11-1-2012

Thirty-Two Sample Audio Search Tasks

Nigel G. Ward
University of Texas at El Paso, nigel@utep.edu

Steven D. Werner
University of Texas at El Paso, sdwerner@miners.utep.edu

Comments:
Technical Report: UTEP-CS-12-39

UTEP Technical Report UTEP-CS-12-39

Thirty-Two Sample Audio Search Tasks

Nigel G. Ward and Steven D. Werner

Department of Computer Science
University of Texas at El Paso
500 West University Avenue
El Paso, TX 79968-0518

email: nigelward@acm.org, stevenwerner@acm.org

November 19, 2012

Searching in audio archives is potentially very useful, and good evaluations can guide development to realize that promise. However most current evaluation programs are technology-centric, rather than user-oriented and task-centric. This paper examines current and potential audio search needs and scenarios, and presents a sample set of thirty-two diverse audio search tasks to support more realistic evaluations.

Index Terms: information retrieval, spoken content retrieval, dialog

1 Motivation

Audio search algorithms and systems are today usually evaluated at the technology level. For example, tools based on speech recognition a system can be evaluated on their ability to retrieve all and only the locations of occurrence of a query word. Although well-suited to comparing the performance of two rival teams proposing incremental advances, such evaluations can be limiting.

Ultimately it's not the quality of the technology that matters; for example no one really cares about finding *words* in audio archives. What people want is generally information or insight. Evaluating audio search with respect to such ultimate goals is important but not easy.

One extreme approach is to leave the evaluation entirely up to users. They could formulate their own goals and report their own satisfaction with their ability to search, across various systems. However judgments of satisfaction on self-selected tasks will vary widely and be sensitive to factors extraneous to the search itself, such as motivation and interest.

Accordingly there is a need for evaluations that enable controlled experimentation, and yet approximate real end-to-end search in support of realistic tasks. An essential component of any such evaluation will be a set of tasks. In this report Section 2 discusses the nature of speech versus text, noting that there is more to audio than words, Section 3 explains the importance of evaluating search based on tasks, not queries, Section 4 overviews some types of audio archive and the associated types of search needs, Section 5 highlights some important dimensions of variation in audio search, and Section 6 presents some practical considerations in the creation of a specific evaluation set. Based on this analysis, Section 7 presents our set of 32 diverse search tasks, and Section 8 discusses their feasibility for use in evaluations on searches over a subset of the Switchboard corpus.

2 The Nature of Speech

This section discusses two different perspectives on searching in audio: as a problem of working with a corrupted form of text or with a record of a human interaction.

One perspective is that audio is fundamentally just a nuisance, that ideally all information would exist as text, and that the purpose of audio search technology is to minimize this pain of dealing with audio. On this view audio is just noise-corrupted form of text, where some of the words are wrong, and there is no "audio search" problem as such; rather there are just two general problems — speech recognition and search in texts — which together can solve this problem [Chelba et al., 2008]. This implies that audio search involves first running a speech recognizer over the input, and then having the searcher work primarily with the resulting transcripts, with occasional reference to the actual audio.

Good transcripts include not only the words in sequence, but also capitalization, punctuation, and boundaries between topics or segments. These are lacking from raw speech recognition output, and in recent years each of these issues has become a popular research topic, often addressed with the application of prosodic information [Hakkani-Tur et al., 1999, Shriberg, 2005]. Prosody can also be used to help the speech recognizer do its job better [Kompe, 1997, Shriberg and Stolcke, 2004, Ward et al., 2011], and turn audio search into a less-impaired version of textual search (although the value of prosody for speech recognition has not so far been high).

Within this general perspective there are varied approaches. One important dimension is when to do the transcription work: ahead of time or just in time, running a wordspotter over the audio in response to a query. Another important dimension is whether to have the transcript be the main product for the searchers, or whether to go beyond that to richly index it, for example by topic [Mishne et al., 2005], or even to extract pieces of information [Diao et al., 2010]. Another interesting question is whether the searcher should specify the words to be searched for by inputting words as orthographic objects or as sound objects [Metze et al., 2012, Oard, 2012]. Across all these variants, however, the word is taken to be the focus of search.

