

OPTIMAL PARAMETERS FOR SEGMENTING A STREAM OF AUDIO INTO SPEECH DOCUMENTS

Gerard. Quinn & Alan. Smeaton

School of Computer Applications
Dublin City University
Dublin, Ireland.

ABSTRACT

Indexing and retrieval of spoken documents is a desirable feature in a digital library as there is a wealth of contemporary information available to us uniquely in this medium. In this paper we describe experiments carried out on the TREC 6 SDR collection to determine the optimal parameters for our speech IR system. In the TREC task the data corresponded to complete news broadcasts and the boundaries between news stories were marked up manually. In an operational news speech retrieval system such as our own, news story boundaries are not always part of the speech data, making it difficult to automatically detect shifts in stories being broadcast. We describe our approach of splitting the stream of audio into speech documents of fixed length and analyse the results from each method culminating in an optimal solution.

1. INTRODUCTION

Up until about a decade ago most of the information stored on computers was text, while this is still the case, multimedia information (a combination of text, audio and images) now constitutes a large proportion of the stored information. Information Retrieval (IR) on text documents has now reached a stage where improvements in effectiveness are usually quite small. For this reason and because of the sudden explosion of multimedia type information being stored on computers more attention has been paid to IR from audio [1] [2] [3] [4].

An initial method of extracting information from speech was word spotting [5] [6] a technique of recognising a set of words from a small vocabulary which are content bearing. Another method was called Topic Identification [7] [8], a technique of partitioning an archive of spoken messages into a small number of predefined topics. These techniques allow for only a small vocabulary of words to be searched for. With the advances in speech recognition it is now possible for systems to accurately recognise most words spoken in studio environments [9] [10] [11] [12]. These systems are

called Large Vocabulary Continuous Speech Recognition systems (LVCSR). The output of LVCSR systems is either a word transcript or phonetic transcript. In layman terms matching a sequence of phonemes against a pronunciation dictionary generates word transcripts and the best matching word is chosen to represent the sequence. Phonetic transcripts don't use a dictionary and a phoneme is chosen based on the probability of it occurring after a phone and before another. Both have their advantages and disadvantages. The main disadvantage of word transcriptions is the problem of Out of Vocabulary words (OVV), words not in the pronunciation dictionary, which then recognised incorrectly. This problem doesn't occur at the speech recognition stage of phoneme transcripts although it has to be handled at the query stage. The main problem with this method is the problem of matching a user's query against a sequence of phonemes, where there are incorrectly matched phonemes, deleted phonemes and extra phonemes added. Although the accuracy of these LVCSR systems is high for read speech, the accuracy drops considerably when the speech is spontaneous or conversational speech. Most of the research in speech recognition is currently concentrating on this problem.

With the advances made in LVCSR systems it is now possible to implement a content based information retrieval system on audio archives. This involves a two stage process, the first being the recognition of the speech and the second is to apply proven text IR techniques [13] [14] [15] on the output of the LVCSR system, where the output is either text or phonemes.

Many IR research groups are now participating in speech IR without a speech recognition system. This is mainly due to the TREC (Text Retrieval Conference) track called Spoken Document Retrieval (SDR). This track allowed for research centres with or without a speech recognition system to investigate how effective content based retrieval can be applied to speech and since all participating groups were working with the same data it allowed for comparison of results between the different

groups. More details of the SDR are given in the next section.

2. TREC 6 SDR

The Text Retrieval Conference 6 Spoken Document Retrieval (TREC 6 SDR) test collection [16] formed the basis of the research presented in this paper. The aim of the SDR track was to bring together IR and Speech groups to work on the problem of content based retrieval from continuous speech. A synopsis of the SDR track is given here.

The SDR track involved implementing IR techniques to do known item retrieval (only one document is deemed to be relevant for each query) on a set of speech documents. There were two modes of participation in this track, full SDR and quasi SDR, where full SDR is for groups with a speech recogniser and who operate on the raw waveform of the audio data and QSDR is for groups interested in getting involved in speech retrieval but who don't at present have a speech recogniser. This latter is the route we took as our speech recogniser was only in its infancy at the time of TREC6, and also it was trained on a set of Hiberno-English news stories, so it would have been prone to high levels of recognition errors.

The corpus of data in the TREC-6 SDR task consisted of about 50 hours of speech from US radio and TV news broadcasts containing approximately 1000 stories with a set of 49 known-item search topics. There were four types of data sources provided: raw audio data (.sph), "truth" transcripts in text that were transcribed and marked up with SGML tags (.dtt), another "truth" transcription that contains the same information content from the news stories as the .dtt transcripts but with most of SGML tag information removed (.lft) and output from an IBM speech recognition system, i.e. including recognition errors [17], in the same format as LTT but with each word being bracketed with an extra SGML tag indicating its time tag occurrence (.srt).

