Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents

Michael J. Witbrock and Alexander G. Hauptmann

Carnegie Mellon University

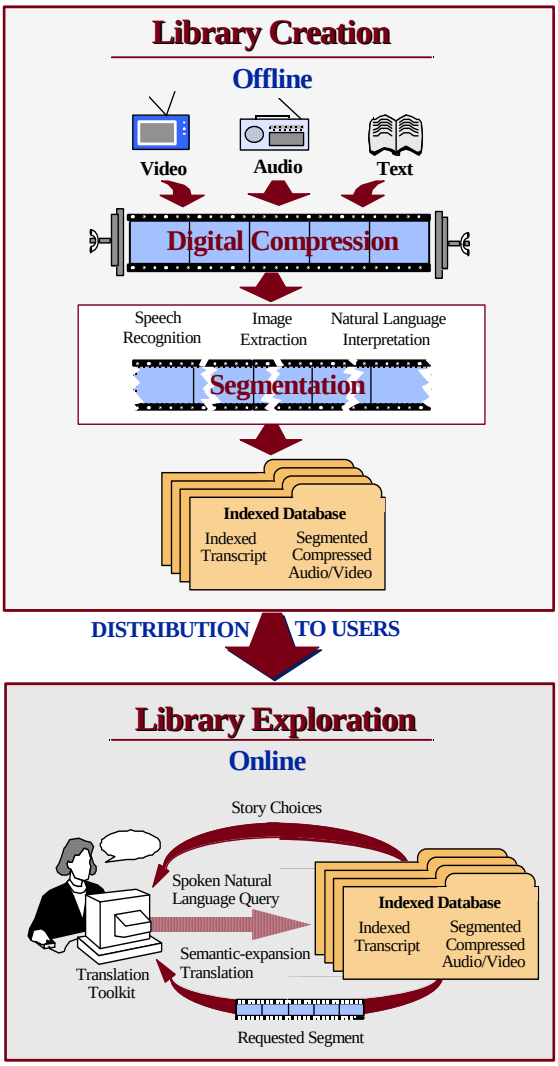

Figure 1. Overview of the Informedia Digital Video Library System

ABSTRACT

Searching for relevant material in documents containing transcription errors presents new challenges for Information Retrieval. This paper examines information retrieval effectiveness on a corpus of spoken broadcast news documents. For documents transcribed using speech recognition, a substantial number of retrieval errors are due to query terms that occur in the spoken document, but are not transcribed because they are not within the speech recognition system's lexicon, even if that lexicon contains twenty thousand words. It has been shown that a phonetic lattice search in conjunction with full word search regains some of the information lost due to out-of-vocabulary words. In this paper an efficient alternative to this search is proposed that does not require a complete search of the phoneme lattices for all documents at run-time. By using fixed length strings of phonemes instead of phonetic lattices, an information retrieval system can search the phoneme space of a spoken document just as efficiently as a normal word document collection. Experimental evidence is presented that this technique permits the system to recapture some of the information lost due to out-of-vocabulary words in the speech recognition transcripts.

INTRODUCTION TO INFORMEDIA

Vast digital libraries of video and audio information are becoming available on the World Wide Web and elsewhere as a result of emerging multimedia computing technologies. However, it is not enough simply to store and play back information as many commercial video-on-demand services intend to do. New technology is needed to organize and search these vast data collections, retrieve the most relevant selections, and permit them to be reused effectively.

The Informedia digital video library project is developing a large digital library of video and audio material. Through the integration of technologies from the fields of natural language understanding, image processing, speech recognition, information retrieval and video compression, the Informedia digital video library system [Wactlar96, Informedia95, Christel94] allows a user to explore multimedia data in depth as well as in breadth. An overview of the system is shown in Figure 1. The process automatically segments hours of video programming into small coherent pieces and indexes them according to their multimedia content. Users can actively explore the information by finding sections of content relevant to their search, rather than by following someone else's path through the material or by serially viewing a single large chunk of pre-produced video. This active exploration is far more flexible than that provided by video-on-demand, where only one way of viewing the content is permitted. It is also more flexible than the interfaces provided by the current

generation of educational CD-ROM's, where users follow a designed path through the material in a more or less passive manner. The goal in Informedia is have the computer serve as more than just a sophisticated video delivery platform. The Informedia Digital Video Library provides the user with a tool with which to assemble, from a large corpus, an instructive set of video segments relevant to a particular information need. Using this tool, a large library of video material can be searched with very little effort.