The other perspective is 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 choices about how to communicate, with texting increasingly popular. Special properties of spoken dialog include its utility for establishing rapport, for allowing self-expression, for dialog activities that involve emotion or interpersonal interactions (such as persuading, apologizing, justifying, explaining preferences, and reaching decisions), for conveying and appreciating personality, and for talking about personal and private matters. (In passing, it's worth noting that prosody and the dynamics of interaction are a large part of the underpinnings of the ways people do these things, and that these properties are characteristic of dialog rather than monolog.) From this perspective, it seems worth exploiting such aspects of speech for audio search.

Fully exploiting this perspective is a long-term project, but it has already inspired a handful of explorations, starting with the observation that important words and phrases can be prosodically distinctive and that we may be able focus search on such important places [Hakkani-Tur et al., 1999]. Other things we may be to identify are interactional “hotspots” where the speakers are unusually involved [Wrede and Shriberg, 2003, Oertel et al., 2011], conflicts [Kim et al., 2012], decision points, various emotional states, and dialog acts of various types [Larson et al., 2011], such as question, apology, and promise. (Research based on this perspective shares much with sister fields that try to extract more information from speech, including social signal processing [Pentland, 2008, Vinciarelli et al., 2009, Jurafsky et al., 2013], sentiment mining [Mairesse et al., 2012], personality inference [Mohammadi and Vinciarelli, 2012], diagnosis of autism and depression, and so on.)

Unfortunately the types of events and states to be retrievable are usually defined *a priori*, and these techniques are evaluated based on their ability to find these dialog acts; again a technology-centric evaluation. We want to know in addition when and how useful such features might be for satisfying user needs, and more generally, whether the use of dialog-specific or other novel techniques can be valuable, in contrast to or in concert with lexically-based search. Thus the advancement of audio search requires better evaluations.

3 The Nature of Search Needs

This section considers the relationship between queries and underlying search needs.

Most empirical study of audio search needs and types has focused on queries at the document level, for example for retrieving Youtube videos [Larson et al., 2012, Hanjalic et al., 2012]. Although the findings are not entirely relevant to search intended to find specific information at the utterance level, there are some general lessons to learn. These results of these studies, aligning with findings about web search [Rose and Levinson, 2004], show that search for specific facts, although often thought of as the typical search task, in fact accounts for less than half of searches, with other common types including background search (for example to gather information to make a decision) and more unfocused browsing (for example to be entertained or to stay informed). The same is probably true, or potentially true, for audio search, so an evaluation set should include also non-factoid search tasks.

Another lesson to take from the general information-retrieval literature is that queries are not adequate as search-task categorizations. Concisely specified searches, in the extreme stated with a single search term, are popular and often thought of as typical. Conciseness is a virtue, but over-concise queries are a problem. This is well-known in the web search world, where users frequently use single-term queries, although this makes it hard for the search engine to know what they’re looking for. In the web search world, enormous effort has gone into techniques for inferring actual user goals, and for presenting a diverse sampling of different types of web pages to provide coverage for all likely goals. As a result users are spoiled, and we have come to think of this as a normal search scenario, although it is anything but.

For evaluation purposes, therefore, it is preferable to describe the search tasks in terms of underlying goals, the actual user need, and to evaluate based on the ability to satisfy that need. For example, *“a query for XJQ-17”* would be less suitable for evaluation than *“the intention to discover whether a rumored new material called XJQ-17 really exists, and if so what it is, who knows about it, and how far it is from commercialization.”* In sum, task sets for evaluation should be designed around fully elaborated search goals, not single-term queries.

4 Audio Search Scenarios

Getting more concrete now, this section presents a rough overview of the types of audio archives and some likely common search scenarios, building on the list in [Larson and Jones, 2012]. As audio search is not common today, and some times of recordings are rare probably because there is no ability to search on them, we do not limit ourselves to known scenarios, but include speculative uses.