Each participating [18] [19] group in the QSDR was required to run 49 known-item searches on the Baseline collection (.srt files) and the Reference collection (.dtt files) and submit the top ranked list of story identifiers for each topic. A known-item retrieval is one in which, for each query or topic, there is one and only one relevant story. This means that topics have to be carefully crafted and checked to ensure there is only one relevant story in the whole 50 hours of broadcast archive. When it comes to evaluating the performance of various "known item retrieval runs" conventional precision and recall is inappropriate with only one relevant document (story) per topic so TREC used

performance methods based on the ranks given to the known item for each search topic. Three of the methods used were:

- 'Mean Rank', the average rank at which the known items were positioned within the ranked lists of returned story identifiers.
- 'Mean Reciprocal', the mean of the reciprocal of the rank at which the known items are retrieved for each of the topics.
- 'Percentage at rank 1', The percentage of the topics that are returned at rank 1.

3. EXPERIMENTS

The objective of our participation in the QSDR track in TREC-6 and our Post TREC experiments which are reported in this paper were to develop the best input parameters to use in our "Taiscéaláí" system [19], which carried out content based retrieval on an archive of RTÉ (the Irish National radio station) radio news broadcasts. In the TREC task the data corresponded to complete news broadcasts and the boundaries between news stories were marked up manually. In an operational news speech retrieval system such as our own, news story boundaries are not normally part of the speech data, making it hard to automatically detect shifts in stories being broadcast. Our approach to handling the fact that a single broadcast contains multiple concatenated stories was to divide the entire broadcast into fixed length windows, either overlapping or not overlapping and with or without a weighting function with the content of the windows being words or overlapping triphones (triples of adjacent phones). Each window was treated as an independent document for indexing and retrieval in a traditional information retrieval system using TF*IDF (Term Frequency Inverse Document Frequency) weighting.

Table 1 gives details of the parameters we varied during our experiments. Words are the words in the baseline and reference files and triphones are overlapping groups of three phonemes generated by our pronunciation dictionary for each word in the baseline and reference files. The pronunciation dictionary had approximately 160,000 terms in the format of 'computer' represented as '*k a x m p y u u t a x r*'. The size of the windows, the documents which are indexed, are of fixed length for each of the experimental runs, where the fixed lengths are 15, 30, 60 or 90 seconds, the time tags in the reference and baseline files being used to calculate the size of the windows. The windows either overlap or do not overlap, and if they overlap the size of the overlap is 1/3 the size of the window i.e. 5, 10, 20 or 30 seconds. We used three weighting formulae when

generating the windows which are termed 'pyramid', 'middle tower' and 'equal' weights. The Pyramid assigns most of its importance to triphones in the middle of each window with sliding scales each side. This was achieved by linearly increasing the number of triphones or words in the middle of the window by a factor of 10 and decreasing this factor proportionally as the distance from the centre increased. Middle Tower assigned the greater importance to triphones in the middle third of each window by a factor of 10. Equal weight as the name implies assigns equal weight to all terms within a window.

Indexing Unit:	Words or Triphones
Window Size:	15, 30, 60 or 90 seconds
Window Overlap:	1/3 the window size or No overlap
Weighting:	Equal, Middle Tower or Pyramid weighting

Table 1 – Details of variables used in our experiments.

In all we ran 40 different experiments varying the length of the speech documents (window size), overlap, weighting function and content of windows. The 40 runs [22] consisted of every combination of the above parameters except for middle tower with no overlap. The reason for this was time and has no bearing on the results. Each experiment consisted of creating the speech documents for the baseline and reference documents, indexing these documents using our text search engine [21] and running the 49 queries against the indexes and logging the results. The search engine is based on the Inverted File approach with special attention to efficient retrieval without losing effectiveness. Each unique triphone/word and the number of speech documents it appears in is stored in the lexicon, and also stored in the lexicon is an offset value into the posting file for each triphone/word. The posting file contains an entry for every document that each triphone/word appears in. The data stored in these entries include the document identifier and the number of times the triphone/word has occurred in that document. The number of unique documents each triphone/word appears in and the number of occurrences in each document are used when retrieving spoken documents using the TF*IDF term weighting method (Term Frequency Times Inverse Document Frequency). We did not use stop word removal (removal of common triphones/words) during the indexing phase as it has been shown that retrieval

performance was improved by less than one percent if stop word removal was done on triphones [23].

