Retrieving Radio News Broadcasts in Danish: Accuracy and Categorization of Unrecognized Words

Morten Hertzum, Haakon Lund
University of Copenhagen
Birketinget 6, Copenhagen, Denmark
hertzum@hum.ku.dk, hl@hum.ku.dk
+45 3234 1344, +45 3234 1454

Rasmus Troelsgård
Technical University of Denmark
Richard Petersens Plads, Lyngby, Denmark
rast@dtu.dk
+45 4525 3922

ABSTRACT
Digital archives of radio news broadcasts can possibly be made searchable by combining speech recognition with information retrieval. We explore this possibility for the retrieval of news broadcasts in Danish. An average of 84% of the words in the broadcasts was recognized. Most of the unrecognized words were compounds, names, and other words that appear of value to retrieval. Thus, the set of words describing a broadcast has to be expanded to compensate for the recognition errors. We discuss doing this by exploiting the alternative matches from the speech recognizer and by extracting words from a related corpus.

Author Keywords
Information retrieval, speech recognition, audio archives

ACM Classification Keywords
H3.7. Information storage and retrieval: Digital libraries.

INTRODUCTION
Systems for retrieving textual information have become ubiquitous, and users expect them – often rightfully – to be effective and efficient. For non-textual information the situation is different: The content-based retrieval of audio, images, and video faces many challenges (Datta et al., 2008; Hu et al., 2011; Larson and Jones, 2011). Yet, an increasing amount of the content in digital repositories is non-textual. In many of these repositories the users can only retrieve the non-textual content by searching the accompanying textual information, which may be sparse compared to the non-textual content. This study focuses on the retrieval of spoken content, specifically radio broadcasts. Research on spoken content retrieval combines automatic speech recognition with information retrieval (Larson and Jones, 2011). We explore spoken content retrieval for Danish speech and assess how valuable the words not recognized by the speech recognizer would be to users of the broadcast archive.

The Danish Broadcasting Corporation, DR, is Denmark's largest electronic media enterprise. Since it was founded in 1925 radio broadcasts have been a core part of its business. More than 800,000 hours of radio broadcasts from DR are digitally archived. However, the possibilities for searching this vast source of information about the Danish cultural heritage are limited (Lund et al., 2013). For most broadcasts the only searchable information is their title and the time of broadcast. This paucity of the information available for searching the archive severely restricts its usefulness. The size of the archive precludes manual indexing of the content, and its cultural-heritage value makes it an important application of Danish spoken content retrieval. This study is part of a long-term effort to make the content of the broadcast archive searchable.

In recent years speech recognition has become more robust and more standardized (Yu and Deng, 2015). The increased maturation has made it feasible to develop speech recognizers for languages spoken by fewer people, such as Danish (5.7 million speakers). However, an evaluation of speech recognition in Danish healthcare about a decade ago revealed considerable problems with the technology (Alapetite et al., 2009). One challenge in Danish compared to, for example, English is that semantically complex concepts are expressed by joining multiple words into one compound word. This increases the number of distinct words the speech recognizer must be able to recognize. Even for English, the word error rate for recognizing broadcast news is typically 10-20% and it increases to 20-40% for conversational speech (Larson and Jones, 2011).

It is an important research question whether high recognition accuracy is required for effective spoken content retrieval. Sparck Jones et al. (1996) argued that word error rates are not indicative of the quality of spoken content retrieval because they treat all words equally. For retrieval purposes it is only the meaning-bearing words that need to be recognized correctly. In addition, a word needs only be recognized once within a broadcast to be available for retrieving the broadcast. This indicates that broadcast-level measures are better suited for spoken content retrieval, whereas the accuracy of speech recognition is best measured at the word level. It also indicates that modest speech-recognition performance may be sufficient for effective spoken content retrieval. For example, Munteanu et al. (2006) found that users' retrieval performance and their experience of the retrieval process were better at a word error rate of 25% than without the system. Allen (2003) reported that retrieval effectiveness fell less than 10% with a word error rate of around 30%. We investigate the retrieval value of the words not in the speech-recognizer output by manually categorizing such words.

