Using Dialog-Activity Similarity for Spoken Information Retrieval

Nigel G. Ward, Steven D. Werner

Department of Computer Science, University of Texas at El Paso

nigelward@acm.org, stevenwerner@acm.org

Abstract

We want to enable users to locate desired information in spoken audio documents using not only the words, but also dialog activities. Following previous research, we infer this information from prosodic features, however, instead of retrieval by matching to a predefined finite set of activities, we estimate similarity using a vector space representation. Utterances close in this vector space are frequently similar not only pragmatically, but also topically. Using this we implemented a dialog-based query-by-example function and built it into an interface for use in combination with normal lexical search. In an experiment searchers used the new feature and considered it helpful, but only for some search tasks.

Index Terms: audio information retrieval, spoken content retrieval, prosody, vector-space model, query by example

1. Two Views of Audio Search

Most current spoken dialogue retrieval systems are based on the view that speech is essentially just noise-corrupted text [1]. 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 this approach.

We take a different perspective on audio search. In most cases users are probably not really interested in finding *words* in audio archives. What people want is often information of some type, characterized in part semantically, in part by intent [2], and in part by dialog process [3] or activity, for example recommending, answering a question, agreeing, forming a decision, telling life stories, making plans, hearing surprising statements, giving advice, attempting to persuade, explaining, and so on.

Dialog activities being numerous and complex, users are unlikely to be able to explicitly specify them in formulating a query. However they can be expected to recognize the sorts of things that they are interested in when they hear them, and thus can be expected to benefit from a "more like this" function [4, 5, 6] that returns results similar to a "seed," that is, an audio snippet used as a query.

This paper presents a model of dialog-activity similarity for search. Section 2 introduces our vector space representation, Section 3 discusses how proximity in this space provides a similarity metric, Section 4 describes an experimental investigation of the utility for search, Section 5 presents the results, and Sections 6 and 7 discuss and conclude.

2. A Vector-Space Representation of Dialog Activity

Following previous research, we chose to use prosodic features as the source of information about dialog activity. As is often noted, prosody can encode information that may not be expressed, or even expressible, by lexical means. Work using prosody in search goes back at least to the observation that important words and phrases can be prosodically distinctive and that this can be used to focus search [7]. Dialog activities that prosody can reveal include interactional "hotspots" where the speakers are unusually involved [8, 9], conflicts [10], agreements on action items [11], various emotional and attitudinal states and stances [12], and dialog acts of various types [13, 14], such as question, apology, promise, and persuade.

While this work has largely been motivated by search, findings to date have shown only that prosodic information can be used to detect such regions, not that this functionality is actually useful for searchers. This may be because the inadequacy of any finite taxonomy [15] of dialog activities for supporting most search needs. We employ instead an empirically-derived representation of dialog activities.

This representation, as described in [16], is obtained by applying Principal Component Analysis to 76 local prosodic features. While using the common features pitch height, pitch range, speaking rate, and volume, this feature set is novel in being computed 1) over different windows across six-seconds of context, thus capturing significant local context, 2) for both participants in the dialog, thus capturing both speaker and listener behavior, and 3) over fixed windows, rather than being word-, syllable-, or phrase-aligned, thus better capturing dialog-activity information. These features were computed every 10 milliseconds throughout the corpus. After PCA this gave 76 dimensions, ordered by how much of the variation they explained.

Upon examination [16, 17, 18], it turned out that most of the top dimensions aligned with various aspects of dialog. These aspects were diverse, including dialog situations, transient dialog states, cooperative dialog acts, simple dialog actions, and apparent mental states. For example, there were dimensions that related to grounding, to turn-taking, to seeking and expressing sympathy, to degrees of novelty and interest, to topic shifts and closings, to emphasis, to humor versus regret, to personal versus impersonal topics, to facts, to explanations, and to opinions. We can thus 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. Proximity and Similarity

Given this vector-space model we can use proximity as a measure of similarity. Specifically we use simple Euclidian distance, over all 76 dimensions, without weighting, although of course the top dimensions exhibit higher variance and therefore affect the results more strongly. As a preliminary exploration, we selected a few seeds and examined what positions the model found as most similar. As 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 two dialogs using ten family-related terms gave only 4 hits, 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 a property tax, gun laws, and public schools.

