features in total, as seen in Table 6. PCA was applied as before.

We examined three ways of computing similarity: Euclidean distance,

$$dissimilarity_E = \sum_{i=1}^{78} (x_i - y_i)^2 \tag{1}$$

city-block distance,

$$dissimilarity_C = \sum_{i=1}^{78} |x_i - y_i| \tag{2}$$

and a weighted distance,

	Past		Future	
	Self	Interlocutor	Self	Interlocutor
Volume	–3200∼–1600	–3200∼–1600	0∼50	0∼200
	–1600∼–800	–1600∼–800	50∼100	200∼400
	–800∼–400	–800∼–400	100∼200	400∼800
	–400∼–300	–400∼–200	200∼300	800∼1600
	–300∼–200	–200∼0	300∼400	1600∼3200
	–200∼–100		400∼800	
	–100∼–50		800∼1600	
	–50∼0		1600∼3200	
Pitch	–800∼–400	–800∼–400	0∼50	0∼200
Height	–400∼–200	–400∼–200	50∼100	200∼400
	–200∼–100	–200∼0	100∼200	400∼800
	–100∼–50		200∼400	
	–50∼0		400∼800	
Pitch	–800∼–400	–800∼–400	0∼50	0∼200
Range	–400∼–200	–400∼–200	50∼100	200∼400
	–200∼–100	–200∼0	100∼200	400∼800
	–100∼–50		200∼400	
	–50∼0		400∼800	
Speaking	–1600∼–800	–1600∼–800	0∼50	0∼200
Rate	–800∼–400	–800∼–400	50∼100	200∼400
	–400∼–200	–400∼–200	100∼200	400∼800
	–200∼–100	–200∼0	200∼400	800∼1600
	–100∼–50		400∼800	
	–50∼0		800∼1600	

Table 3: The 78 prosodic features input to PCA for the final experiments.

$$dissimilarity_W = \sum_{i=1}^{78} w_i |x_i - y_i| \tag{3}$$

The reason for trying weighting was that some of the dimensions seemed especially useful for the similarity computations and/or especially revealing of dialog activities, and we wanted our models to reflect this. However we note that the models without explicit weights anyway give more importance to the top dimensions, because they have higher variance and therefore affect the results more strongly.

6.1. Training Procedures

While the first two models have no free parameters, for the third the weights needed to be determined. Ideally the weights should have been set to maximize our metric, the searcher utility ratio. However, to simplify training, we used a simpler goal in this stage. Specifically, we trained to improve models' ability to determine, given any pair of timepoints, whether they were similar or not, that is, whether both appeared in regions that shared a tag given by some annotator. This we did using linear regression, where the target was a distance of 0 if x and y were similar, and 1 if not.

Thus, for example, if two selected timepoints x and y both were located in regions that had been tagged as talk about #favorite-movies, then x and y were counted as similar. If x and y shared no tags, they were counted as not similar. This is of course not an infallible indication, since a point pair might be similar even if they did not belong to regions that were felt to be worth tagging. Nevertheless, overall this gave us a set of point pairs guaranteed to be similar and a set that was probably not, which was good enough for training.

When generating the sets of point-pairs for training, we experimented with various more restrictive definitions of similar. One type of constraint was to require agreement by multiple annotators in order to consider two points similar. For this the tag names were ignored (as always), and so the annotators might have considered the points to be similar in different ways entirely. The second type of constraint was to require a minimum quality label for each tag, for example, to include only pairs whose shared tag was rated 3, excluding those rated 0, 1, or 2. Requiring higher quality labels and more agreement gave higher-quality training data, but at the cost of reducing the variety. For all selection schemes, the sets of similar and non-similar time-point pairs were obtained by randomly sampling points over the training set, with oversampling to ensure equal numbers of similar and non-similar pairs.

To evaluate the utility of these various schemes, we used a simple figure of merit representing the separation between the similar and non-similar sets, namely the difference between the mean distance between the non-similar point pairs and the mean distance between the similar point pairs, divided by the standard deviation of all distances.

6.2. Preliminary Findings

Using this separation measure, we evaluated various models and training schemes. We found several things. First, we confirmed the value of simple Euclidean distance: it gave a 0.11 separation, compared to the 0.0 a non-informative model would give.