OzCHI '16, November 29 – December 2 2016, Launceston, TAS, Australia
© 2016 ACM. ISBN 978-1-4503-4618-4/16/11…$15.00
DOI: http://dx.doi.org/10.1145/3010915.3010972.

METHOD

To investigate the effectiveness of automatic speech recognition for retrieving radio broadcasts in Danish we conducted an empirical study.

Broadcasts

For this exploratory study, 19 broadcasts were selected and transcribed verbatim. The broadcasts were news broadcasts with an anchor person in a studio, assisted by journalists and sources who reported from the field. While the sound quality in the studio was high, it varied in the field due to, for example, phone connections and background noise. To ensure variation in their content the news broadcasts were selected from the period 1965 to 2010. In total, the 19 news broadcasts contained 27130 spoken words.

Speech-Recognition System

We used a Danish speech recognizer developed on the basis of the Kaldi toolkit (Povey et al., 2011). It had a vocabulary of 53914 words.

The output from the speech recognizer was a textual representation of the audio broadcasts. This textual representation consisted of the words that produced the best match between the language model of the speech recognizer and the audio content of the broadcasts. In addition, the speech recognizer produced a number of alternative matches, which gave the competing, but less likely, proposals of how to transform the audio content into text. Each alternative match had a confidence score that indicated the level of agreement between the language model and the spoken audio.

Procedure

Our analysis consisted of comparing the output from the speech recognizer to the transcription of the broadcasts. We initially aligned the speech-recognizer output and the transcription by, automatically, determining the minimum number of word substitutions, insertions, and deletions needed to transform the speech-recognizer output into the transcript. On this basis we calculated the word error rate, which is the standard measure of accuracy in automatic speech recognition (Larson and Jones, 2011):

$$Word\ error\ rate = \frac{Substitutions + Insertions + Deletions}{Number\ of\ words\ in\ transcript}$$

After calculating the word error rate – a word-level measure – we turned to broadcast-level measures. As a precursor to calculating these measures we stemmed the words using the Snowball stemmer (snowballstem.org). Stemming is standard in information retrieval (to relieve the searcher from the task of specifying all inflections of the query terms). Two broadcast-level measures central to information retrieval are the coverage and noise of the speech-recognizer output:

$$Coverage = \frac{Distinct\ transcript\ words\ in\ recognizer\ output}{Distinct\ words\ in\ transcript}$$

$$Noise = \frac{Distinct\ nontranscript\ words\ in\ recognizer\ output}{Distinct\ words\ in\ transcript}$$

We determined the coverage and noise for two different outputs from the speech recognizer: (a) the words in the best match between the language model and the broadcasts and (b) the words in the best and alternatives matches. We included the alternative matches that had a confidence score of at least 0.98.

Across the broadcasts 1198 distinct words were, as the analysis will show, never recognized correctly. We manually categorized these words to learn about their content and be able to assess their value for information retrieval. Table 1 shows the categories. In addition to this categorization we indicated whether each word was a noun (including whether it was a name), a verb, an adjective, or something else (including ambiguous).

The categorization was done in five steps. First, the list of unrecognized words was explored to create the set of categories. Second, 10% of the unrecognized words were randomly selected and coded by the first and second authors independently. Third, the two authors met and discussed their codings to reach consensus about the categorization of the words and to arrive at a shared understanding of the categories. Fourth, the two authors independently coded the remaining 1079 (90%) of the unrecognized words. Cohen's (1960) kappa of the agreement between the two authors was .81 for the categorization according to the categories in Table 1 and .74 for the categorization into word classes. Both kappa values met the criterion that kappa values above .60 indicate satisfactory reliability (Lazar et al., 2010). Fifth, the two authors discussed the words they had categorized differently and a consensus was reached.

Category	Definition	Example
Name	The name, or part of the name, of a person, geographical location, organization, or another entity	"Vladivostok"
Group of persons	A word that denotes a group of persons by a common characteristic	"sukkersygepatienter" (diabetics)
Title	A word added to a person name to signify the person's official position or academic qualification	"vicepolitidirektør" (deputy police director)
Numeral	A word denoting a number	"syvogfirs" (eighty-seven)
Compound	A word made by joining two or more words into one longer word	"sikkerhedszone" (security zone)
Error	A misspelling in the transcription of the audio file	

Table 1. Categories of unrecognized words.