Some types of archive are already common:

Surveillance Recordings Intelligence agencies have an interest in finding information in wire-tapped telephone conversations etc. The search needs here probably include: cutting through the voluminous irrelevant talk to find discussions of past or planned nefarious activities; compiling lists of people or events; gathering factual, behavioral, and personality information to form a profile of a person; mapping out social roles and networks, gauging individual and group opinions, moods and proclivities; and so on.

Broadcast News Intelligence agencies also have an interest in monitoring news from sources around the world. As most news is scripted, such audio is truly just noise-corrupted text [Garofolo et al., 2000]. Nevertheless, when the scripts are not available, audio transcription and/or search is necessary.

Customer-Service Interaction Recordings Many call-center telephone conversations are monitored or recorded for audit and quality-control purposes. This typically involves supervisors reviewing calls to monitor agents' attitudes and performance, looking for problems that need correction. In the call center world, there are semi-automated "Speech Analytics" tools for this. A (potential) application for advanced analytics systems is the discovery of "business intelligence" such as the discovery of recurring issues and new trends in customer desires and behavior.

Lectures Classroom lectures and other information-rich presentations are sometimes recorded; and some aspects of the search needs here have been examined [Stifelman et al., 2001].

Legislative and Court Proceedings Law and custom often require depositions, testimony, deliberations and so on to be transcribed, at great expense, usually by humans but with some exceptions [Kawahara, 2012]. If recordings could be made searchable, this cost could perhaps be avoided. The common search needs are, however, not clear, except for finding matter relating to specific topics and the classic example of seeking contradictory statements made by the same person at different times.

Performances A system for searching political speeches was deployed experimentally by Google in 2008. Search in entertainment datasets (for example to find the funniest bits in recent talk shows) might be well appreciated.

Interviews Some aspects of the search needs of professional historians in oral histories have been considered [Oard et al., 2004], and such recordings might be more generally useful, if more kinds of search were supported.

Amateur Content Amateur product or restaurant reviews can be mined for sentiment, positive or negative [Morency et al., 2011, Mairesse et al., 2012], among other things.

There are also some types of audio archive that may in future become common:

Personal Recordings People who record their own interactions may want to search them. Goals may including: finding and reliving enjoyable moments, finding evidence (*"I was right, you did say that!"*), retrieving half-forgotten information, and diagnosis of one's own social flaws [Sellen and Whittaker, 2010, Vemuri et al., 2004].

Family Recordings Photos, videos, and other keepsakes are becoming digital, and audio may go the same way [Petrelli et al., 2010]. Imagine that a child wishes to create an audio memory book about a deceased aunt. To create it he might want to search in recordings of family gatherings for examples of what made her special: her humor, her memories of the war, her obsessive talk about her cats' personalities, her attempts to life-coach her nieces and nephews, and so on.

Customer-Service Recordings Beyond call-center interactions, many other types of service encounters could be recorded and archived. If this were done, for example, for healthcare interactions, one would then be able to search to answer queries such as: did the doctor ever caution him about taking the medicine with alcohol? what words did the patient use to describe the chest pain he felt last month? where did the patient say her daughter lived? Today some aspects of doctor-patient interactions are laboriously recorded textually, and the rest are lost, but if all were instead recorded as audio and searchable, it could be possible to both reduce the up-front cost and make more information available.

Meeting Recordings A few years ago there was much research interest in supporting retrieval from recorded business meetings. The use cases were never really clear (perhaps because in most meetings the important things, those that need to be referred to later, are generally noted down, so the value proposition for search is not obvious), and indeed uptake of the technology has been very limited [Whittaker et al., 2008]. (Indeed, research on meeting corpora has largely moved on to other tasks, such as diarization, summary generation, and social role analysis.) However there are meetings for which notes are not available, namely surreptitiously recorded meetings; thus meetings recorded with visible microphones can be a proxy for the sort of dialogs that electronic bugs may pick up and intelligence agencies may want to exploit.

Other Workplace Recordings Today interactions among coworkers are not generally recorded, except for cockpit recordings and police communications, which can be used for forensic purposes after something goes wrong, and voicemails [Whittaker et al., 1999], a form of monolog which appears to be dying out in favor of email or texting. Nevertheless there is potentially much useful information in workplace interactions.