Second, we confirmed our hunch that using only selected, good-quality data could give better results: a weighted model trained this way gave a separation of 0.22, compared to 0.19 when trained on all the data.

Third, we discovered that linear regression consistently gave negative weights to some of the dimensions, for example 67, which, when we listened to it, seemed to encode the difference between calm, indifferent speech and energetic explaining. This led us to try pruning, that is, feature selection. The first method tried was simply leaving a dimension out of the model (setting its weight to zero), if doing so improved performance on a held-out subset of the training data. This was repeated for the other dimensions, resulting in dropping about a third. This also improved results, however there was an interaction, in

that feature selection was generally highly valuable only when the training data was lower quality. Our second pruning method was the simpler one of just dropping any dimension to which regression assigned a negative weight.

During these preliminary studies we found that good separation did not always predict good task performance, so we returned to the searcher utility ratio for the final tuning and evaluation. This required an extra step in our system, to convert from the point-similarity judgments of the model to similar-region finding. We solved this translation problem simplistically, by using the middle point of the query region as the seed, and then returning the most similar time points, across the entire corpus, as the ranked list of jump-in points.

7. Results

	searcher utility ratio	
	training data	test data
Random	0.25	0.12
Euclidean Distance	0.21	0.12
City-block Distance	0.22	0.17
Weighted, all	0.17	0.19
Weighted, good	0.17	0.18
Weighted, all-p	0.17	0.18
Weighted, good-p	0.17	0.15
Weighted, all+	0.26	0.22
Weighted, good+	0.27	0.18
Weighted, all-p+	0.27	0.22
Weighted, good-p+	0.28	0.20

Table 4: Model Performance. All = trained on all tagsets, Good = trained on only selected tagsets; p = iterative-leave-one-out pruning applied to dimensions, + = only positively-weighted dimensions retained.

Table 4 shows the results for the two simple models, for the best of the weighted-distance models, and for a few others to show the effects of training-data selection and of dimension pruning. As a baseline, results are also shown for randomly choosing the jump-in points for each query.

Before considering the implications of the testset results for the different models, it is worth commenting on the generally worse performance seen on the training-set data. We think this is probably because the training set included a few very-long regions, which made it easier for any method (even random) to get a high score.

Surprisingly, Euclidean distance performed no better than random. Apparently its ability to discriminate similar from non-similar points did not translate to the ability to find the

onsets of similar regions. However the city block distance did perform better than the baseline, and most of the weighted models did even better

The various training and pruning techniques mattered. Training using all the tagsets was best. Pruning was effective, with dropping negatively-weighted dimensions being the best method, and usually with some additional benefit from also selectively dropping dimensions that did not contribute to training set performance.

8. Comparison to Lexical-Similarity Models

The natural next questions are, is prosody-based similarity better than lexical similarity? and can the two be combined to perform better than either alone?

To investigate this we developed four lexical models. These were not strawmen to make the prosodic models look good; on the contrary, they were developed independently and competitively, demonstrated best-in-class performance (Garcia et al., 2013), and were subsequently further improved.

The lexical models varied in two ways: either they used thesaurus information or they didn't, and either they used good transcripts or poor ones. The former were human-generated transcripts and the latter automatic speech recognition (ASR) output, kindly provided by Steve Renals using an experimental system. Although using acoustic models and a language model trained on meeting data, for Switchboard it gave transcripts with word error rates of about 30% to 60%.

The lexical methods involve two main processing phases: one to find the words that best represent the meaning of the segments (a bag-of-words representation), and one to then compute a similarity measure between pairs of segments.

In the first phase, the method first divided all the data into segments. For the human-generated transcriptions these were the phrases as the transcribers produced them, and for the ASR transcriptions these were pause-delimited regions. Second, it applied part-of-speech tagging (Toutanova et al., 2003), both to disambiguate some words and to set up for the next step. Third, it applied morphological processing to get the base forms of the words, as inflections (number, tense, etc.) were not generally semantically relevant. Fourth, it filtered, using a large list of 500+ stopwords, including standard stopwords and also some words typical of spontaneous speech. All of these were done for all segments, including all queries and all potential results.

