occur when large-vocabulary speech recognizers transcribe unrestricted audio such as broadcast news or informal speech.

IR improvements

Similar to text retrieval, spoken document retrieval is not concerned with uninformative stop words such as *the*, *it*, *a*, and so forth. These short words occur frequently, are poorly articulated, hard to recognize, and in general add little value when searching for relevant documents. Therefore, removing them from the index improves retrieval performance. Similarly, for semantically related words, suffix stripping or stemming to a common root can facilitate matching between different word forms. For example, the words *document*, *documents*, *documentation*, and *documented* can easily confuse a speech recognizer; mapping these words to the common stem *document* typically improves retrieval.

A second IR method for preventing speech recognizer errors combines techniques such as relevance feedback and query expansion. Relevance feedback is a two-pass method. In the first pass, users enter a query and select those hits they consider relevant. In the second pass, the system uses the selected documents to compose a more powerful query. The relevance feedback from the first pass helps to improve retrieval performance in the second pass.

Pseudorelevance feedback removes user intervention by assuming that the top documents from the first pass are relevant and then using the two-pass method. In general, pseudorelevance feedback is not as effective as traditional relevance feedback, but it speeds up the search significantly.

Query expansion uses semantically related terms to expand the query. For example, query expansion might augment the query *George Bush* by adding *White House* and *President*, which the search can extract from offline text collections. Another form of query expansion uses acoustic similarity to account for possible mistakes in the speech recognizer.

Speech recognition improvements

Word-based speech recognition systems use preset vocabularies including 60,000 to 100,000 words.[3] By definition, the system cannot hypothesize words outside this vocabulary. While a vocabulary of 100,000 words includes most spoken words, every document includes a small percentage of out-of-vocabulary (OOV) words that are likely to be content-bearing terms, and not including them has an adverse effect on retrieval performance.

To circumvent this problem, the system can tailor the vocabulary by examining documents related to the task. For example, a speech recognizer used for court hearings could use legal documents to learn the appropriate dictionary words. While these specialized vocabularies can reduce the number of OOV words, they cannot guarantee their elimination.

Word spotting is an attractive approach because the systems are simple and have low computational requirements.

Word spotting. An alternative to large-vocabulary transcription is word spotting. A word-spotting system has a limited vocabulary that typically includes fewer than 50 keywords selected for a particular task. The word spotter transcribes the audio material that passes through it as a predefined keyword, other speech, audio, or silence. The word-spotting approach is attractive because the systems are simple and have low computational requirements. A disadvantage is that if a new search word is introduced, the word spotter must reprocess the entire document.

Subword recognition. The vocabulary limitations of the word-spotting approach have led to a number of open vocabulary-indexing strategies based on subword recognition. Rather than recognizing spoken words, these approaches recognize subword units—typically, phonemes or syllables—from which all words are constructed. The IR system decomposes search terms into their constituent subword strings, then scans the recognized terms for strings corresponding to the search unit.

Retrieval systems can use two approaches to perform a subword match between query terms and documents: *n*-gram matching and approximate- or fuzzy-string matching. In *n*-gram matching, the system extracts fixed-length phoneme sequences from the search words and scans the sequences for these *n*-grams. Approximate string matching substitutes, inserts, or deletes phonemes. These replacements take into account the most likely errors observed on training data. As the possibilities for a match increase, the search will find more relevant documents at the cost of more false positives. Clearly a tradeoff is needed to maximize retrieval performance.

Thus far, we have assumed that the retrieval system represents documents as a linear sequence of (sub)words. Other alternative intermediate representations are graphs or word lattices, which are readily available because large-vocabulary speech recognizers often use them.[4] Lattice nodes repre-