The Informedia project is developing new component technologies and embedding them in a video library system primarily for use in education and training. To establish the effectiveness of these technologies, the project is establishing an on-line digital video library consisting of over a thousand hours of video material.

News-on-Demand [Hauptmann96] is a particular collection in the Informedia Digital Library that has served as a proving ground for automatic library creation techniques. In News-on-Demand, complete automation is the principal goal. Motivated by the timeliness required of news data and the volume of material to be indexed every day, the project has applied speech recognition, natural language processing and image understanding to the creation of a fully content-indexed library and to interactive querying. While this work is centered around processing news stories from TV broadcasts, the Informedia library creation process exemplified in News-on-Demand represents an approach that can make any video, audio or text data more accessible.

A basic premise of the Informedia project was that speech recognition generated transcripts could make it possible to search for and retrieve multimedia material [Hauptmann96]. The Informedia Digital Library System depends on text transcripts that allow effective indexing and retrieval of segments relevant to a query. If perfect, manually created transcripts were available, the success of information retrieval in the Informedia Digital Library System would be assured; there are many examples of successful document retrieval systems. However, large amounts of the video and audio data in the real world do not have associated perfect transcripts. It is therefore necessary to develop and evaluate efficient techniques for information retrieval that can be applied to large collections of imperfectly, and automatically, transcribed transcripts.

This paper concerns itself with evaluating the effectiveness of retrieval from automatically transcribed documents. After a review of other work done in information retrieval of spoken documents, the metrics used in this paper are described, together with the data used in the experiments. This paper reports on some initial experiments which compared retrieval effectiveness of spoken documents transcribed manually and by speech recognition. An analysis of the sources of errors in the information retrieval system, showed that words not in the lexicon of the speech recognition system constituted the single largest source of retrieval error, since they were completely missing from the transcripts. A review of other approaches to this so-called out-of-vocabulary (OOV) problem in the literature indicated that searches that take the phonetic representation or the words into account, instead of the usual orthographic word representation, can partially alleviate the OOV problem. The two main approaches to using phonetic matching, indexing fixed length phonetic strings, and searching a phone lattice, are reviewed in the next section. In this paper, an efficient solution to the problem through a combination of phonetic sub-string matching and full word retrieval is described.

PREVIOUS RESEARCH IN IR FROM SPOKEN DOCUMENTS

Some early work in information retrieval from a corpus of spoken documents was performed at ETH Zurich [Schäuble95]. Without the aid of a large vocabulary continuous speech recognizer, Schäuble and Wechsler resorted to a phonetic string representation. Using five thousand phonetic strings of lengths varying between three and six phonemes instead of words, they reported precision and recall from a relatively small database of Swiss radio broadcast news. Although the Hidden Markov Model (HMM) phone recognition system they used achieved reasonably accurate performance, recognizing 60-70% of phones correctly, more information could be brought to bear on this recognition task by using a strong language model as is common in large vocabulary recognition. Schäuble and Wechsler's phone recognition system used biphone probabilities to constrain recognition, but a full large vocabulary recognition system can use word trigram probabilities to augment the information extracted from the audio. In fact, the Cambridge work described in the following paragraph has shown that a large vocabulary recognition system, if used alone, enables better retrieval effectiveness than an all-phone transcription.

James [James95, James96] performed a first evaluation of the benefits of combining both recognized word and phoneme representation to improve information retrieval. Initially, he attempted a real-time reprocessing of the corpus by the speech recognition engine using the words in the query that were previously missing in the recognizer's vocabulary, but found this to be extremely inefficient. Instead, James developed an approach in which a statically computed phone lattice was searched, during retrieval, for matches to the words in the query. Speech recognition systems, because of the difficulty in assigning a unique transcription to an utterance, conventionally produce a lattice giving the probabilities of the most likely few phones or words at each time. Usually, the most probable path through this lattice is output as the recognition result. However, by keeping a reasonably deep phone lattice around, the system can retain the possibility of matching with several similar sounding phone strings at each point in time. When a novel word, such as a proper name, was input as part of a search to James' system, the entire phone lattice for all spoken documents was scanned to see whether a path matching the phonetic representation of the new word could be found in the lattice for any document. Although this scan can be done quite quickly, it still requires searching linearly through the complete document collection. Thus it is impractical for all but small corpora and infrequent searches.