4. Experiment Design

To explore the utility of this similarity metric in audio search we designed an evaluation. In general, information retrieval can be evaluated using either user-centric or technology-centric methods. Standard technology-centric metrics such as recall, precision, and refinements thereof [1, 19, 20] are useful for comparing rival implementations of an approach that is already known to be useful, but less so for evaluating novel approaches. Therefore we adopted a user-centric method, one that directly measures actual value to searchers.

To enable a controlled experiment, we provided searchers with a set of tasks. Whereas in the examples above the seeds were found by listening to all the audio from the start, we provided the searchers with a lexical search function so they could find seeds faster. Thus, for example, if the task was to find talk about problems with pets, a searcher could first use lexical search, for example on words like *pet*, *dog*, and *cat,* to find possible seeds for dialog-based query-by-example searches.

We framed our hypotheses in terms of the advantages of dialog-activity search as an additional function, on top of traditional word-based search. Specifically, we hypothesized: A) that lexical-plus-dialog-activity search would yield more hits than lexical-only search, and B) that it would be preferred by and helpful to searchers.

4.1. Tasks and Data

Current standard evaluation infrastructures for audio search are mostly of two kinds: those that support evaluation only at the level of component technologies, such as lexical search or dialog-act search; and those which involve tasks, but which retrieve only "documents," such as news segments or Youtube videos [2, 13, 21, 22]. The exception only addresses search for factoids in monologs [23]. Wanting to do a task-based evaluation, and one that evaluates support for searchers to "jump in" directly to relevant utterances, we had to create our own evaluation suite. Considering 14 diverse audio search scenarios, we created a set of 32 tasks that could be tested on easily obtainable data but which were otherwise broadly representative [24].

Wanting a data set with a wide variety of dialog activities and topics, we chose Switchboard [25]. In this dataset the speakers were given suggested topics, but in practice talked mostly about whatever they wanted to. Although Switchboard is certainly not a universal corpus — it is exclusively telephone conversations; it is exclusively two-party dialog (although sometimes almost monolog in style); the subjects are aware they are being recorded and so tend to avoid self-identifying information; the subjects are all strangers and so the interpersonal dimensions of interaction are limited; and many common dialog activity types are not found — it is nevertheless quite varied.

We took a 4.5 hour subset of the Switchboard corpus and listened to decide which tasks to use. After excluding unsatisfiable searches, to avoid demoralizing the searchers, we selected 20 tasks, chosen to maximize diversity. These 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.