The second phase compared the bag of words of the query and the bag of words of each possible result, to obtain a score for each. Using these scores the method picked the best segments, and returned their starting points as the jump-in set. We tried several methods for computing similarity (Garcia et al., 2013), and found that measures based on the number of words common to both bags of words worked well. These included, in particular, the dot product, when representing the segments as vectors of word frequencies:

$$\overrightarrow{q} \cdot \overrightarrow{s} = \sum_{i=1}^{|V|} q_i s_i \tag{4}$$

	searcher utility ratio test data
ASR transcripts, exact match	.34
ASR transcripts, soft match	.42
human transcripts, exact match	.41
human transcripts, soft match	.46

Table 5: Performance of the lexical models on the test set. Exact is with $kappa = 1.0$; soft with $kappa = 0.6$, that is, including the WordNet::Similarity.

where V is the vocabulary, and $\overrightarrow{q}$, $\overrightarrow{s}$ are the word frequency vectors representing the query and the segment respectively.

This measure can be expected to work well in terms of precision, as it ranks highly segments that share many words. In order to improve coverage, we augmented it with measures that take into account lexical and semantic generalizations. These were based on those in the software package WordNet::Similarity (Pedersen et al., 2004) that gives a measure of the semantic similarity and relatedness between any pair of words. Specifically, the Lesk measure did best; this reflects the overlap between the glosses of any two concepts, as well as concepts that are directly linked to them in WordNet.

We then defined a new measure, the cl6 soft match measure, based on a linear combination of the Lesk measure and the dot product.

$$\kappa \cdot (\overrightarrow{q} \cdot \overrightarrow{s}) + (1 - \kappa) \cdot \text{lesk}(q, s) \tag{5}$$

where q, s are the bags of word representing the query and candidate segment respectively. Best performance was obtained with $\kappa = 0.6$. Finally, only results expected to have a high probability of correctness were returned, this was done by limiting the number of results to the top 4, 5, or 7, depending on the model used and experience with the training data. These limits significantly contributed to performance, especially for the thesaurus-using models, where dozens or hundreds of results might be considered similar to some small degree, but it was found better to not return results beyond a certain point.

Table 5 shows the results. As expected, performance was better on the higher-quality transcripts, and better with the soft match. All the lexical models far outperformed the prosodic models.

9. Combined Models

Since the lexical and the prosodic models use different dimensions of similarity, a combined model should be able to do better than either alone.

The potential here is seen by the existence of queries where only the prosodic search yielded hits, for example for queries taken from the tagsets #hard-classes and #making-choices, with lexical search (using the human-generated transcripts) yielding none. Other

Base Model; Evaluation Data	lexical-only	lexical+prosodic
human, soft; training	.31	.32
ASR, exact; test	.28	.23
ASR, soft; test	.26	.22
human, exact; test	.30	.30
human, soft; test	.27	.30

Table 6: Performance for the Lexical and Combined Models

tagsets where prosody did better included #entertainment and #friends-and-family. Conversely, there were tagsets where lexical search did better, including #scholarships/research, #video-games, and #weekend-plans. There seemed to be a tendency for prosodic search to do better for similarity sets involving attitudes, feelings, and activity types, with lexical search doing better for narrower topics and those where common keywords can be expected.

The combination we chose was a simple rescoring, where we obtained probability estimates from the two models and combined them.

$$likelihood_c = P_p^{\lambda} P_l^{(1-\lambda)} \tag{6}$$

For this we first wrote equations to convert the raw scores of the two models into probability estimates, based on simple curve-fitting. Because for longer query regions the lexical overlap scores tended to be higher, these estimates were computed as a function of the size of the intersection divided by the duration, in seconds, of the query region. Second, we matched up the jump-in points proposed by the lexical model with those proposed by the prosodic model. For this we used all top lexical results, regardless of the list-length limits mentioned above. In essence, this was a retreat to a simpler evaluation measure, one that more directly judges rankings, which is more useful for gaining insight into the raw power of the models and their combinability (but which of course, relates less directly to users' needs, as it gives no credit for a model's ability to flag some results as probably not relevant).

Again here we needed an extra step, as the lexical models proposed regions but the prosodic model proposed timepoints. We tried a few simple methods, without seeing much variation in the performance. We settled on simplistically matching up each lexical-candidate region with the prosodic-candidate timepoint within that region, if any, with the highest probability estimate.