Nevertheless, a substantial body of subsequent work has been done at Cambridge to demonstrate the utility of this sort of matching when it is available to a retrieval system. Jones *et al.* [Jones96, Brown95] used a specially constructed test set of 50 queries and 300 voice mail messages, from 15 speakers, designed to have on average 10.8 highly relevant documents per query. They measured precision at rank 5, 10, 15 and 20 and reported the average precision. For this data, the best performance on a hand-transcribed version of the data was an average precision of 36.8%, given a speech recognition error rate of 47%. Relative precision for the best speech data was 85.6% of the text retrieval precision. This figure was achieved by combining a speech recognizer

transcript (based on a 20,000 word North American business news language model) with a phone-lattice scanning word-spotter based on speaker independent biphone models. A number of techniques for combining the two representations were evaluated. The various combination techniques included searching for only the OOV words in a phone lattice, searching for all words in the phone lattice as well as in the word transcripts, combining raw scores or complete document rankings, and combining the results of the word transcript search and the phonetic lattice search with different normalization formulas. The different combination techniques were almost equivalent at about 82-85% relative precision compared to perfect text retrieval, and all were better than the individual word search (72% of perfect text precision) or phone lattice scanning search (75% of perfect text precision) separately. It should be noted that these results are comparable to the results given in this paper in terms of the techniques used, but not in terms of the data sets. Because each corpus has significantly different characteristics, one should resist the temptation to compare precision values. In particular, because of substantial differences between the corpus used in the present work, and the Cambridge corpus, the four point rank precision and recall measures used by Jones *et al.* are not entirely suitable, and will not be reported.

The present work attempts to extend the advantages of mixing large vocabulary recognition with phonetic matching to a system that can handle very large corpora. It avoids the necessity for an expensive lattice search by augmenting a word based index with one based on phonetic sequences, and it avoids some of the accuracy problems of all-phone recognition by generating these sequences from the output of the more constrained large vocabulary speech recognition system.

EVALUATING RETRIEVAL EFFECTIVENESS IN THE FACE OF RECOGNITION ERRORS

In the current work, information retrieval experiments were performed using data sets consisting of perfect text transcripts and transcripts created by the Sphinx-II speech recognition system [Hwang94]. Two evaluation experiments were carried out. Firstly, the magnitude of the effect on retrieval effectiveness of the speech recognition system's limited vocabulary was estimated. Then a new system including phonetic sub-strings, along with the original automatic transcription was evaluated, showing that it was possible to partially recover from limited vocabulary using this method of phonetic patching. The details of these experiments are given in the following sections.

THE INFORMEDIA SEARCH ENGINE

The Informedia search engine, SEIDX, does statistical document retrieval based on the vector-space retrieval model [Salton71, Witten94]. Each document and each query is represented as a weighted sparse vector, with one element per possible term in the lexicon. Retrieval is done by ranking documents by the dot product between the query vector and the document vector. Two different weighting schemes were adopted for the elements in the document vectors. The first, and simpler scheme simply assigned to each term element in the vector the *term frequency (tf)*, the count of occurrences of that term in the document. This count is normalized by the relative importance of the terms in distinguishing documents. This former is achieved using the *document frequency (df)* of a term, the relative proportion of documents in which the term occurs at least once. Words that occur in many different documents are considered less likely to be useful for identifying a particular document. Elements for particularly common "*stop*" words, such as "a" or "the" were replaced by zero. The weighting commonly referred to as TFIDF (term frequency by inverse document frequency) is given by the formula:

$$tf \quad \times \quad -\log(df)$$

The more complex weighting scheme attempts to normalize the terms both by their relative importance in distinguishing documents and so that retrieval is not biased in favor of documents that have more query terms simply by virtue of being longer. Each element in the document vector has the following form:

$$\frac{tf \quad \times \quad -\log(df)}{dl \times \sum_{i=1}^{dl} 1/(df_i)}$$

The numerator is again the typical TFIDF weighting, and the denominator is the *document length* (dl), the number of words in the document, multiplied by the total inverse document frequency weight for the document. This latter term amounts to cosine normalization for the vector, and has proved helpful for speech documents, where large numbers of relatively rare words may be inserted by the recognition system. For the system using this scheme, stop words and stemming [Porter80] are also used.