4.2. Interface

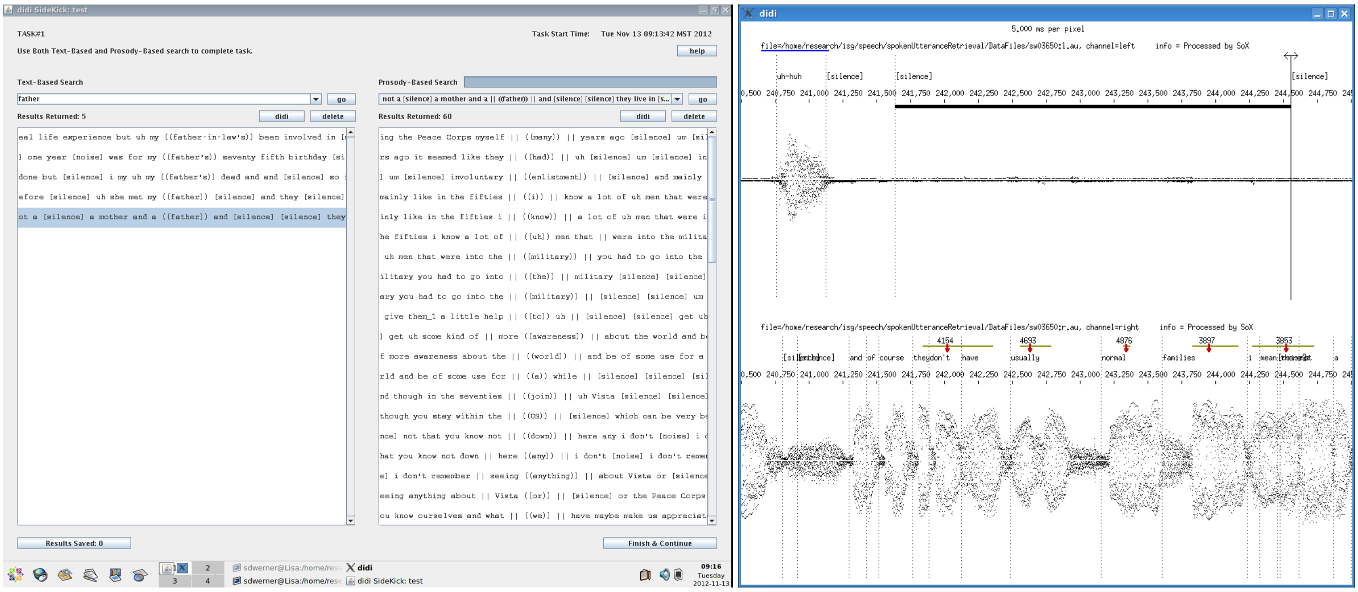
Best practice in interfaces for audio search includes the ability to jump to search results in the audio and the ability to listen to and navigate within the audio [26]. Our interface was accordingly built on top of a simple audio listening application [27] that displays the transcribed text and audio. This was augmented with a sidekick window to provide the search functions, as seen in Figure 1.

Lexical search is provided in the left side of the sidekick with an input box for the search terms and a display area for matches. Search is simple substring search, with no options for exact match, multi-term search, boolean search, etc. Each match is displayed as a few words of context centered on the matching word. Users can click on a match to jump to that position in the audio display, where they can hear the result and see the full transcript. After determining that a phrase or utterance is relevant, the searcher can click and sweep over those few seconds to specify the region and save it as a result.

For dialog-activity search, the user selects a point in the audio display as a seed and invokes the search command. Similar points are then found and presented for the user to peruse. Only points similar enough, being within the proximity threshold determined during the initial informal exploration, are shown; and nearby matches are grouped into regions, overlooking gaps of up to 50 millisecond, in order to avoid multiple fragmentary close results. To make the dialog-based search experience resemble that of word search, the matches are similarly displayed as a list. In the audio file currently being viewed, matches are also decorated with a red arrow at the highest-proximity point, the match score, and a horizontal bar over the entire region.

4.3. Subjects and Procedure

We hired four students as searchers. Each was given an initial explanation and an hour of training, including explanation of the new "prosodic search" feature and a practice search. They then did the twenty tasks, coming in to the lab at their conve-



didi SideKick: test

TASK#1

Task Start Time: Tue Nov 13 09:13:42 MST 2012

Use Both Text-Based and Prosody-Based search to complete task.

Text-Based Search: father

Results Returned: 5

```
eal life experience but uh my ((father-in-law's)) been involved in [s
] one year [noise] was for my ((father's)) seventy fifth birthday [si
done but [silence] i my uh my ((father's)) dead and and [silence] so i
efore [silence] uh she met my ((father)) [silence] and they [silence]
ot a [silence] a mother and a ((father)) and [silence] [silence] they
```

Results Saved: 0

Prosody-Based Search: not a [silence] a mother and a || ((father)) || and [silence] [silence] they live in [s...

Results Returned: 60

```
ing the Peace Corps myself || ((many)) || years ago [silence] um [si
rs ago it seemed like they || ((had)) || uh [silence] um [silence] in
] um [silence] involuntary || ((enlistment)) || [silence] and mainly
mainly like in the fifties || ((i)) || know a lot of uh men that were
inly like in the fifties i || ((know)) || a lot of uh men that were i
he fifties i know a lot of || ((uh)) men that || were into the milita
 uh men that were into the || ((military)) || you had to go into the
ilitary you had to go into || ((the)) || military [silence] [silence]
ary you had to go into the || ((military)) || [silence] [silence] um
 give them_1 a little help || ((to)) uh || [silence] [silence] get uh
] get uh some kind of || more ((awareness)) || about the world and be
f more awareness about the || ((world)) || and be of some use for a
rld and be of some use for || ((a)) while || [silence] [silence] [si
nd though in the seventies || ((join)) || uh Vista [silence] [silence]
though you stay within the || ((US)) || [silence] which can be very be
nce] not that you know not || ((down)) || here any i don't [noise] i d
hat you know not down || here ((any)) || i don't [noise] i don't remem
e] i don't remember || seeing ((anything)) || about Vista or [silence
eeing anything about || Vista ((or)) || [silence] or the Peace Corps
ou know ourselves and what || ((we)) || have maybe make us appreciat
```

