Ecologically Valid Evaluation of Speech Summarization

Anthony McCallum
University of Toronto
Toronto, ON, M5S 3G4
mccallum@cs.toronto.edu

Cosmin Munteanu
National Research Council
Canada
Fredericton, NB, E3B 9W4
cosmin.munteanu@nrc-cnrc.gc.ca

Gerald Penn
University of Toronto
Toronto, ON, M5S 3G4
gpenn@cs.toronto.edu

Xiaodan Zhu
National Research Council
Canada
Ottawa, ON, K1A 0R6
xiaodan.zhu@nrc-cnrc.gc.ca

ABSTRACT

Speech summarization is in demand for most large-scale audio-visual corpora, particularly broadcast news and meeting recordings. The former contains well-formed non spontaneous speech with low automatic speech recognition (ASR) word error rates and can often be summarized by excerpting the first N minutes of every story, whereas the latter contains spontaneous speech with astronomical word error rates, and (depending on who runs the meeting) a structure that can be difficult to do justice to in an excerpt-based summary. We propose an ecologically valid evaluation of speech summarization in the university lecture domain, as a means of differentiating between the beneficial properties of speech summaries in general and those of summaries of broadcast news in particular. The lecture domain is far better structured than meeting proceedings and has a far better word-error rate because the lecturer can be wired with a head-worn microphone and adapted to his or her own acoustic model, but it also contains speech that is more spontaneous than an anchored news broadcast and has higher error rates. The lecture domain also lends itself well to a task based evaluation metric, namely university level exams or quizzes, in which browsing or searching for factoid answers to questions is very similar to some uses of news corpora.

On the other hand, our research so far [5, 9] has demonstrated that the state of the art for evaluating summarizers in any domain is such a disaster that even summarizers that make use of every linguistic and acoustic feature we know of perform no better than those that simply fill an N-minute summary with the longest utterances, (the "length baseline") in spontaneous domains, including lectures, or than the aforementioned "positional baseline" in the broadcast news domain. This different baseline is a distinctive property of the broadcast news domain – the length baseline does not perform well here. What we need is a better evaluation protocol. Our working hypothesis is that an ecologically valid evaluation may show a convergence of baseline performances across domains, and that current, linguistically-rich experimental systems do in fact outperform them.

Keywords

speech summarization, utterance selection, spontaneous speech, user study, extrinsic evaluation

1. SPEECH SUMMARIZATION

Being one of the most natural means of communication between humans, it is not surprising that speech has continued to be used effectively in a wide variety of applications. In fact, it can be argued that instead of making speech obsolete, advancing technology actually makes speech a more viable means of information dissemination and storage. The widespread availability of recording equipment, as well as affordable digital storage, means that more and more spoken data can be collected and stored cost effectively. In addition, the pervasiveness of high bandwidth Internet connectivity means that such data can be made widely accessible if desired. For example, companies archive past conference calls and meetings, and universities often make recordings of lectures available online. The result is an increase in the number of speech documents available, a trend we expect to continue. Such an increase naturally results in a demand for methods to retrieve and navigate speech, which being inherently more linear than text, is not a trivial task when there is a constraint on time.

In the past, various methods have been proposed to navigate spoken documents, including time compression and playing two separate portions of audio simultaneously, one in each ear [1, 6]. Although some success has been reported in speeding up audio by up to two times, [7] shows that this is unfavourable to users.

We choose to focus instead on speech summarization. Summarization maintains the ability to review a representation

of an entire spoken document but focuses on those utterances (sentence-like units) that are most important and therefore does not require the user to process everything that has been said. Speech summarization is a natural extension of text summarization. We focus on extractive summarization where a selection of utterances is chosen from the original spoken document in order to make up a summary. This is in contrast to abstractive summarization where new utterances are generated. Extractive speech summarization is more feasible and as such makes up the majority of current speech summarization research. In fact, it has be shown that extractive speech summarization is quite effective for tasks requiring the use of spontaneous speech archives, and qualitative evaluation indicates extractive summaries are intuitive and efficient [4].