INFORMATION RETRIEVAL METRICS USED

The standard metrics for retrieval effectiveness in the information retrieval literature are precision and recall [Salton71]. Precision is defined as the number of correct (relevant) hits returned by the system divided by the number of total hits returned to the user. Recall is defined as the number of correct hits returned to the user divided by the number of hits a perfect retrieval should have returned. A conventional way of evaluating retrieval effectiveness is to measure the average precision at a variety of recall percentages. In this paper, eleven-point average interpolated precision, as described, for example, by Witten *et al* [Witten94] will be reported. However, it is difficult to compare precision results of different queries in different document collections. Therefore, retrieval effectiveness for automatically transcribed spoken documents is conventionally also reported as the percentage to a comparable text retrieval system applied to perfect transcripts [Jones96]. Reporting the relative precision to retrieval from perfect text transcripts also has the advantage of abstracting from a particular metric and search engine used.

The precision/recall metric has the drawback that a person (or, preferably, a number of people) must manually provide relevance judgments for the test data. This is extremely time-consuming and is therefore usually only done thoroughly for small data sets. Any results are simply assumed to scale to larger sets. Within the Informedia project, an effort is being undertaken to measure precision and recall for a data set of 602 news stories given a list of 105 queries. This involves having each human judge make 63210 relevance judgments. To date, a full set of these evaluations has been completed only by a single judge; however, partial sets of evaluations have been made by several judges.

Making these relevance evaluations is not an easy task, even for human judges. Over a subset of 100 stories, for which three judges completed relevance judgments for all 105 queries, the judges agreed on 10443 judgments and disagreed 57 times. However, of the 85 cases where at least one judge thought a story was relevant to a query, the two other judges agreed only 28 times.

DATA

This section discusses the data used in the experiments. The first data subset consisted of manually created transcripts obtained through the Journal Graphics (JGI) transcription service, for a set of 105 news stories from 18 news shows broadcast by ABC and CNN between August 1995 and March 1996. The shows included were ABC World News Tonight, ABC World News Saturday and CNN's The World Today. The average news story length in this set was 418.5 words.

For each of these shows with transcripts, a speech recognition transcript was generated from the audio using the Sphinx-II large vocabulary continuous speech recognition system [Hwang94] running with a 20,000 word dictionary and language model based on the Wall Street Journal from 1987-1994. Sphinx-II is a fairly standard Hidden Markov Model (HMM) based recognizer.

Speech recognition for this data has a 50.7% Word Error Rate (WER) when compared to the JGI transcripts. WER measures the number of words inserted, deleted or substituted divided by the number of words in the correct transcript. Thus WER can exceed 100% at times.

The Journal Graphics transcription service also provided human-generated headlines for each of the 105 news stories. Each headline was matched to exactly one news story. The headlines were used as the query prompts in the information retrieval experiments. Thus the rank of the correct story was defined as the rank, returned by the search engine, of the news story for which the headline used as the query was created by JGI. Recall that this does not ensure that no other story is relevant to the title. In fact, in the 63,210 relevance judgments, a human judge assigned an average of 1.857 relevant documents to each headline. The average length of a headline query was 5.83 words.

In all the experiments described here, the stories being indexed were segmented by hand. Automatic segmentation methods could be expected to generate errors that decrease retrieval effectiveness.

This set of 105 text or speech recognized stories was augmented with 497 "distracter" Journal Graphics transcripts of news stories from ABC and CNN in the same time frame (August 1995 - March 1996. Corresponding speech transcripts were *not* obtained for this set. These distracter news transcript texts had an average length of 672 words per news story.

The comparison described below used the 602 story corpora to evaluate retrieval effectiveness for the manual transcripts and the speech recognition transcripts in the set. Recall that only for the 105 stories corresponding to the headline queries were the two transcription sources available, with the remaining 498 "distracter" stories coming from "perfect" manual transcripts. Thus the data set was biased against the speech recognition data in that it mixed perfect text transcripts, some of which may have been relevant to the query headline, with the targeted speech recognized transcripts. Since the speech recognized transcripts can be expected to have lost some query terms to recognition errors, their relevance ranking is likely to have been somewhat lower than that of a comparable, relevant text story. Because of the limited number of stories for which both speech and manual transcriptions were available, accepting this bias was necessary to permit experimentation on a sizable retrieval corpus. This bias is also present, of course, in the data sets with larger numbers of distracters.