Figure 1: User interface used in the experiments, with the audio listening tool at the right and the sidekick window at the left, and within that the lexical search in the left side and the dialog activity search on the right. Shown is a search through all 4.5 hours of dialog.

nience until all were complete. To encourage thoughtful sifting, they were given a small bonus for each answer found, less half the number that we later deemed irrelevant, but most of their compensation was by the hour. They were allocated 30 minutes for each task, but could move ahead to the next if they felt they had already found everything.

For half of the tasks the dialog-activity search function was turned off. Each searcher did two tasks in one condition before switching to the other. Two of the searchers started with lexical-only, and two with lexical plus dialog-activity.

After the searchers completed their work, we checked all hits for relevance according to the full description of each task. While almost all were clearly on-task, there were exceptions, including some clear misunderstandings and some lack of attention to the detailed task specifications. We accordingly excluded a total of 34 results: 8 from the lexical+prosody condition and 26 from the lexical-alone condition.

5. Results

Regarding hypothesis A, that the system version that included dialog-activity search would enable searchers to find more results: as seen in Table 1 there is a slight tendency for more results with the dialog-activity search, but the difference is far from significant.

Regarding hypothesis B, that the system version that included dialog-activity search would be preferred by searchers, there were three types of evidence, all supporting the hypotheses. First, in the free comments section of the final questionnaire all four searchers noted that it was beneficial, although with different nuances. 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),” another “Sometimes the prosodic search was necessary because it helped to find new results, not specifically the word on text-search,” the third “Lexical plus prosodic brought a different technique to search. If a limited number of results was returned using lexical, then searched prosodically a whole new set of results would be returned,” and finally the last searcher said “Prosodic search would find other results than lexical search found, plus more; with more work it could be better than lexical only.”

Second, the per-task questionnaires included an open-ended question about how the features available had helped or hindered. By design each searcher had dialog-activity search unavailable half the time, and all said, for at least 2 of these 10 tasks, that they would have liked to have had it available.

Third, although there was no requirement to use dialog-activity search, all searchers used it at least occasionally throughout the experiment; no one abandoned it.

To explore specifically when dialog-activity search was useful, we first looked at successes. Based on the logfiles, each result saved was attributed to one of three sources: dialog-activity search, lexical search, or browsing, depending on the most recent preceding search, and whether it had happened within the previous two minutes. Although rough, this indicated that 110 of the results were due to dialog-activity search. These were 40% of the 272 when dialog-activity search was available, indicating that it was frequently used and useful.

We then looked to see which specific tasks were best served by dialog-activity search. For this we considered the number of searchers who, in the written questionnaire for that task, volunteered a comment to the effect that they found prosodic search helpful for the results (when they had it) or that they would have liked to have it (when they didn't): these numbers are seen in the last column of Table 1. We also considered for each task whether or not searchers found more results when they had dialog-activity search available. For some of the tasks dialog-activity search clearly had value. For example, for task 6, finding a place where someone became tongue-tied, 3 of the 4 searchers thought dialog-activity search useful, and those who had it indeed found more results. On the other hand, there were three tasks where no searcher thought dialog-search helpful and having it was not advantageous. Generally the results were ambiguous: indeed, for many tasks one searcher liked this function but another explicitly said that it was not helpful.