It is common to first use ASR to transcribe the speech, and then perform summarization on the resulting transcript. However, efforts have been made to rely less on transcripts, by incorporating acoustic features. This can be especially useful in situations where ASR word error rates are high. Extractive summaries can be displayed visually as text, or also audially. The latter may be most effective in situations with relatively high ASR word error rates, allowing users to listen to the original audio where reading an error laden transcript would be unfavourable. This is a case where extractive summarization may be preferable to abstractive summarization.

2. SPONTANEOUS SPEECH

Most speech summarization research has been conducted on broadcast news. This is at once an easy domain, because of nearly perfect studio acoustic conditions and professionally trained readers, and a difficult domain, because the widespread availability of written sources on the same topics means that a speech summarizer must not just objectively summarize the events of the news, but provide a perspective unique to the particular broadcast being summarized to be worth doing. In addition, a positional baseline performs very well in summarizing broadcast news [2]. That is, those utterances occuring near the beginning of a broadcast are often more suitable for inclusion in a summary than those occuring later, meaning that simply taking the first N utterances provides a very challenging baseline.

In our current research, we seek to distinguish the broadcast news domain from university lectures, in which we trade off less than ideal acoustic conditions for a structured presentation in which deviation from written sources (e.g., textbooks) is commonplace, and yet a positional baseline performs very poorly. [9] point out several important characteristics of spontaneous speech for speech summarizers, including that it is often not linguistically well-formed, and contains disfluencies and false starts. Additionally, spontaneous speech is more vulnerable to ASR error, resulting in higher word error rates.

3. TASK-ORIENTATED EVALUATION

As pointed out by [5], current speech summarizers have been optimized to perform an utterance selection task that may not necessarily reflect how a summarizer is able to capture the goal orientation or purpose of the speech data. In our study, we follow the prevailing trend in HCI towards extrinsic summary evaluation, where the value of a summary is determined by how well the summary can be used to perform a specific task rather than comparing the specific content of a summary to an artificially created gold standard [4]. [4] have used the meeting domain to perform a study where users are asked to use summaries in order to perform a decision audit task. Such a task required a general understanding of the spoken documents and required the gathering of information that could not easily have been found by simply performing a search on a transcript.

The university lecture domain is another example of a domain where summaries are an especially suitable tool for navigation. Like the decision audit task, simply performing a search will not result in the type of understanding required of students in their lectures. Lectures have topics, and there is a clear communicative goal. By using actual university lectures as well as university students representative of the users who would make use of a speech summarization system in this domain, any results obtained will be ecologically valid.

As an evaluation measure, we will have teaching assistants familiar with the lecture content create quizzes that are representative of what students are expected to learn in a given lecture. This provides an ecologically valid quantitative measure of whether a given summary is useful. Having this evaluation metric in place, automated summaries will be compared to manual summaries created by annotators familiar with the quiz content as well as summaries where the annotator has not previously seen the evaluation quiz. This allows us to demonstrate the value of task-oriented summarization, as well as determine which utterances an ideal summary would consist of.

4. EXPERIMENTAL DESIGN

4.1 Overview

Ecological validity is thus crucial to our experimental protocol. Earlier results in this area have been based on arbitrary numerical annotations of utterances by people, with no attempt to calibrate or contextualize the task of labeling utterances as "salient." [5] show that, under this earlier protocol, automatic summarization using any combination of various features generally does not perform better than a simple length baseline in any domain except broadcast news. When summarizing spontaneous university lecture data, performance is in fact significantly worse than when using SWITCHBOARD, a corpus collected for speaker identification research, in which participants were paid to speak aimlessly on a single topic with no goal or communicative intention to work towards. It was suggested that this is because SWITCHBOARD contains a greater variance in utterance length. Given these results, it is possible that our study will demonstrate that current state-of-the-art automatic summarizers, particularly those which rely on linguistically rich features, will be far more effective, when evaluated using an extrinsic evaluation task, than length or positional baselines. In the event of such an outcome, our now ecologically validated manual summaries can be used as a sort of gold standard for future optimizations of automatic speech summarizers.