RESULTS

Across the broadcasts the word error rate was an average of 22% (*SD* = 5%). Table 2 shows the coverage and noise of the best match from the speech recognizer and of the best-and-alternative matches. Coverage was significantly higher for the best-and-alternative matches than for the best match, Welch's $t(35.75) = 8.46$, $p < .01$. However, the noise was also significantly higher for the best-and-alternative matches, Welch's $t(18.50) = 293.85$, $p < .001$.

Measure	Best match	Best and alternative matches
Coverage	84±3%	87±3%
Noise	19±6%	204±47%

Table 2. Recognition accuracy (mean ± standard deviation).

An example of the noise included with the alternative matches was the phrase "værre end aktiemarkedet frygtede" (worse than the stock market feared). The best match of this phrase was "værre en aktiemarked frygtede", which from a retrieval point of view was close to correct. Including the alternative matches added the following words: "verden" (world), "hverken" (whether), "hverdagen" (everyday), "vejret" (weather), "værn" (protection), "væren" (being), "værten" (host), "værdien" (value), "værende" (being), "han" (he), "and" (duck), "er" (is), "værtinden" (hostess), "men" (but), "af" (of), "markedet" (market), "og" (and), "aktiemarkedet" (stock market), and "over" (over). About half of these words were plainly off topic; the other half would add little value to information retrieval because they were either very common or near identical to the words in the best match.

A total of 1198 distinct words occurred in the broadcasts but not in the best-match output from the speech recognizer. These unrecognized words constituted the gap between the achieved coverage and a perfect 100% coverage. We initially assessed the value of the unrecognized words for information retrieval by determining whether they were content bearing. To this end we compared the unrecognized words with the most frequent words in the newswires from the Danish newswire service Ritzau. We assumed that the words in the Ritzau newswires were not content bearing if they occurred with a frequency of more than 0.1% of the total Ritzau corpus of 500+ million words. By this definition, only 2 of the 1198 unrecognized words were not content bearing.

Table 3 shows the categories of unrecognized words. Compounds (32%) were the largest category. Words such as "arbejdsgiverforeningsformanden" (the president of the employers' association) illustrated how compounds increased the number of distinct words the speech recognizer needed to be able to distinguish. A further indication of the prevalence of compounds among the unrecognized words was the larger average length of unrecognized (9.6 characters) compared to recognized (4.9 characters) words. Many of the compounds appeared of clear relevance to information retrieval.

The second largest category of unrecognized words was names (31%). Persons, places, as well as organizations were frequently mentioned by name in the broadcasts. Names are indispensable in specifying the content of the news and thereby also important to information retrieval. Titles (3%) were used for introducing persons and as a convenient alternative way of referring to a person who had previously been mentioned by name.

Groups of persons (7%) were another and – compared to names – more aggregate way of specifying who the news was about. Sometimes concrete groups of persons were referred to using broad terms such as "lægegruppe" (medical group) but mostly the words indicating groups of persons were highly indicative of the news story, such as "borgerrettighedsforkæmpere" (civil rights activists) and "boligspekulanter" (housing speculators). These words would be valuable for information retrieval.

Numerals (3%) were a small category, and words such as "toogtrediveårig" (32-year-old) appear unlikely to matter much to information retrieval. The 'other' category (23%) contained a variety of words, including "juni" (June), "opfordrede" (encouraged), and "tradition" (tradition). While the failure to recognise many compounds and names was probably that they were unknown to the speech recognizer, it was less apparent why many of the 'other' words had not been recognized. Finally, a mere 1% of the unrecognized words were the result of errors in the transcription of the broadcasts.

Category	Count	Percentage
Compound	384	32
Name	367	31
Group of persons	84	7
Numeral	41	3
Title	31	3
Error	13	1
Other	278	23

Table 3. Categorization of the 1198 unrecognized words.