task	lex- only results	lex+ dialog results	dialog- search likers
1 daytrip idea	6	4	
2 birthday gift idea	20	6	1
3 government complaints	10	17	
4 speaker's name	4	3	1
5 becoming at ease	24	7	2
6 being tongue-tied	7	14	3
7 pleasant laughter	18	56	
8 hobbies	16	14	
9 family structure	13	13	1
10 current place lived	18	14	2
11 places lived before	19	21	
12 education level	22	19	
13 speaker age	12	8	1
14 relative status	11	5	2
15 trusted media	16	17	
16 expensive purchases	2	6	
17 planned activities	5	6	2
18 teaching or explaining	10	12	
19 the word *nine*	14	14	
20 TI as employer	7	10	1
Total	254	266	

Table 1: Per-Task Results: Number of hits in each condition, and number of searchers expressing a desire for or an appreciation of prosodic search.

Finally, in order to better understand how dialog-activity search actually worked, we looked at the per-task distributions of results in dialog-activity space. Without doing a full investigation, we just sought dimensions important for each task. To do this we first found the average value on each dimension of all results for each task, averaged over all the result regions. We did this for all four searchers, and noted the dimensions for which the averages of all four fell in the same halfspace.

For task 9, 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 in [16], 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 very different tasks the distributions were also very different. For task 3, 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 task 10, 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. 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 do tend to fall in different regions of dialog-activity space, and using proximity is well founded.

6. Discussion and Future Work

While the results were mixed and the sample size too small to allow confident conclusions, overall the new method does appear to be valuable for some types of search. This is despite methodological limitations which suggest that the experimental results understate the true value. In particular: the searchers were hugely more familiar with lexical-only search; our implementation was slow (using linear search without indexing, rather something like a nearest-neighbors algorithm); the weight for the different dimensions were not tuned to correlate with perceived similarity; the search results were presented in straight temporal order, rather than ordered by match quality; and the searchers, unrealistically, had an accurate hand-labeled transcription of the audio [28] (although in most usage scenarios they would have only speech recognition output, which tends to be errorful; for example for conversational speech an error rate of 18% would be unusually good [29]). All of these limitations should be removed in future investigations.

Another direction to explore would be an under-the-hood combination of dialog-activity search with lexical search, with ranking based on both prosodic and lexical similarity to the seed, perhaps using lexical vector-space representations from information retrieval, semantic modeling, and/or language modeling [30, 31].

Yet another direction would be to add user-interface support for query refinement, in particular, for searching on composite seeds, such as the average of two locations, the difference between two locations, and other kinds of 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, say, frustration as it appears in call-center dialogs, or humor as it appears in radio call-in programs each day.

Another scenario in which dialog-activity search could be valuable would be for under-resourced languages, since it requires no knowledge, apart from that automatically derivable by applying PCA to the prosodic features extracted over about 2 hours of sample dialog to generate the vector space. While the specific dimensions are unlikely to be universal, as are their mappings to prosodic features, as long as the query example is in the same language and speaking style as the recordings to search, proximal positions should still be pragmatically and often topically similar. We plan to test this.

7. Summary

We have shown that query-by-example search, using a prosody-based vector-space representation of dialog acts, often returns places matching the query in dialog activity and/or topic, and that this can be a useful and user-appreciated addition to standard word-based search. This new method may find uses in search of workplace recordings, surveillance recordings, personal recordings, multimedia as social media [32], and so on.

8. Acknowledgements

We thank Olac Fuentes, David Novick, Alejandro Vega, Shreyas Karkhedkar, and Mary Harper for ideas, comments, and assistance, and the National Science Foundation for a REU supplement to IIS-0914868.

9. References

[1] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Mag.*, vol. 25, pp. 39–49, 2008.

[2] A. Hanjalic, C. Kofler, and M. Larson, "Intent and its discontents: The user at the wheel of the online video search engine," in *ACM Multimedia*, 2012.

[3] V. Pallotta, V. Seretan, and M. Ailomaa, "User requirements analysis for meeting information retrieval based on query elicitation," in *ACL*, vol. 45, 2007, pp. 1008–1015.

[4] Z. Liu and Q. Huang, "Content-based indexing and retrieval-by-example in audio," in *IEEE Multimedia*, 2000, pp. 877–880.