Social Recordings We can imagine a club or organization creating a set of recordings among its members, for subsequent use by new members as a quick way to familiarize themselves with group norms and know-how, as well as to answer specific questions. For example, in an academic computer science department one might collect dialogs between students and expect many of the topics in such dialogs to be of interest to others: newcomers might search for information about, for example, internship experiences, linux installation options, what classes to take, how to balance school with work and family obligations, thoughts about applying to graduate schools, and so on, among more general topics of interest to young people such as good local bands, parking-ticket avoidance strategies, and where to eat.

5 Characteristics of Audio Search Situations

This section briefly notes some other dimensions of variation, orthogonally to the classification of archive types above.

First, there are differences among properties of the recordings themselves, likely in turn to affect search needs, behavior, and results. Most significant is probably the number of people involved, with the key distinction being between dialog and monolog, as suggested earlier. Other aspects include whether the interaction was semi-scripted versus spontaneous, whether the language was controlled or semi-standardized versus free, whether video was present or absent, and how much meta-data is available.

Second, properties of the user/searcher population are important, including whether they are experienced or novices, searching for their job or for their own purposes; whether they are searching on their own voice, on those of people they know, or on those of strangers; and whether they are seeking factual information, seeking to learn something about the speaker, or wanting something else.

6 Evaluation Set Development

Based on the above analysis, we set out to develop a set of search tasks for evaluation. We had several additional considerations:

Our primary consideration was breadth; we wanted to cover as many as possible of the types of search, as surveyed above. However we made no attempt to make the set representative; there is no attempt to include more tasks for search types that are more common, since we don't know which tasks are more common in any specific application. Therefore one would not, for example, want to use average success over all the tasks as a quality metric, rather this evaluation set is suited for exploring which sorts of searches new methods can perform well on.

Practicality, in the sense of supporting affordable evaluations, was also an important consideration, in three ways. First, we wanted to be able to run tests using casual labor, without requiring the involvement of searchers with special knowledge, skills, and experiences. However most people are good at imagining themselves in different roles, and able to search as if they were in different situations and for various purposes, so this was not extremely limiting.

Second we wanted to reduce overhead time by choosing tasks that could be satisfied using a single data collection, to reduce the time cost of searchers familiarizing themselves with new genres of data. To our knowledge, the broadest single type of dialog is free, unstructured conversations. The specific corpus with which we have the most experience is Switchboard, which is also widely available [Godfrey et al., 1992]. We therefore chose search tasks that we thought would work well with this corpus. Only one task, task 32 below, is truly Switchboard-specific, but this and the other tasks could be easily generalized or specialized to work better for other corpora. Switchboard is unusual in that the speakers were given topics to discuss, but in practice most of the speakers talked mostly about whatever they wanted to, ranging widely over topics and dialog styles. Switchboard has some other properties worth mentioning: it is telephone-based, not face-to-face; it is not task oriented; it is exclusively two-party dialog (although with some segments essentially monolog in style); as the subjects are aware they are being recorded, they tend to avoid self-identifying information; there are no repeated same-pair conversations and so is no past context and no expectation of future interactions; and the subjects are all strangers and so the interpersonal dimensions of interaction are limited, with many dialog activity types missing, including planning, persuading, and decision-making.

Third, we wanted to make it easy to recruit and retain workers. In particular, we didn't wish to make searchers feel they were engaged in immoral activities, such as violating privacy, so such tasks are minimized, and any tasks which seek personal information about the speakers are framed as part of scenarios in which the information found is to be used in beneficial ways. Of course such tasks can serve as proxies for less innocent searches, and indeed all our tasks except 1-4, 6-8, and 10-11 have analogs in intelligence-type searches.

Our final consideration was one of avoiding bias; we didn't want to have the tasks slanted in some way that would intrinsically make some search techniques look bad and others good. In the interest of full disclosure, we must mention that before finalizing the list of 32, we informally tried out a new dialog-based search technique on two, namely family structure and complaints about the government, for both of which it seemed to work better than word-based search. However we chose not to therefore exclude these tasks, since we don't know whether any other search technique would do equally well or better for them.