EFFECT OF THE RECOGNITION LEXICON

The first three rows of Table 1 give the results of retrieval experiments for corpora for which 105 of the stories were produced in three different ways. The first row shows that for careful human-generated transcripts, the eleven-point interpolated precision for the better retrieval system was 0.799. Transcripts produced using speech recognition reduced this figure to 0.644, a 19.4% reduction. In the third row of the table, all the words in the 105 careful transcriptions that were outside the recognition system's 20,000-word vocabulary were omitted, and the corpus was re-indexed. The interpolated precision for this corpus was 0.692. 31% of the decrease in retrieval effectiveness in going from manual to automatic transcription can therefore be attributed to limitations in the speech recognition lexicon, regardless of any other speech recognition errors that might be present.

Table 1: Improvements in eleven-point average precision for speech recognized stories using phoneme recognition. This experiment is based on a set of 602 stories. For each condition, a base-line (TFIDF + stop words) is shown, as well as the best information retrieval using TFIDF, stop words, document length normalization, document weight normalization and suffix stripping. The conditions contrast words from manually transcribed text, words from a 20,000 word speech recognition system, words from the manual transcripts after the out-of-vocabulary words were removed, retrieval given only a phonetic representation of the text transcript, and retrieval given only a phonetic representation of the speech recognized transcript. The bottom rows show the improvements in precision obtained when both words and phonemes are used for retrieval.

	TFIDF + stop words		**Full system with all IR features**	
Type of transcription	Average Precision	% Text retrieval	Average Precision	% Text retrieval
Words from Text	0.570	100%	0.799	100%
Words from SR	0.330	57.8%	0.644	80.6%
Words from Text excluding words not in SR dictionary	0.435	76.3%	0.692	86.6%
Phonemes from Text	0.508	89.1%	0.737	92.2%
Phonemes from SR	0.325	57.0%	0.600	75.1%
Text words, Text Phonemes	0.688	121%	0.804	101%

SR words + SR Phonemes	0.457	80.2%	0.676	84.6%
Text-SR OOV + Phonemes	0.592	104%	0.747	93.5%

MIXING LEXICAL AND PHONETIC STRING INDEXING

It is often the case that when speech recognition fails on a particular word, whether because the word is out-of-vocabulary, or for other reasons, the hypothesized word differs from the correct one by way of only one or two changed phonemes. If the retrieval engine could match on the remaining phonemes, it might be able to correctly retrieve the document. The technique presented here works in exactly this manner. Instead of forming a phone lattice, or using an all-phone (as opposed to word) recognition system, the system starts with the normal automatically generated transcript of words. Using the Sphinx-II lexicon a new phonetic transcript is produced by replacing each word with its phonetic pronunciation. From this phonetic transcription all phonetic sub-strings between 3 and 6 phones in length are produced and compiled into an inverted index for this phonetic transcript corpus. This index has a larger vocabulary than usual, but this does not present a serious problem for the retrieval system. During retrieval the scores of the documents in this corpus are merged with the documents retrieved from the word transcript corpus. The combined, weighted scores provide the new ranking for the final list of retrieved documents.

During retrieval, the words in the query are converted to a phonetic representation, in this case using a pronunciation server that is not limited to just the recognition vocabulary, and again, phonetic sub-strings are extracted. These are used as a phonetic sub-string the query. Search proceeds in parallel over both the word document index and the phonetic transcription index, with the results merged.

The results of the purely phonetic sub-string searches on speech recognized documents, along with similar searches performed, for comparison, on purely manual transcriptions, are given in Table 1, rows 4 and 5.

The final three rows of Table 1 show the improvements that were obtained after combining the word and phonetic searches. While retrieval from speech recognition transcripts (combining phonetic sub-strings and word) is still worse than retrieval from perfect text, the difference is less. It is interesting to note that phonetic retrieval also appears to help retrieval from perfect text, presumably due to stemming effects that are recaptured by phonetic sub-string transcription.