[5] J. Mizuno, J. Ogata, and M. Goto, "A similar content retrieval method for podcast episodes," in *IEEE Spoken Language Technology Workshop*, 2008, pp. 297–300.

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

[7] D. Hakkani-Tur, G. Tur, A. Stolcke, and E. E. Shriberg, "Combining words and prosody for information extraction from speech," in *Proc. Eurospeech, vol. 5*, 1999, pp. 1991–1994.

[8] B. Wrede and E. Shriberg, "Spotting 'hot spots' in meetings: Human judgments and prosodic cues," in *Eurospeech*, 2003, pp. 2805–2808.

[9] C. Oertel, S. Scherer, and N. Campbell, "On the use of multimodal cues for the prediction of degrees of involvment in spontaneous conversation," in *Interspeech*, 2011.

[10] S. Kim, S. H. Yella, and F. Valente, "Automatic detection of conflict escalation in spoken conversation," in *Interspeech*, 2012.

[11] M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, S. Noorbaloochi, and S. Peters, "Detecting and summarizing action items in multiparty dialogue," in *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, 2007, pp. 200–211.

[12] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: harvesting opinions from the web," in *ICMI: 13th International Conference on Multimodal Interfaces*, 2011, pp. 169–176.

[13] M. Larson, M. Eskevich *et al.*, "Overview of MediaEval 2011 rich speech retrieval task and genre tagging task," in *MediaEval '11*, 2011.

[14] M. Freedman, A. Baron, V. Punyakanok, and R. Weischedel, "Language use: what can it tell us?" in *49th Annual Meeting of the Association for Computational Linguistics, Volume 2*, 2011, pp. 341–345.

[15] P. Lukowicz, A. S. Pentland, and A. Ferscha, "From context awareness to socially aware computing," *Pervasive Computing, IEEE*, vol. 11, no. 1, pp. 32–41, 2012.

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

[17] ——, "Towards empirical dialog-state modeling and its use in language modeling," in *Interspeech*, 2012.

[18] N. G. Ward and K. A. Richart, "Lexical and prosodic indicators of importance in spoken dialog," in *14th Annual SIGdial Meeting on Discourse and Dialogue, submitted*, 2013.

[19] B. Liu and D. W. Oard, "One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech," in *29th SIGIR*, 2006, pp. 673–674.

[20] M. Eskevich, W. Magdy, and G. Jones, "New metrics for meaningful evaluation of informally structured speech retrieval," *Advances in Information Retrieval*, pp. 170–181, 2012.

[21] J. Garofolo, C. Auzanne, and E. Voorhees, "The Trec spoken document retrieval track: A success story," 2000, NIST Special Publication 246, pages 107–130.

[22] M. Larson and G. J. F. Jones, "Spoken content retrieval: A survey of techniques and technologies," *Foundations and Trends in Information Retrieval*, vol. 5, no. 4-5, pp. 235–422, 2012.

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

[24] N. G. Ward and S. D. Werner, "Thirty-two sample audio search tasks," University of Texas at El Paso, Department of Computer Science, Tech. Rep. UTEP-CS-12-39, 2012.

[25] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of ICASSP*, 1992, pp. 517–520.

[26] S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal, "Scan: Designing and evluating user interfaces to support retrieval from speech archives," in *SIGIR*, 1999, pp. 26–33.

[27] N. Ward, "Didi, a dialog display and labeling tool," 2003, http://www.cs.utep.edu/nigel/didi/.

[28] ISIP, "Manually corrected Switchboard word alignments," 2003, Mississippi State University. Retrieved 2007 from http://www.ece.msstate.edu/research/isip/projects/switchboard/.

[29] D. Yu, F. Seide, and G. Li, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. 29th Int'l Conf Machine Learning*, 2012.

[30] K. Erk, "Vector space models of word meaning and phrase meaning: a survey," *Language and Linguistics Compass*, vol. 6, pp. 635–653, 2012.

[31] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[32] N. G. Ward, S. D. Werner, D. G. Novick, T. Kawahara, E. E. Shriberg, L.-P. Morency, and C. Oertel, "The similar segments in social speech task," in *MediaEval Workshop*, 2013.