Our study is a within-subject experiment where participants

are provided with undergraduate sociology lectures on a lecture browser system installed on a desktop computer. For each lecture, the browser makes accessible video, audio, manual transcript, as well as an optional summary. Our study will make use of first year university sociology lectures. As previously stated, evaluation of a summary will be based on how well the user of the summary is able to complete a quiz based on the content of the original material.

4.2 Human Subjects

Participants will be recruited from throughout one large North American university campus. Participants will be limited to undergraduate students who have had at least two terms of university studies, to ensure that they are familiar with the format of university level lectures and quizzes. In addition, students who have already taken the first year sociology course that we intend to draw lectures from will not be permitted to participate in this study. We aim to recruit a minimum of sixty-four participants, and ideally twice that amount. Recruitment will be through posters advertising the study posted on bulletin boards in university buildings containing a high volume of undergraduate student traffic. An effort will be made to distribute these posters widely in order to get a diverse set of participants.

4.3 Summaries

Our strategy for extractive speech summarization can be viewed as an utterance selection task. The original audio is broken down into a set of utterances which are identified by 200 millisecond pauses. The result are utterances that correspond to natural sentences or phrases. From this point forward, the task of summarization consists of choosing N% of the utterances, where N is typically between 10 and 30, for inclusion in a summary.

The automatic summarizer we use in this study is the one described in [5] which follows a similar structure to [8]. This summarizer makes use of a wide variety of speech features and is representative of the state-of-the-art. The summarizer takes either ASR or manual transcripts as well as an audio file as input which it uses to process disfluenceis and extract various features important to identifying sentences. False starts and repetitions, which occur commonly in spontaneous speech, are detected and removed. A binary logistic regression classifier is used to train an utterance selection module that can make use of various lexical (MMR score, utterance length, etc.), structural (utterance position, etc.), and acoustic (pitch, energy, speaking rate, etc.) features, among others.

4.4 Conditions

Participants will be asked to make use of the browser interface for up to four lectures, one for each of the following conditions:

- no summary
- automatic summary
- generic manual summary
- task-oriented manual summary

A given session will begin by having a participant perform a short warm up with a portion of lecture from the same course as the other lectures are drawn from. This will allow the participant to become familiar with the lecture browser interface. Following this, the participant will be given approximately ten minutes to complete a quiz on one lecture for each of the four conditions. During this time, the participant will be able to browse the audio/video, slides, and manually annotated transcript. However, each original lecture will be around forty minutes in length. This establishes a strong time constraint that we expect will make using a well-prepared summary beneficial to completing the task at hand. In fact, [3] show that providing users with a summary improved performance of a quiz-solving task compared to simply providing transcripts. An opinion survey also shows that there is a qualitative benefit to summarization in this context [3].

Lectures and corresponding conditions will be rotated appropriately for counter balancing. In each of the four conditions, each participant will be exposed to the lecture material for the first time, preventing any possible recall effects. As such, we are simulating a scenario in which someone wants to extract information from a lecture source that he or she has not previously heard, or that he or she has heard far enough in the past that none of the actual informational content or the approximate location of such content in the lecture can be remembered. For those conditions that rely on a manual summary, a separate human summarizer will be employed to create this summary.

4.4.1 No summary

The first condition serves as a baseline where no summary is used at all, and participants only have access to the video and manual transcript. As each participant will be given approximately ten minutes to complete the evaluation quiz, given a properly designed quiz, this should result in a rather low score. It is important to have this baseline so that we can show that given our experimental setup, being provided with summaries as defined in the following three conditions does provide some significant value in terms of the ability to perform well on the evaluation quiz.

4.4.2 Automatic summary

The automatically generated summary will be generated as described above. This essentially makes use of all of the best performing features and is representative of the state-of-the-art in speech summarization. Given an ecologically valid experimental setup and evaluation metric, results obtained for this condition may confirm that current state-of-the-art summarizers will indeed perform better than a simple length baseline on lecture data, a baseline that under previous intrinsic evaluation conditions has been competitive[5].