CONCLUSIONS

Due to errors in recognition for current speech recognition systems, retrieval effectiveness for speech-recognized documents is systematically reduced compared to text retrieval. An important component of this decrease is due to words in the documents that the recognition system cannot possibly get right, because they are not included in its fixed-size vocabulary. In these cases similar sounding words will often be generated. By converting the speech transcription into a phonetic sequence, and by searching an inverted index of subsequences of this transcription along with the word by word transcription, a good amount of the retrieval effectiveness lost to OOV words can be recovered. By using an inverted index instead of a lattice scanner to search the additional phonetic information, these benefits can be efficiently extended even to very large speech corpora.

REFERENCES

[Brown95] Brown, M. G., Foote, J. T., Jones, G. J. F., Spärck Jones, K., and Young, S. J. "Automatic Content-based Retrieval of Broadcast News," *Proceedings of ACM Multimedia*. San Francisco: ACM, November, 1995, pp. 35-43.

[Christel94] Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S., and Wactlar, H.", "Informedia Digital Video Library", Communications of the ACM", 38 (4), April 1994, pp. 57-58.

[Hauptmann96] Hauptmann, A.G. and Witbrock, M.J., Informedia News on Demand: Multimedia Information Acquisition and Retrieval, in Maybury, M. T., Ed, *Intelligent Multimedia Information Retrieval*, AAAI Press/MIT Press, Menlo Park, 1996 (In Press).

[Hwang94] Hwang, M., Rosenfeld, R., Thayer, E., Mosur, R., Chase, L., Weide, R., Huang, X., and Alleva, F., "Improving Speech Recognition Performance via Phone-Dependent VQ Codebooks and Adaptive Language Models in SPHINX-II." *ICASSP-94*, vol. I, pp. 549-552.

[Informedia95] http://www.informedia.cs.cmu.edu/

[James95] James D. A., The Application of Classical Information Retrieval Techniques to Spoken Documents. Cambridge University Ph.D. thesis, 1995.

[James96] James D. A., System for Unrestricted Topic Retrieval from Radio News Broadcasts. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Atlanta, GA, USA, May 1996, pp. 279-282.

[Jones96] Jones, G.J.F., Foote, J.T., Spärck Jones, K., and Young, S.J., "Retrieving Spoken Documents by Combining Multiple Index Sources", SIGIR-96 Proceedings of the 1996 ACM SIGIR Conference, Zürich, Switzerland.

[Porter80] Porter, M.F. 1980, An algorithm for suffix stripping. *Program*, 14(3) 130-137.

[Salton71] Salton, G., Ed, "The SMART Retrieval System", Prentice-Hall, Englewood Cliffs, 1971.

[Schäuble95] Schäuble, P. and Wechsler, M. "First Experiences with a System for Content Based Retrieval of Information from Speech Recordings," IJCAI-95 Workshop on Intelligent Multimedia Information Retrieval, Maybury, M. T., (chair), working notes, pp. 59 - 69, August, 1995.

[Witten94] Witten, I.H., Moffat, A., and Bell, T.C., "Managing Gigabytes : Compressing and Indexing Documents and Images", Van Nostrand Reinhold, 1994.

NOTES

Table 2. Improvements in the average rank of the correct story from speech recognized stories using phoneme recognition. This experiment is based on the set of 602 stories. For each condition, a base-line (TF + stop words) is shown, as well as the best information retrieval using TFIDF, stop words, document length normalization, document weight normalization, proximity weighting and suffix stripping. The conditions contrast words from manually transcribed text, words from a 20,000 word speech recognizer, words from the manual transcripts after the out-of-vocabulary words were removed, retrieval given only a phonetic representation of the text transcript, and retrieval given only a phonetic representation of the speech recognized transcript. The third column shows the improvements in average rank of the correct story obtained when both words and phonemes are used for retrieval.

Transcript type	TFIDF + stop words	Best IR System
Words from Text	0.612	0.859
Words from SR	0.349	0.702
Words from Text less OOVs for SR	0.478	0.752
Text-SR OOV + Phonemes	0.645	0.819
Phonemes from Text	0.544	0.805
Speech Word + Phonemes	0.486	0.746
Text Word + Phonemes	0.737	0.872
Phonemes from SR	0.360	0.680

Should say what the OOV rate *is*.