7 The Thirty-Two Search Tasks

Without further ado, here is our set of 32 diverse search tasks. For the reasons given above, each is presented together with a context and a justification.

7.1 Social Network Dialog Search

Imagine that a new rival to Facebook has introduced a service that features recordings of telephone conversations among friends, and friends of friends, and so on. If you join this service, you get discounted cell phone service, in exchange for agreeing to have at least 100 minutes a month of your conversations recorded and made available to your online friends. (Of course you get to exclude any calls you like from being recorded.)

Members of this service then have the ability to search though conversations by friends and friends of friends. The downside is that these are completely unorganized and often confusing, but the good side is that they are by people you (mostly) trust, and the topics are frequently things that interest you.

1. You are planning a little trip out of town with your mother. You'd like to find references to interesting places to visit, together with descriptions of fun things to do there, potential downsides, and explicit comparisons to other places to visit. You're thinking mostly about a long daytrip or at most an overnighter. You're not into any specific sport, so a conversation about someone's hangliding weekend wouldn't be too interesting. On the other hand a discussion of an arts and crafts festival, or a music festival, or a scenic town with a nice harbor tour would be. It's okay if the

talk is about a place distant from you, as long as it gives you ideas for what you might do locally. For example, a conversation about taking a helicopter over the mountains in Switzerland would be less interesting than a conversation about a two-hour hike in the hill. Similarly, given your mother's age and your budget, talk about skiing, rock concerts, opera, would be less interesting.

2. Your birthday is coming up, and you're interested in finding things that sound fun to have, to add to your wishlist. You'd like to find references to things that people talk about fondly, whether cameras, phones, magazines, plants, kitchen appliances, hats, and so on. Music recommendations would be interesting too. You're not really interested in food, and you know no one's going to buy you anything expensive like a car or a house, and you definitely don't want a pet, but almost anything else that you can buy, and that at least one person likes, would be interesting to hear about and consider.

3. You're looking for jokes and funny stories to amuse yourself, especially concise ones that you could retell at a party.

4. You want to convince a relative that she doesn't want a pet. You want to find talk about people having problems with pets, or to the difficulties and costs of caring for a pet, and so on.

5. You are thinking about forming a group of people to talk about problems in government, and maybe about ways to address them. As a first step, you want to find all audio segments where people are complaining about any aspect of government, at any level. At this stage you're interested in learning what problems people see, and what sorts of things upset them most. You're even interested in hearing complaints by cranks who are bitter and believe that no government can ever be good.

6. You are organizing an event and inviting lots of people, most of whom you've never met in person. In preparation, you would like to learn the pronunciation of each person's name. Find all instances in conversation of people saying their own name. (Hint, most people will say something like "Hi, I'm Martha" early in a conversation.)

7.2 Search in Self-Recorded Dialogs

Imagine that recent changes to the law have made it legal to record all your conversations unless the other person objects.

7. Suppose that you've decided to become a better communicator, and in particular to work on being able to put people at ease. You want to go through your recent conversations with strangers and find, in each conversation, the first point in which the other person seemed to relax and become comfortable.

8. Now try the opposite: this time focusing on all places where your interlocutor seemed to feel uncomfortable or otherwise momentarily stuck for how to respond or what to say.

9. One day you friend tells you she heard some really interesting historical fact, that she intended to tell you about, but now she can't remember what it was. Being bored, you decide you'd like to find it, and she gives you access to the archive of all her recent conversations (excluding some sensitive ones, of course). To track it down systematically, you set out to find all references to historical people and events, of the sort that are in the history books; that is, no myths or legends, no stories of great grandparents, and no discussion of just recently retired politicians.

7.3 Customer Service Monitoring

Imagine that you are a manager at a call center that records all customer-agent interactions.

10. You want to find all instances of an agent being inattentive during a conversation.

11. You want to find all instances of an agent getting tongue-tied or otherwise behaving disfluently.

12. Knowing that establishing rapport with the customer is a good thing, and knowing that one hallmark of trust and rapport is pleasant laughter, you want to find all points where the customer is laughing in a nice way.