Nouns (74%) were by far the most common word class among the unrecognized words. Verbs (5%) and adjectives (5%) were fairly rare, and the word class of the remaining words (16%) was unknown or ambiguous.

The speech recognizer can only recognize words it knows, that is, words in its vocabulary. As much as 81% of the 1198 unrecognized words were not in the vocabulary (including 97% of the compounds but only 57% of the 'other' words). The remaining 19% of the unrecognized words could, in principle, have been recognized but, in practice, their pronunciation did not match the speech model or was distorted by noise.

DISCUSSION

With a word error rate of 22% the speech recognizer we are using for Danish speech performs similarly to the 10-20% word error rates typically achieved for broadcast

news in English. The coverage of 84% is better than the word error rate immediately suggests it would be because many words occur multiple times in a broadcast and, thereby, provide multiple opportunities for being recognized at least once. It is, however, not satisfactory for information retrieval that 16% of the words in the broadcasts are not available for retrieval purposes. The categorization of the unrecognized words shows that most of them are names and compounds, which appear to be of value to information retrieval. What can be done about this problem?

When Allen (2003) reported that retrieval effectiveness fell less than 10% with a word error rate of 30% he also reported that the main means of compensating for the recognition errors was automatic expansion of the queries and the speech-recognizer output with similar terms. One approach to expansion would be to add the alternative matches from the speech recognizer to the set of words used for characterizing each broadcast. This approach increases the coverage, but only by 3 percentage points. In addition, the modest increase in coverage is achieved at the cost of a dramatic increase in noise. We could set an even higher threshold for including the alternative matches than a confidence score of 0.98. However, upping the threshold would not only reduce the noise but also the already modest improvement in coverage. This approach to expansion appears unproductive.

Another approach to expansion would be to enrich the speech-recognizer output with words from the Ritzau newswires. An established way of identifying such expansion words is by co-occurrence (Singhal and Pereira, 1999): The words that often occur in the same newswires as the words in the speech-recognizer output are likely to be about the same topic. It is important to this approach that the news broadcasts and the Ritzau newswires are both about contemporary news from a Danish perspective. Previous work emphasizes the importance of drawing the expansion words from a topically related corpus (Allen, 2003; Singhal and Pereira, 1999). The topical relatedness may be the characteristic that explains why this approach appears more effective than expansion with the alternative matches, which were produced without topical information beyond the audio features of the speech. In our future work with spoken content retrieval from DR's news broadcasts, we will experiment with using the Ritzau newswires for expansion. However, effective retrieval across the full range of broadcasts in the digital archive will either require expansion with words from different text corpora depending on the type of broadcast or another approach to expansion.

It is of particular concern that the majority of the unrecognized words are nouns because previous studies find that nouns (including names) are the most common word class in queries (Anick, 1994; Jadhav et al., 2014; Jiang et al., 2013). For example, Anick (1994) finds that "users tend to simply concatenate sequences of nouns into queries, without using function words or verbs to separate them". A third approach to improve coverage would simply be to extend the vocabulary of the speech recognizer, in particular with more nouns. The Ritzau newswires suggest themselves as a rich source of additional vocabulary words.

To the extent that users include verbs or even multi-word phrases in their queries, the process of specifying the queries in text may lead to wordings that differ from how people express themselves orally (see, e.g., Leech et al., 2001). Such differences and the accompanying risk of missing relevant broadcasts are, however, more likely for broadcasts containing conversational speech than for news broadcasts, which in part are read from a manuscript. We speculate that names, in particular, are more likely to occur in specific and fact-finding queries than in broader and subject-oriented queries. Thus, the categories of unrecognized words may harm some kinds of information needs more than others. In relation to news broadcasts we have no basis for estimating whether fact-finding queries will be more, or less, frequent than other kinds of queries. Rather, an archive of news broadcasts appears relevant to a wide variety of information needs.