For the combined systems we used the prosodic model which performed best on the training data, as shown on the last line of Table 4, even though we knew that it was overfitted. Best results were obtained with $\lambda = 0.17$. On the training data, addition of prosodic information consistently improved the utility ratio.

Table 6 shows the results. The addition of prosodic information gave a benefit for one testset condition (last line), which was the one analogous to what it was trained for (first line), but not for the other conditions.

Re-examining our data, we noted a factor that may have contributed to this lack of

consistent benefit: the presence of more non-native English speakers in the test set than in the training set (75% versus 20%). While the non-natives had no major lexical or grammatical weaknesses, all had noticeably limited prosody, something often true of even advanced learners (Barraja-Rohan, 2011; Zimmerer et al., 2014), and this likely reduced the amount of information available to our method.

When present, the benefit was mostly from prosodic information boosting the probability estimate for lexically-proposed candidates that were in fact correct. Some additional benefit was seen for queries where lexical similarity proposed no candidates but some prosodic results were correct.

10. Conclusion and Future Work

We have shown that prosodic information is of value for search in dialog archives.

As future work there are several avenues likely to improve on our results. Beyond late fusion, other ways to combine prosodic and lexical similarity should be tried (Wollmer et al., 2013; Bruni et al., 2014). For example, recent developments in vector space representations of words (Turian et al., 2010; Erk, 2012; Mikolov et al., 2013; Huang et al., 2013), suggest that it could be productive to build a unified lexico-prosodic vector-space model of both meaning and dialog activity. It would also be valuable to seek a better fundamental understanding of how the prosodic and lexical aspects of dialog relate. This could lead to better ways to translate from estimates of similarities between timepoints to estimates of similarities between regions. Further study here could enable use of prosodic information not only for query-by-example, but also for searchbox interfaces, where prosodic information may be useful for query expansion (Chen et al., 2013). A more perceptually-valid measure of prosodic similarity could also be valuable (Reichel et al., 2009).

Future work might also explore ways to give users more control over prosody-based search, especially for situations where no lexical information is available. In particular, this could support query refinement, to enable searching on composite seeds, such as the average of two locations, the difference between two locations, and others based on relevance feedback. This would be especially useful for users repeating the same search tasks again and again, for example to create a well-refined detector for frustration as it appears in call-center dialogs, or humor as it appears in radio call-in programs each day.

Other applications might also be considered. We have only looked at cases where both the query and the possible targets are taken from dialog. However one might also use prosodic information for other purposes, including monolog search or search over multi-party interaction. It is even possible to imagine prosody supporting information retrieval using spoken queries (Oard, 2012), if searchers speak deliberately in the style of the content they want to find.

Finally, other languages are of interest. Like other query-by-example and unsupervised methods (Metze et al., 2014), dialog-activity search could be especially valuable for under-resourced languages. In particular, use of the city-block metric requires no knowledge and no labeled training data, only about 2 hours of unlabeled dialog to run through PCA to automatically derive the vector space. The specific dimensions of dialog activity are unlikely

to be universal, as are their mappings to prosodic features. However as long as the query example is in the same language and speaking style as the recordings to search, prosodically similar regions should still generally be pragmatically and topically similar.

References

Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T., Matsui, T., 2011. Overview of the IR for spoken documents task in NTCIR-9 workshop. In: Proceedings of the NII Test Collection for IR Systems Workshop. pp. 223–235.

Barraja-Rohan, A.-M., 2011. Using conversation analysis in the second language classroom to teach interactional competence. Language Teaching Research 15, 479–507.

Bruni, E., Tran, N.-K., Baroni, M., 2014. Multimodal distibutional semantics. Journal of Artificial Intelligence Research 49, 1–47.

Buckel, T., Thiesse, F., 2013. Predicting the disclosure of personal information on social networks: An empirical investigation. In: 11th International Conference on Wirtschaftsinformatik. pp. 1619–1633.

Bunt, H., 2011. Multifunctionality in dialogue. Computer Speech and Language 25, 222–245.

Chelba, C., Hazen, T. J., Saraclar, M., 2008. Retrieval and browsing of spoken content. IEEE Signal Processing Mag. 25, 39–49.