7.4 Statistical Market Research

Imagine that you work for a market research company which has obtained thousands of minutes of normal telephone dialogs among normal consumers (perfectly legally, by offering them a free cell phone and premium Skype use for a year in exchange). Unlike most market research companies, yours does not require participants in its consumer programs to fill out lengthy forms. Instead they simply infer that information from the phone calls.

For each individual:

13. Find at least one hobby they have or used to have.

14. Find out how many people are in his/her household, including how many adults and how many children.

15. Find the location they are currently living (city and state, if possible).

16. Find at least one location where they lived in the past. City names are best, but any information can be included, for example, “we lived abroad when I was young” or “I grew up in a small town.”

17. Find any references, direct or indirect, to the education level of the speaker.

18. Find any mention, exact or approximate, of the age of the speaker.

19. Find all segments revealing employment status, such as working, not working, in school, retired, busy raising children, etc. Indirect statements are of interest (“every day I'm busy at home with the kids”) if they seem intended to indicate employment status. Only the first mentions are of interest; for example a statement that merely refers to an already-known employment status (for example “my boss's wife once had that disease” or “I listen to NPR on the way to work”) is not of interest.

7.5 Social Network Analysis

Imagine that your company has now decided to try to determining the relations between the people in a group, and between people in the group and external influences.

20. In many conversations, even among strangers, before long the participants figure out which one is “higher status.” In each conversation, find one segment which indicates somehow

that one participant is “higher status” (for example, older, wiser, bossier, more managerial, or otherwise more dominant) than the other. There may be segments where one participant appears to be claiming higher status but the other is not necessarily accepting that: these are interesting too.

21. Find how this group of people might be informed or influenced. Specifically, find sources of information that are considered trustworthy by (some of the) people in this group. These may include magazines, websites, blogs, television shows, radio stations, political figures, and clergy. Do not include friends and relatives, no matter how trusted. Of course people disagree on what’s trustworthy, but just identify all segments where a person is expressing trust in a source.

7.6 Individualized Marketing

Imagine that you are now working for a marketing organization. Understanding peoples’ needs helps make it possible to reach them more effectively, for example for deciding who might like to receive discount coupons for specific things — investment opportunities, retirement planning, home refinancing, vacation destinations, magazines, insurance, concerts, and so on — and only sending coupons to people likely to be interested. Alternatively this might help in forming profiles used to customize what advertisements appear when they browse the web.

22. Find any mentions of major recent expenditures, for example for a house, a car, education, medical expenses, or an exotic vacation. Distant past expenses are not of interest, nor are planned or possible future expenses. Savings and investments are not of interest, nor are tax payments.

23. Find any mention of the current local weather at the time of the call (as a way of deciding what discount vacation coupons they might like to receive).

24. Find all segments in which the speaker mentions planned activities, or possible activities, for example, planning to go to a restaurant tomorrow, to go to the mall on the weekend, to do some gardening when it gets warmer, to go to a concert next moth, and so on. Possible, habitual, vague or distant plans are not of interest (e.g. “my wife wants to try out the new winery”, “every morning I go jogging,” “I want to try horseback riding someday,” or “I plan to retire to Florida”).

7.7 Modeling Individuals

Imagine that job applicants start submitting “audio dialog portfolios” of themselves to prospective employers, and that employers want to mine this data to build psychological profiles for each person they decide to interview.

25. Find all mentions of rule breaking, no matter how major or minor (remodeling without a building permit, speeding, littering, etc.). Include cases where the speaker did something questionable or was was accused of something, even if they don’t think they did anything wrong.

26. Find all mentions of scary or dangerous experiences.

27. Find one time per speaker where he or she tries to teach the interlocutor something, or explain something. Only consider specific knowledge; general wisdom or prejudices, such as “always wear a smile,” or “never trust teenage boys,” is not of interest.