4.4.3 Generic manual summary

In this condition, each participant will be provided with a manually generated summary. Each manual summary will be created prior to the execution of the study by a human summarizer who has listened to a given lecture in its entirety. Summaries will be formed by selecting a set of those utterances that are believed to be most important or relevant. This condition will demonstrate how a manually created summary is able to aid in the task of taking a quiz on

the subject matter. In addition, along with the next condition, this summary provides a comparison for evaluating the performance of current state-of-the-art automatic summarizers.

4.4.4 Task-oriented manual summary

Similar to the previous condition, prior to the execution of the study, a summary will be created manually by selecting a set of utterances from the lecture transcript. In the case of the task-oriented summary, the annotator responsible for generating the summary will be shown the quiz to be used as an evaluation measure for a particular lecture, before being asked to create a summary. As a result, it is expected that these summaries should very accurately pick out those utterances that will directly aid a participant in answering questions on the quiz successfully. This will determine the value of designing summaries with a particular task in mind, as opposed to simply choosing utterances that are felt to be most important or salient. If such summaries result in participants performing this task well, then we will be able use these summaries to gain a better understanding of what an ideal summary should contain.

4.5 Summary Evaluation

A teaching assistant for the sociology class from which our lectures have been drawn will be employed to generate quizzes to be used for evaluation. These quizzes will be designed to incorporate factual questions only, avoiding those questions that would require comprehension of the topic. This ensures that variation in participant intelligence will have a minimal impact on our results. In addition, we will ensure that questions involve information that is distributed equally throughout the lecture, but at the same time not linearly in the transcript or video slider, which would allow participants to predict where the next answer might be located. Finally, all questions will be non-trivial to minimize the chance of the participant having previous knowledge of the answer. We are also likely to include a normalizing quiz at the start of the experiment to determine any questions for which participants already know the answer.

5. PROGRESS TO DATE

All lecture data has already been recorded and manually transcribed so that the utterances which make up the summaries are legible and ASR errors do not interfere with the participants' ability to perform the tasks at hand. We have a fully functional lecture browsing interface and lectures are currently being chosen and preprocessed to work with this system. We have begun recruiting participants and will begin scheduling shortly. Outstanding work involves the creation of evaluation quizzes as well as any other evaluation metrics we decide to implement. Following that, we will begin running the experiment which is likely to take place during the first portion of the summer.

6. CONCLUSION

We have proposed an ecologically valid evaluation on speech summarization using the university lecture domain. We will evaluate the value of task-oriented summaries as well as use a task-based metric for determining the value of a summary. In addition, the resulting verified summaries will set a new high-water mark for evaluation within spoken language processing research.

7. REFERENCES

[1] E. Cherry and W. Taylor. Some further experiments upon the recognition of speech with one and with two ears. *Journal of the Acoustical Society of America*, 1954.

[2] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. From text summarisation to style-specific summarisation for broadcast news. *Advances in Information Retrieval*, pages 223–237, 2004.

[3] L. He, E. Sanocki, A. Gupta, and J. Grudin. Comparing presentation summaries: slides vs. reading vs. listening. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 177–184. ACM, 2000.

[4] G. Murray, T. Kleinbauer, P. Poller, S. Renals, J. Kilgour, and T. Becker. Extrinsic summarization evaluation: A decision audit task. *Machine Learning for Multimodal Interaction*, pages 349–361, 2008.

[5] G. Penn and X. Zhu. A critical reassessment of evaluation baselines for speech summarization. *Proceedings of ACL-HLT. Columbus, OH*, 2008.

[6] A. Ranjan, R. Balakrishnan, and M. Chignell. Searching in audio: the utility of transcripts, dichotic presentation, and time-compression. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 721–730. ACM, 2006.

[7] S. Tucker and S. Whittaker. Time is of the essence: an evaluation of temporal compression algorithms. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 329–338. ACM, 2006.

[8] K. Zechner. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 199–207. ACM, 2001.

[9] X. Zhu and G. Penn. Summarization of spontaneous conversations. In *Proc. of Interspeech*, pages 1531–1534. Citeseer, 2006.