4. ANALYSIS

The analysis phase involved statistically analysing the results from the SDR collection. We ran a series of ANOVA (Analysis of Variance, where variance is the standard deviation squared) tests and Mean comparison tests. As the name implies, ANOVA is a technique which enables us to test two or more population means simultaneously and it gave us details of which variable had a statistical affect and which did not. The other test we ran was Mean comparison of the individual parameters. This gave us details of the individual variables and which parameter worked best. An example being the comparison of window size and how they faired against each other, this is achieved by grouping the runs based on the window size used, getting the means of the results in each of these pools and comparing the results to figure out which performed the best.

Throughout our ANOVA tests the factor refers to the method of preparing the spoken documents for indexing by our search engine. The options of the factor are the combinations of the data type (words or triphones), window size, overlap and weighting algorithm

4.1 Analysis of the percentage of baseline runs at rank 1

Variable	F value	Sig of F.
Overlap	70.095	.000
Weight	1.315	.283
Window Size	6.724	.001
Data	8.809	.006

Table 2 ANOVA of percentage of baseline runs at rank 1 results.

The Null hypotheses are that overlapping windows, weighting algorithms, window size, and the data do not have an effect on the results. Since the significance of F for overlap is less than 5% this statistically implies that overlap has an effect on the results, this is also the case for window size and data, .01% < 5% and .6% < 5% for window size and data respectively. The significance of weighting algorithms is over 5% (28.3% > 5%) this implies that weighting does not statistically affect the results.

We now use the average mean of overlap for the baseline runs at rank 1 to extract information to figure out if having a 1/3 overlap gave better results than not having any overlap. Looking at Table 3 we see that 1/3 overlap has a mean of 54.67 and no overlap has a mean of 38.39, these numbers represent the mean of the percentage of baseline runs at rank 1 with 1/3 overlap and no overlap respectively. This implies that 1/3 overlap gives the better results. We also note that since the significance of F for overlap is very close to 0% it implies that overlap has the largest effect on the results.

Overlap	Mean
1/3 overlap	54.67
No overlap	38.39

Table 3 - Comparison of means when grouping the runs by 1/3 overlapping and none overlapping documents.

The window size has the next greatest affect on the results as its significance of F is equal to .001%. We see from Table 4 that as the window size decreases from 90 seconds to 30 seconds there is an improvement in the results. The results reach a peak at 30 seconds and then fall again when the window size is reduced further to 15 seconds.

Window Size	Mean
90 sec	41.83
60 sec	48.99
30 sec	52.24
15 sec	49.59

Table 4 - Comparison of means when grouping the runs by window size.

The final parameter that has a statistical affect on the results is the data used to index the collection, which are words or triphones. The significance of the F value is .006% and looking at the Table 5 which shows the average mean for the data we see that words have the greatest effect at 50.7 and triphones at 45.6.

Data	Mean
Triphones	45.6
Words	50.7

Table 5 – Comparison of means when grouping the runs by data used.

The weighting algorithm does not statistically affect the results but we still wanted to know which weighting algorithm had the best effect even if it was minor. We grouped all the runs which had overlapping together and got the average mean for each weighting algorithm. Table 6 shows details of means. From this data we note that pyramid weighting has the highest mean although it is not significantly different from the others.

Weighting	Mean
Middle Tower	55.1
Equal	51.78
Pyramid	57.13

Table 6 – Comparison of weighting algorithms when grouping overlapping runs by weighting algorithm used.

Analysis of the baseline ‘mean rank’ results and the baseline ‘mean reciprocal’ results showed similar results to the above [22].

4.2 Difference between baseline and reference results

From Table 7 we see that the average results from the reference runs are better than the baseline runs. This was expected as the reference runs were hand transcribed from the news broadcasts and the baseline was the output from the IBM speech recogniser which included recognition errors.

	Reference	Baseline
% at rank 1	56	48
Mean Rank	236	200
Mean Reciprocal	.65	.575

Table 7 – Average reference and baseline results

The reference average mean rank was 118% of the baseline average mean rank, the reference average

mean reciprocal was 113% of the baseline average mean reciprocal, and the reference percentage at rank 1 was 116% of the baseline average percentage at rank 1.

4.3 Baseline best runs

From the ANOVA test and mean comparison test we would assume that our best results would come from 1/3 overlapping 30 second windows with the data type being words and the weighting algorithm not really mattering, and this is the case. From the statistical tests carried out on the runs we would also assume that if the best runs for each data type (words and triphones) were separated then the above rules would still apply. That is, the best runs would have 1/3 overlapping 30 second windows with a random weighting algorithm and again this is the case as can be seen in Table 8 and Table 9.

Rank	Run No.	Mean Recip	Run Type
1	36	.66	30 sec, 10 overlap, pyramid
2	10	.65	60 sec, 20 overlap, middle
3	38	.64	15 sec, 5 overlap, pyramid
4	34	.63	60 sec, 20 overlap, pyramid
5	8	.62	30 sec, 10 overlap, middle

Table 8 – Top 5 Triphone Baseline Runs.