28. Find all places where a speaker uses a foreign words somewhere in a dialog. Well-known place names don’t count (“Paris”) but specific places if pronounced in native style do (“le Bois de

Boulogne"). Words and phrases commonly used in English don't count ("tortilla", "croissant", "deja vu"), but if the speaker a word pronounces as if it were foreign to him and/or unfamiliar to his listener (such as perhaps, "creche", "elan", "parvenu", "oeuvre"), then it should be included.

29. Find out how much a speaker knows about China: over all his conversations, find all mentions of China or things Chinese.

7.8 Other

Finally, imagine that you are working for a boss who just gives you some queries, with no real justification or explanation.

30. "Find me all sentences including the word *bank.* I don't care if it's river banks or money banks, just that word. Oh, and include *banks and banking*, but not *banked* or *banker* or anything else."

31. "Find me all occurrences of the word *nine.*"

32. "Find me all segments discussing working for Texas Instruments, as either a current or a potential employer."

8 Feasibility

After drawing up this list, we took a 4.5 hour subset of the Switchboard corpus and listened to discover which of these tasks were suitable. Specifically, we counted the number of regions in this subset which served each search task, and thus should count as hits if returned by a search algorithm. While there is no in-principle reason to exclude unsatisfiable searches from an evaluation, in practice such can be demoralizing to human searchers, and so we wanted to weed those out.

After doing so we had 24 candidate tasks remaining, and from these we selecting out a nice round number, 20, chosen to maximize diversity. Specifically, these are tasks 1-2, 5-7, 11-18, 20-22, 24, 27 and 31-32. We are currently using this subset to measure performance of a new search method [Ward and Werner, 2013].

Acknowledgment

We thank Martha Larson for comments. This work was supported in part by the NSF by a REU supplement to project number IIS-0914868.

References

[Chelba et al., 2008] Chelba, C., Hazen, T. J., and Saraclar, M. (2008). Retrieval and browsing of spoken content. *IEEE Signal Processing Mag.*, 25:39–49.

[Diao et al., 2010] Diao, M., Mukherjea, S., Rajput, N., and Srivastava, K. (2010). Faceted search and browsing of audio content on spoken web. In *ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1029–1037.

[Garofolo et al., 2000] Garofolo, J., Auzanne, C., and Voorhees, E. (2000). The trec spoken document retrieval track: A success story. NIST Special Publication 246, pages 107–130.

[Godfrey et al., 1992] Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520.

[Hakkani-Tur et al., 1999] Hakkani-Tur, D., Tur, G., Stolcke, A., and Shriberg, E. E. (1999). Combining words and prosody for information extraction from speech. In *Proc. Eurospeech, vol. 5*, pages 1991–1994.

[Hanjalic et al., 2012] Hanjalic, A., Kofler, C., and Larson, M. (2012). Intent and its discontents: The user at the wheel of the online video search engine. In *ACM Multimedia*.

[Jurafsky et al., 2013] Jurafsky, D., Ranganath, R., and McFarland, D. (2013). Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech and Language*, 27:89–115.

[Kawahara, 2012] Kawahara, T. (2012). Transcription system using automatic speech recognition for the Japanese parliament (Diet). In *IAAI: Innovative Applications of Artificial Intelligence*.

[Kim et al., 2012] Kim, S., Yella, S. H., and Valente, F. (2012). Automatic detection of conflict escalation in spoken conversation. In *Interspeech*.

[Kompe, 1997] Kompe, R. (1997). *Prosody in Speech Understanding Systems*. Springer.

[Larson et al., 2011] Larson, M., Eskevich, M., et al. (2011). Overview of MediaEval 2011 rich speech retrieval task and genre tagging task. In *MediaEval '11*.

[Larson and Jones, 2012] Larson, M. and 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.

[Larson et al., 2012] Larson, M., Soleymani, M., Eskevich, M., Serdyukov, P., Ordelman, R., and Jones, G. J. F. (2012). The community and the crowd: Multimedia benchmark dataset development. *IEEE Multimedia*, July-September.

[Mairesse et al., 2012] Mairesse, F., Poifroni, J., and Di Fabbrizio, G. (2012). Can prosody inform sentiment analysis? Experiments on short spoken reviews. In *IEEE ICASSP*.

