Title	
Author(s)	
Citation	
Issue Date	2001-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1550
Rights	
Description	, ,

Information retrieval from spoken document database

Yuki Maeda

School of Information Science,
Japan Advanced Institute of Science and Technology

August 15, 2001

With the enlargement of memory capacity and the popularization of broad-band networks, the digitalization of spoken documents is beginning. However, there is not enough good methods to retrieve spoken documents. Thus, researches about spoken document retrievals techniques using speech recognition are being carried out. But there is a problem of restriction on the number of vocabularies which cab be recognized. Then, in order to remove such restriction on the number of vocabularies, paying attention to phonemes of words, researches which use the degree of similarity between words based on the degree of similarity between phonemes are being carried out. In the researches, retrieval techniques without vocabulary restrictions of speech recognition for news story spoken documents are being studied. The retrieval methods by the researches use confusion matrix (CM) to match query phonemes with document phonemes.

In order to transcribe spoken documents in kana-kanji characters , vocabularies used in spoken documents must be contained in the dictionary of a speech recognizer. However, there is a problem that new words always appear. In this research, we try to use phonemes for information retrieval from spoken document database. Since a dictionary is not required for phoneme recognition, retrieval without vocabulary restriction is possible. Here, RWCP speech news corpus and the large lexical Japanese speech recognizer Julius was used to transcribe phonemes. The error rate of the phoneme transcription is about 34%.

Retrieval queries are given as phoneme sequences. Distance between phoneme sequences of spoken articles generated by the speech recognizer and query phonemes sequences are calculated, and whether there is any relation is judged. However, since there are recognition errors, it is necessary to search not only phoneme sequences that are successfully recognized but also phoneme sequences where some phonemes are missed and some are added. For modeling recognition errors, using recognized phoneme sequences

and manually transcribed phonemes, we calculated probabilities of how phonemes are confused, and stored them in a CM. Distance between query phonemes and phonemes in articles is calculated using the CM. Moreover, for comparison with the method using CM, we try to matching phonemes using the binominal method(BM) where the matching score is assigned to one when two phonemes are the same and 0 otherwise. Distance between query phoneme sequences and article phoneme sequences are calculated using the continuous DP matching method for speech recognition. Moreover, we extended the DP matching method to treat probability. Using the method, we calculate distance between queries and articles and judged whether queries relate with articles.

In order to verify the validity of the retrieval technique, the spoken document full-text search engine JAISearch was implemented as CGI which runs on a web server. We chose 91 articles arbitrarily as the training data from all 246 articles uttered by six speakers so that as the number of articles is averaged for every speaker. 89 articles were chosen arbitrarily as the evaluation data, and 50 queries was prepared. The number of mean phonemes per one query are 6.46. The number of retrieved articles per one query are about 11 articles. First we tried retrieval using exact matching regardless of recognition errors. The result was recall 32%, precision 77%, F-measure 24%. Next, we used the continuous DP and BM. The result was recall 51 %, precision 59% and F 29%. The third, we used the continuous DP and CM. The result was recall 65%, precision 52% and F 31%. Finally we used the probabilistic continuous DP and CM. The result was recall 62%, precision 68% and F 36%. The best combination was CM and the probability continuation DP matching.