

Spoken Language Condensation in the 21st Century

Klaus Zechner

Educational Testing Service
Center for Assessment Design and Scoring
Rosedale Road, MS 11-R
Princeton, NJ 08541, USA
kzechner@ets.org

Abstract

While the field of Information Retrieval originally had the search for the most relevant documents in mind, it has become increasingly clear that in many instances, what the user wants is a piece of coherent information, derived from a set of relevant documents and possibly other sources. Reducing relevant documents, passages, and sentences to their core is the task of text summarization or information condensation. Applying text-based technologies to speech is not always workable and often not enough to capture speech specific phenomena. In this paper, we will contrast speech summarization with text summarization, give an overview of the history of speech summarization, its current state, and, finally, sketch possible avenues as well as remaining challenges in future research.

1. Introduction

Computer-based text summarization has a long history -- long in relation to the existence of modern electronic calculating machines. The first publication on that topic is most likely the paper of Luhn [16] who lays out a set of heuristics and statistics to create an extract summary of a text. Of course, information was measured in Kilobytes in those days, whereas we look at Giga- and sometimes even Terabytes now. Exactly that explosion of textual information in electronic form was the factor that made a small research project for few insiders into one of the major fields of natural language processing (NLP) and information retrieval (IR). The status of that field can also be seen in the fact that an additional strand of the Text REtrieval Conference series (TREC, [16]) was established a couple of years ago, the Document Understanding Conferences (DUCs, [18]). What happened to text summarization in the past 5-10 years, might well happen with speech summarization in the next 10-20 years. But speech summarization is a comparatively much younger discipline, its origins dating mostly from the 1990s. It was not until then that sufficiently large volumes of spoken language were available (e.g., the Broadcast News corpus of the Defense Advanced Research Project Agency (DARPA, [19])) to make summarization an essential task. Another development coincided with the data availability: for the first time, commercial speech recognition engines could

successfully recognize continuous speech without a speaker-specific training phase, in unrestricted domains, and in real-time, the latter mostly due to the exponential increase in processing power. With two important preconditions met and with the rapid rise of text summarization research, it is understandable why speech summarization got some momentum, primarily in the last 5 years.

Speech summarization has a number of significant challenges that distinguish it from general text summarization. The next section will discuss some of these challenges and briefly suggest ways for overcoming them. In Section 3, we will provide a brief historical perspective on the genesis of different strands of speech summarization in the last decade. Finally, we will discuss major challenges that still remain to be successfully addressed in Section 4, and the last section concludes the paper.

2. Speech vs. Text Summarization

Speech is a different medium from text. Initially, unlike the black characters on white paper or the binary representations of characters on a storage medium (such as a CD-ROM, for example), speech is nothing but a transient stream of sound waves, generated by the human vocal tract, and quickly dispersing, unless recorded, typically in digital form.

The first approximation of speech summarization is to use the tools of text summarization and apply them to the textual form of the speech samples. The textual representation is achieved by transcribing the speech -- either by a human transcriber (which is very time consuming and expensive, but fairly accurate), or by a transcribing machine, a speech recognizer (in the long run much cheaper but far less accurate). In many practical contexts, human transcription is not a real option, e.g., because of time constraints, and one is forced to rely on the fairly faulty output of a speech recognizer. Word error rates range from 5-15% (clean broadcast news, good signal quality) to up to 50% (spontaneous and possibly noisy telephone conversations).

What are the main differences between text and speech as they pertain to summarization?

- Speech contains a number of spontaneous effects, which are not present in written language, such as hesitations, false starts, or fillers. These extra word tokens that contain virtually no meaning should be removed from the transcript before summarization.
- Speech is a continuous phenomenon that comes without unambiguous sentence boundaries. In order to determine “minimal pieces of information” (which could be sentences or clauses, e.g.), sentence boundaries have to be inserted automatically, before further processing.
- Many forms of speech involve more than one speaker engaged in an interactive conversation. In this case, content may well be distributed across two or more speakers – the task at hand being the automatic identification of cross-speaker information linkages or coherence links.
- Last but not least, one has to deal with transcription errors of automatic speech recognition engines, which can be, as stated above, quite substantial.

In previous work [14] we demonstrated some methods for addressing these mentioned differences between text and speech and also showed in how far they benefit the resulting summary. The two main areas where the adjustment helped were (a) automatic detection and removal of speech disfluencies, and (b) automatic detection of regions with higher confidence (in their correctness, as assigned by the speech recognizer) and using this information to re-rank (or re-weight) sentences (or information units).

3. A Brief History of Speech Summarization

3.1. Narrow domain summarization

There has been substantial research effort and progress in the domain of summarization within a pre-specified narrow genre, in particular within the framework of VerbMobil [12], a multi-million Euro project funded by the German Ministry of Education, Science, Research and Technology (1993-2000). The core goal of VerbMobil was machine translation of speech between German and Japanese and back. A domain model then allows a particular speech act and/or its semantic content to expand or extend the dialogue representation. To generate a summary, the core information in the domain model is selected and then transformed into sentences of a particular language (the summarization module can, in principle, handle multiple languages) [9]. The coherence, cohesion and fluency of these

summaries are quite impressive; they show what state-of-the-art domain summarizers are capable of. However, there are obvious disadvantages to this approach: not only does it require a dedicated and time-consuming effort to build a domain model and then associate with it rules for generation of summaries, it also stays limited to that domain and is not flexible when the domain changes.