[Metze et al., 2012] Metze, F., Rajput, N., et al. (2012). The spoken web search task at Mediaeval 2011. In *IEEE ICASASP*.

[Mishne et al., 2005] Mishne, G., Carmel, D., et al. (2005). Automatic analysis of call-center conversations. In *ACM International Conference on Information and Knowledge Management (CIKM)*, pages 453–459.

[Mohammadi and Vinciarelli, 2012] Mohammadi, G. and Vinciarelli, A. (2012). Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transations on Affective Computing*, to appear.

[Morency et al., 2011] Morency, L.-P., Mihalcea, R., and Doshi, P. (2011). Towards multimodal sentiment analysis: harvesting opinions from the web. In *ICMI: 13th International Conference on Multimodal Interfaces*, pages 169–176.

[Oard, 2012] Oard, D. (2012). Query by babbling: a research agenda. In *Proceedings of the first workshop on Information and knowledge management for developing region*, pages 17–22. ACM.

[Oard et al., 2004] Oard, D. W., Soergel, D., Doermann, D., et al. (2004). Building an information retrieval test collection for spontaneous conversational speech. In *SIGIR*, pages 41–48.

[Oertel et al., 2011] Oertel, C., Scherer, S., and Campbell, N. (2011). On the use of multimodal cues for the prediction of degrees of involvment in spontaneous conversation. In *Interspeech*.

[Pentland, 2008] Pentland, A. (2008). *Honest Signals*. MIT Press.

[Petrelli et al., 2010] Petrelli, D., Vilar, N., Kalnikaite, V., Dib, L., and Whittaker, S. (2010). Fm radio: Family interplay with sonic momentos. In *ACM CHI*, pages 2371–2380.

[Rose and Levinson, 2004] Rose, D. E. and Levinson, D. (2004). Understanding user goals in web search. In *WWW '04: 13th International Conference on World Wide Web*, pages 13–19.

[Sellen and Whittaker, 2010] Sellen, A. J. and Whittaker, S. (2010). Beyond total capture: a constructive critique of lifelogging. *Communications of the ACM*, 53(5):70–77.

[Shriberg and Stolcke, 2004] Shriberg, E. and Stolcke, A. (2004). Prosody modeling for automatic speech recognition and understanding. In *Mathematical Foundations of Speech and Language Processing, IMA Volumes in Mathematics and Its Applications, Vol. 138*, pages 105–114. Springer-Verlag.

[Shriberg, 2005] Shriberg, E. E. (2005). Spontaneous speech: How people really talk, and why engineers should care. In *Interspeech*. Lisbon.

[Stifelman et al., 2001] Stifelman, L., Arons, B., and Schmandt, C. (2001). The audio notebook: Paper and pen interaction with structured speech. In *CHI 2001 Conference Proceedings*, pages 182–189.

[Vemuri et al., 2004] Vemuri, S., Schmandt, C., Bender, W., Tellex, S., and Lassey, B. (2004). An audio-based personal memory aid. *UbiComp 2004: Ubiquitous Computing*, pages 400–417.

[Vinciarelli et al., 2009] Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27:1743–1759.

[Ward et al., 2011] Ward, N. G., Vega, A., and Baumann, T. (2011). Prosodic and temporal features for language modeling for dialog. *Speech Communication*, 54:161–174.

[Ward and Werner, 2013] Ward, N. G. and Werner, S. D. (2013). Using dialog-activity similarity for spoken information retrieval. In *IEEE ICASSP, submitted*.

[Whittaker et al., 1999] Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F., and Singhal, A. (1999). Scan: Designing and evluating user interfaces to support retrieval from speech archives. In *SIGIR*, pages 26–33.

[Whittaker et al., 2008] Whittaker, S., Tucker, S., Swampillai, K., and Laban, R. (2008). Design and evaluation of systems to support interaction capture and retrieval. *Personal and Ubiquitous Computing*, 12(3):197–221.

[Wrede and Shriberg, 2003] Wrede, B. and Shriberg, E. (2003). Spotting ‘hot spots’ in meetings: Human judgments and prosodic cues. In *Eurospeech*, pages 2805–2808.