Chen, Y.-W., Chen, K.-Y., Wang, H.-M., Chen, B., 2013. Effective pseudo-relevance feedback for spoken document retrieval. In: ICASSP. pp. 8535–8539.

Erk, K., 2012. Vector space models of word meaning and phrase meaning: a survey. Language and Linguistics Compass 6, 635–653.

Eskevich, M., Aly, R., Ordelman, R. C. S., Jones, G. J. F., 2013. The search and hyperlinking task at MediaEval 2013. In: MediaEval Workshop.

Eskevich, M., Magdy, W., Jones, G., 2012. New metrics for meaningful evaluation of informally structured speech retrieval. Advances in Information Retrieval, 170–181.

Freedman, M., Baron, A., Punyakanok, V., Weischedel, R., 2011. Language use: what can it tell us? In: 49th Association for Computational Linguistics, Volume 2. pp. 341–345.

Galuščáková, P., Pecina, P., 2014. Experiments with segmentation strategies for passage retrieval in audio-visual documents. In: Proceedings of International Conference on Multimedia Retrieval. ACM, p. 217.

Galuscakova, P., Pecina, P., Hajic, J., 2012. Penalty functions for evaluation measures of unsegmented speech retrieval. In: Catarci, T., et al. (Eds.), CLEF: Information Access Evaluation. Springer, pp. 100–111.

Garcia, F., Sanchis, E., Calvo, M., Pla, F., Hurtado, L.-F., 2013. ELiRF at MediaEval 2013: Similar segments in social speech task. In: MediaEval Workshop.

Garofolo, J., Auzanne, C., Voorhees, E., 2000. The TREC spoken document retrieval track: A success story, NIST Special Publication 246, pages 107–130.

Godfrey, J. J., Holliman, E. C., McDaniel, J., 1992. Switchboard: Telephone speech corpus for research and development. In: Proceedings of ICASSP. pp. 517–520.

Hakkani-Tur, D., Tur, G., Stolcke, A., Shriberg, E. E., 1999. Combining words and prosody for information extraction from speech. In: Proc. Eurospeech, vol. 5. pp. 1991–1994.

Hanjalic, A., Kofler, C., Larson, M., 2012. Intent and its discontents: The user at the wheel of the online video search engine. In: ACM Multimedia.

Huang, C.-L., Hori, C., Kashioka, H., 2013. Semantic inference based on neural probabilistic language modeling for speech indexing. In: ICASSP (IEEE). pp. 8480–8484.

Jung, S., Na, S.-H., 2013. Refining sentence similarity with discourse information in dialog systems. In: Interspeech. pp. 3742–3745.

Kim, S., Yella, S. H., Valente, F., 2012. Automatic detection of conflict escalation in spoken conversation. In: Interspeech.

Larson, M., Eskevich, M., et al., 2011. Overview of MediaEval 2011 rich speech retrieval task and genre tagging task. In: MediaEval '11.

Larson, M., Jones, G. J. F., 2012. Spoken content retrieval: A survey of techniques and technologies. Foundations and Trends in Information Retrieval 5 (4-5), 235–422.

Levow, G.-A., 2013. UWCL at MediaEval 2013: Similar segments in social speech task. In: MediaEval Workshop.

Liu, B., Oard, D. W., 2006. One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In: 29th SIGIR. pp. 673–674.

Liu, Z., Huang, Q., 2000. Content-based indexing and retrieval-by-example in audio. In: IEEE Multimedia. pp. 877–880.

Lukowicz, P., Pentland, A. S., Ferscha, A., 2012. From context awareness to socially aware computing. Pervasive Computing, IEEE 11 (1), 32–41.

Mairesse, F., Poifroni, J., Di Fabbrizio, G., 2012. Can prosody inform sentiment analysis? Experiments on short spoken reviews. In: IEEE ICASSP.

Metze, F., Anguera, X., Barnard, E., Davel, M., Gravier, G., 2014. Language independent search in MediaEval's spoken web search task. Computer Speech & Language 28, 1066–1082.

Mikolov, T., Yih, W.-T., Zweig, G., 2013. Linguistic regularities in continuous space word representations. In: Proceedings of NAACL-HLT. pp. 746–751.