3.2. Emphasis detection

There has been some research going on in the field of automatic emphasis detection by using prosodic features. This research typically ignores the verbal content of the speech sample and instead focuses solely on its prosodic characteristics. This would have the advantage of being more stable and robust, since it does not have to deal with speech recognition errors.

Chen and Withgott [3] present a Hidden Markov Model (HMM) based approach to detecting “emphasized regions” in a speech file. The HMM was trained using a manually annotated set of recordings. On an independent test set, the agreement between human emphasis and automatic emphasis was fairly high ($K > 0.5$).

Stifelman [10] uses a pitch-based emphasis detection program based on Arons’ [1] algorithm) and finds that the detected emphasized regions match well with the beginnings of manually marked discourse segments (based on [5]).

While we feel that this technology may have its applications in certain fields, its main benefit will most likely be to complement summarizers that work on textual representations of speech. A unification between the text based technology and the prosody-based emphasis detection remains yet to be demonstrated.

3.3. Summarization of news broadcasts

When DARPA started its Broadcast News Recognition series in the last 90’s, a new task was created: SDR, Spoken Document Retrieval, where speech recognizer transcripts serve as indices to an acoustic database. At that time, another extension of speech transcription was envisioned: speech summarization. Waibel, Bett, and Finke [13] presented a simple system to summarize SwitchBoard [4] dialogues using a metric derived from MMR [2] (Maximum Marginal Relevance). They found that the information reduction due to summarization for the most part did not prevent a user from correctly identifying the main concepts of the document, i.e., these concepts were retained in the summaries.

Hirschberg et al. [7] presented a system that supports local navigation for browsing and information extraction from acoustic databases, using speech recognizer transcripts together with the original audio recording. While it was demonstrated that the system shows benefits when used in different information

retrieval scenarios, it has failed to show an effect in a summarization task.

Valenza et al. [11] present a summarizer, which combines confidence scores from a speech recognizer with generic relevance scores. Experiments showed that the summary word error rates were significantly smaller than the global error rate for the entire transcript.

Hori and Furui [8] built a summarizer whose scores are based partly on linguistic information such as position in a parse tree, and whose search is a dynamic programming approach. They use it to reduce Japanese TV captions by about 30-40% in length.

Our own work [14] extended the genre of news broadcast with typically one speaker per segment to spoken dialogues: (a) conversations between family and friends over the phone; and (b) multi-person discussions in research group meetings. The summarization engine's core is based on the MMR technology and we use machine-learning techniques to identify and remove speech disfluencies, to automatically insert sentence boundaries and to determine cross-speaker information in question-answer pairs. The system can further also use confidence scores in combination with the information content related scores to accommodate at least to some extent the erroneous output of the speech recognizer [15]. In an experiment, we showed that for more informal genres, the system significantly outperformed two baselines (LEAD=taking the leading part of a conversation segment, and MMR).

4. Remaining Challenges

While early successes in spoken language summarization are encouraging, it is very clear that much remains to be done in a wide variety of fields. We will briefly go over some of the main points, as they are presenting themselves to the research community.

- Robust and improved speech recognition: While state-of-the-art speech recognizers exhibit acceptable performance (90% and above) for some domains and genres (e.g., news anchor in noise-free environment), the accuracy on many other genres is markedly lower (it can be 50% and below). Summaries based on that erroneous input are hard to read and hard to understand. We can hope that further advances in speech recognition will improve this situation for spontaneous speech data, dialogues/discussions, and for speech in noisy environment and/or the presence of sound sources in the background.
- Integration of prosodic and word-based information: As we mentioned above, there have been several attempts to detect emphasized stretches of speech automatically which then can be used to form a summary (an

excerpt in the truest sense). However, what is missing is a thorough examination of the interplay between prosodic features and derived entities (such as "emphasis") on the one hand and text-based summary generation on the other hand. One would surmise that the two avenues for summary creation could mutually enhance and benefit each other.

- Pulling together AI (Artificial Intelligence) and IR (Information Retrieval) based summarization: While AI-style summarization has its firm place in closed domains, IR can handle unlimited domains well at the expense of lacking any "understanding", any meaning representation of the underlying text. How can AI approaches made more robust so that they can "escape" their confinements of domain? And how can IR-style summarization be enhanced by using NLP methods, but without sacrificing their broad applicability? That at least a tentative merging might be feasible shows the related field of Question-Answering where some groups use, e.g., syntactic parses, semantic frames and relationships to arrive at a better content representation, which is all done in a framework of domain independent information retrieval [6].
- Bridging the gap between summarization and question answering (QA): Summarization and question answering perform a similar task, in that they both map an abundance of information to a (much) smaller piece to be presented to the user. Both can work either on single documents or on a collection of documents where the issue of information fusion and reconciliation becomes important. But one difference still remains: a summary is in most cases a short "story" about the contents of one or multiple texts or passages, whereas the output of a QA-system is typically a fact expressed in a phrase, description, or a sentence (this is enforced by the TREC space constraints of e.g., 50 characters per answer). We could now conceive of both QA developing branches where the information can be longer, say, 2 or 3 paragraphs --- which is likely necessary as questions become more complex and intricate --- and summarization being reduced, in some instances, to a much smaller form than usual.
- Devising new meaningful evaluation criteria and metrics for spoken language summarization: Summarization evaluation has