The next step in our efforts to make the digital archive of DR broadcasts searchable is to run 20,000 hours of news broadcasts through the speech recognizer and have experimental participants solve information-retrieval tasks about the content of these broadcasts. The present, exploratory study shows that the speech recognizer should be sufficiently accurate for this purpose. However, the present study has also clarified that we need to extend the vocabulary of the speech recognizer and to expand the speech-recognizer output with words from another news corpus. Access to the broadcasts by querying their spoken content will be valuable to historians, journalists, media and communication researchers, and many others. Eventually, we hope to be able to investigate the search process of users as they pursue their information needs in the full 800,000 hours of digitally archived broadcasts: What are they looking for? How do they search? Which broadcasts are retrieved more often? How can the retrieval system be improved?

CONCLUSIONS

The word error rate of the speech recognizer was 22% for the Danish news broadcasts. This accuracy is comparable to the accuracy achieved for news broadcasts in English but meant that 16% of the words in the broadcasts were not recognized any of the times they were spoken. Most of these words appeared valuable to retrieval. Using the alternative matches from the speech recognizer to expand the set of words describing a broadcast dramatically increased the amount of noise. It appears more promising to extract expansion words from a topically related corpus. The results of this exploratory study feed into a long-term effort to make the cultural heritage expressed in Danish radio broadcasts since 1925 available to users.

ACKNOWLEDGEMENTS

This study is part of the CoSound project, which was co-funded by the Innovation Fund Denmark (grant no. 0603-00475B). We are grateful to Morten Højfeldt Rasmussen for access to his speech recognizer and to Ritzau for access to their newswires from 1988 to 2015.

REFERENCES

Alapetite, A., Andersen, H.B. and Hertzum, M. Acceptance of speech recognition by physicians: A survey of expectations, experiences, and social influence. *International Journal of Human-Computer Studies* 67, 1 (2009), 36-49.

Allen, J. Robust techniques for organizing and retrieving spoken documents. *EURASIP Journal on Applied Signal Processing* 2003, 2 (2003), 103-114.

Anick, P.G. Adapting a full-text information retrieval system to the computer troubleshooting domain. In *Proc. SIGIR 1994*, ACM Press, New York (1994), 349-358.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37-46.

Datta, R., Joshi, D., Li, J. and Wang, J.Z. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys* 40, 2 (2008), 5:01-5:60.

Hu, W., Xie, N., Li, L., Zeng, X. and Maybank, S. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews* 41, 6 (2011), 797-819.

Jadhav, A., Andrews, D., Fiksdal, A., Kumbamu, A., McCormick, J.B., Mistano, A., Nelsen, L., Ryu, E., Sheth, A., Wu, S. and Pathak, J. Comparative analysis of online health queries originating from personal computers and smart devices on a consumer health information portal. *Journal of Medical Internet Research* 16, 7 (2014), e160.

Jiang, D., Pei, J. and Li, H. Mining search and browse logs for web search: A survey. *ACM Transactions on Intelligent Systems and Technology* 4, 4 (2013), Article 57.

Larson, M. and Jones, G.J.F. Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval* 5, 4&5 (2011), 235-422.

Lazar, J., Feng, J.H. and Hochheiser, H. *Research methods in human-computer interaction*. Wiley, Chichester, UK (2010).

Leech, G., Rayson, P. and Wilson, A. *Word frequencies in written and spoken English - based on the British National Corpus*. Routledge, New York (2001).

Lund, H., Bogers, T., Larsen, B. and Lykke, M. CHAOS: User-driven development of a metadata scheme for radio broadcast archives. In *Proc. iConference 2013* (2013), 990-994.

Munteanu, C., Baecker, R., Penn, G., Toms, E. and James, D. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proc. CHI 2006*, ACM Press, New York (2006), 493-502.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwartz, P., Silovský, J., Stemmer, G. and Veselý, K. The Kaldi speech recognition toolkit. Paper presented at *IEEE Workshop on Automatic Speech Recognition and Understanding* (2011).

Singhal, A. and Pereira, F. Document expansion for speech retrieval. In *Proc. SIGIR 1999*, ACM Press, New York (1999), 34-41.

Sparck Jones, K., Jones, G.J.F. and Young, S.J. Experiments in spoken document retrieval. *Information Processing & Management* 32, 4 (1996), 399-417.

Yu, D. and Deng, L. *Automatic speech recognition: A deep learning approach*. Springer, Heidelberg (2015).