Mizuno, J., Ogata, J., Goto, M., 2008. A similar content retrieval method for podcast episodes. In: IEEE Spoken Language Technology Workshop. pp. 297–300.

Oard, D. W., 2012. Query by babbling: a research agenda. In: Proceedings of the first workshop on information and knowledge management for developing region. pp. 17–22.

Oertel, C., Scherer, S., Campbell, N., 2011. On the use of multimodal cues for the prediction of degrees of involvment in spontaneous conversation. In: Interspeech.

Pallotta, V., Seretan, V., Ailomaa, M., 2007. User requirements analysis for meeting information retrieval based on query elicitation. In: ACL. Vol. 45. pp. 1008–1015.

Pedersen, T., Patwardhan, S., Michelizzi, J., 2004. Measuring the relatedness of concepts. In: Proc. of the Nineteenth National Conference on Artificial Intelligence (AAAI-04). pp. 1024–1025.

Purver, M., Dowding, J., Niekrasz, J., Ehlen, P., Noorbaloochi, S., Peters, S., 2007. Detecting and summarizing action items in multi-party dialogue. In: Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue. pp. 200–211.

Reichel, U. D., Kleber, F., Winkelmann, R., 2009. Modelling similarity perception of intonation. In: Interspeech.

Rose, D. E., Levinson, D., 2004. Understanding user goals in web search. In: WWW '04: 13th International Conference on World Wide Web. pp. 13–19.

Slaney, M., Lifshits, Y., He, J., 2012. Optimal parameters for locality-sensitive hashing. Proceedings of the IEEE 100, 2604–2623.

Toivanen, J., Seppänen, T., 2002. Prosody-based search features in information retrieval. Proceedings of FONETIK 2002, 105–108.

Toutanova, K., Klein, D., Manning, C., , Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of HLT-NAACL. pp. 252–259.

Turian, J., Ratinov, L., Bengio, Y., 2010. Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 384–394.

Ward, N. G., 2014. Automatic discovery of simply-composable prosodic elements. In: Speech Prosody. pp. 915–919.

Ward, N. G., Vega, A., 2012. A bottom-up exploration of the dimensions of dialog state in spoken interaction. In: 13th Annual SIGdial Meeting on Discourse and Dialogue.

Ward, N. G., Werner, S. D., 2012. Thirty-two sample audio search tasks. Tech. Rep. UTEP-CS-12-39, University of Texas at El Paso.

Ward, N. G., Werner, S. D., 2013a. Data collection for the Similar Segments in Social Speech task. Tech. Rep. UTEP-CS-13-58, University of Texas at El Paso.

Ward, N. G., Werner, S. D., 2013b. Using dialog-activity similarity for spoken information retrieval. In:

Interspeech.
Ward, N. G., Werner, S. D., Garcia, F., Sanchis, E., submitted. The value of prosody-based similarity models for information retrieval. In: Interspeech 2014.
Ward, N. G., Werner, S. D., Novick, D. G., Kawahara, T., Shriberg, E. E., Morency, L.-P., Oertel, C., 2013. The similar segments in social speech task. In: MediaEval Workshop.
Werner, S. D., Ward, N. G., 2013. Evaluating prosody-based similarity models for information retrieval. In: MediaEval Workshop.
Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F., Singhal, A., 1999. Scan: Designing and evaluating user interfaces to support retrieval from speech archives. In: SIGIR. pp. 26–33.
Whittaker, S., Tucker, S., Swampillai, K., Laban, R., 2008. Design and evaluation of systems to support interaction capture and retrieval. Personal and Ubiquitous Computing 12 (3), 197–221.
Wollmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., Morency, L.-P., 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. Intelligent Systems, IEEE 28, 46–53.
Wrede, B., Shriberg, E., 2003. Spotting 'hot spots' in meetings: Human judgments and prosodic cues. In: Eurospeech. pp. 2805–2808.
Yuan, J., Liberman, M., Cieri, C., 2006. Towards an integrated understanding of speaking rate in conversation. In: ICSLP.
Zimmerer, F., Jugler, J., Andreeva, B., Mobius, B., Trouvain, J., 2014. Too cautious to vary more? A comparison of pitch variation in native and non-native productions of French and German speakers. In: Speech Prosody Conference.