always been a somewhat problematic issue. It starts with the fact that in almost all cases, an “ideal” summary does not exist and the divergence of even well trained annotators in deciding on summary-relevance of sentences is astounding. Further, there are redundancies and dependencies between sentences, which are hard to enumerate and handle. And finally, while matching sentences can be counted easily, how to evaluate fluency and coherence in an automatic way is still elusive. When we move to speech, a number of additional issues arise: Are summaries to be read while listening to the corresponding audio? How should one deal with word errors from the speech recognizer --- it might well be that a particular sentence is highly relevant in the text, but unfortunately the word error rate is too high to be made sense of. (If the latter sentence’s acoustic correlate is played, however, the word errors can lose their bite). This cursory tour through challenges of summarization evaluation meant to demonstrate that much more work in this area is needed since comparing different summarizers in a meaningful way is as crucial as it is in other fields (e.g., in speech recognition).

5. Conclusion and Outlook

In this paper, we have put the area of speech summarization into its historical perspective, presented its differences to text summarization, and discussed some of its current and future challenges, in part also shared by other fields of summarization.

We believe that the future will bring more intensive research and new technological developments on all branches of summarization, in particular multi-lingual, multi-modal, and multi-media summarization, as well as some merging trends, such as between summarization and question answering or between summarization of different types of media in multi-media summarization. The scope of and the need for summarization will likely widen by a large extent so as to continually generate challenges and opportunities in this discipline.

6. Disclaimer Note

Any findings or opinions expressed in this paper are those of the author and not necessarily of Educational Testing Service.

7. References

- [1] Arons. B. 1994. Pitch-based emphasis detection for segmenting speech. Proceedings of the ICSLP-96, 1931-1934.
- [2] Carbonell, J., Geng, Y., and Goldstein, J. 1997. Automated query-relevant summarization and diversity-based reranking. Proceedings of the IJCAI-97 Workshop on AI and Digital Libraries, Nagoya, Japan.
- [3] Chen, F.R. and Withgott, M. 1992. The use of emphasis to automatically summarize spoken discourse. Proceedings of the ICASSP-92, 292-232.
- [4] Godfrey, J.J., Holliman, E.C., and McDaniel, J. 1992. Switchboard: Telephone speech corpus for research and development. Proceedings of the ICASSP-92, 517-520.
- [5] Grosz, B.J. and Sidner, C.I. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12(3), 175-204.
- [6] Harabagiu, S.M. and Pasca, M. 2000. Mining Textual Answers with Knowledge-Based Indicators. Proceedings of FLAIRS-2000, 214-218.
- [7] Hirschberg, J., Whittaker, S., Hindle, D., Pereira, F., Singhal, A. 1999. Finding information in audio: a new paradigm for audio browsing/retrieval. Proceedings of the ESCA Workshop: Accessing information in spoken audio, 117-122, Cambridge, UK.
- [8] Hori, C. and Furui, S. 2000. Automatic speech summarization based on word significance and linguistic likelihood. Proceedings of ICASSP-00, 1579-1582.
- [9] Reithinger, N., Kipp, M., Engel, R., and Alexandersson, J. 2000. Summarizing multilingual spoken negotiation dialogues. Proceedings of ACL-2000, 310-317.
- [10] Stifelman, L.J. 1995. A discourse analysis approach to structured speech. AAAI-95 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.
- [11] Valenza, R., Robinson. T., Hickey, M., and Tucker, R. 1999. Summarization of spoken audio through information extraction. Proceedings of the ESCA Workshop: Accessing information in spoken audio, 111-116, Cambridge, UK.
- [12] Wahlster, W. 1993. Verbmobil – translation of face-to-face dialogs. Proceedings of MT Summit IV, Kobe, Japan.
- [13] Waibel, A., Bett, M., Finke, M. 1998. Meeting Browser; Tracking and summarizing meetings. Proceedings of the DARPA Broadcast News Workshop.
- [14] Zechner, K. 2001. Automatic Summarization of Spoken Dialogues in Unrestricted Domains. Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, available as Technical Report CMU-LTI-01-168.
- [15] Zechner, K. and Waibel, A. 2000. Minimizing word error rate in textual summaries of spoken language. Proceedings of NAACL-2000, 186-193.
- [16] Luhn, H.P. 1958. Automatic creation of literature abstracts. *IBM Journal*, 159-165.
- [17] TREC. <http://trec.nist.gov/>
- [18] DUC. <http://duc.nist.gov/>
- [19] DARPA Broadcast News. <http://www.nist.gov/speech/publications/darpa99/>