Rank	Run No.	Mean Recip	Run Type
1	7	.71	30 sec, 10 overlap, middle
2	27	.71	30 sec, 10 overlap, equal
3	35	.70	30 sec, 10 overlap, pyramid
4	11	.70	90 sec, 30 overlap, middle
5	29	.69	15 sec, 5 over, equal

Table 9 – Top 5 Word Baseline Runs.

It has been proven statistically that words and triphones affect the results (see ANOVA test above), with words out-performing triphones. The problem is that these tests assume that there are no OOV words. In our system (Taiscéalaí) we were not going to be able to include facilities to detect out of vocabulary words and hence our recogniser would incorrectly recognise words which it did not have in its pronunciation dictionary. So if we include the fact that OOV words would affect the results then the gap between words and triphones would then be reduced and so the decision was taken for our implementation of “Taiscéalaí” system to do triphone recognition and to deal with OOV words at query time. Where pronunciation rules can be used to convert text into phonemes for words that do not appear in a dictionary.

With our decision then to use triphones, 1/3 overlap and 30 second windows our only decision left was to decide which weighting algorithm to use. We chose to use pyramid weighting although it was statistically proven that the weighting algorithm did not have any significant bearing on the results, the reason for this was because its results were slightly better than the other two weighting algorithms.

5. CONCLUSION

In this paper we gave a description of our experiments carried out on the TREC 6 SDR collection. The aim of these experiments was to determine the best parameters to use in our “Taiscéalaí” system, which does content based retrieval on an archive of RTÉ radio news broadcasts. We gave details of our experiments carried out for TREC 6 SDR and POST TREC. The SDR document collection was marked up and segmented into news stories. This did not reflect our situation with the output of our speech recogniser for RTE which was going to be a set of phones or words (at the time of experiments it was not decided if the output would be phones or words), with no tags except for time tags and pause information and speech type, i.e. there was going to be no story segmentation. To reflect our situation we removed all tags and created documents based on document size, overlap and weighting.

We carried out 40 Post TREC experiments varying the parameters in each run. The parameters varied were; data to be indexed (words or triphones), document size (15, 30, 60 & 90 seconds), overlapping documents (no overlap or 1/3 overlap) and weighting (no weighting, middle tower and pyramid weighting). The experiments showed that overlap had the greatest statistical affect on the results and that overlapping documents out-

performed non-overlapping documents. Window size was also shown to statistically affect the results with window size of 30 seconds performing the best. The data also had a statistical affect with words out-performing triphones slightly, but the problem of out of vocabulary words (words not in the pronunciation dictionary) in new broadcasts were not taken into account when generating this data. With this problem and the fact that the speech recogniser would be much slower the decision was taken to use triphones as the output of the speech recogniser. The final parameter weighting was shown not to have a statistical affect and hence picking any weighting function would result in similar results. We decided to pick pyramid weighting because it very slightly (not statistically) out performed the others.

So the parameters picked were, data = triphones, overlap = 1/3, window size = 30 seconds, weighting = pyramid weighting. The results obtained from using these parameters were about the average at TREC 6. The percentage correct at rank 1 is 59%.

6. REFERENCES

- [1] Ng, K. "Survey of Approaches to Information Retrieval of Speech Messages", Working paper, 1996.
- [2] Ng, C., Zobel, J., "Speech Retrieval using Phonemes with Error Correction", in Proceedings of SIGIR 1998
- [3] Ng, C., Wilkinson, R., Zobel, J., "Speech Document Retrieval using Phonetic Strings", in Proceedings of SIGIR 1998.
- [4] Sanderson, M., Crestani, F., "Mixing and Merging for Spoken Document Retrieval", in Proceedings of SIGIR 1998.
- [5] Jones, K., S., Jones, G., Foote, J., Young, S. "Experiments in Spoken Document Retrieval" Information Processing & Management, Vol. 32, No. 4, pp. 399-417, 1996.
- [6] Knill, K., Young, S., "Fast Implementation of Viterbi-based word-spotting", in Proceedings ICASSP'96, pp 520-523, Atlanta, 1996.
- [7] R.C. Rose, E.I. Chang, R.P.Lippmann, "Techniques for Information Retrieval from Voice Messages" in Proceedings of IEEE ICASSP, Volume I, pp. 317-320, 1991.
- [8] McDonough, J., Ng, K., Jeanrenaud, P., Gish, H., Rohlicek, R., "Approaches to Topic Identification on the Switchboard Corpus" in Proceedings of ICASSP, pp I-385-I-388, 1994.
- [9] Young, S. "Large Vocabulary Continuous Speech Recognition: a Review", IEEE Workshop on Speech Recognition, Utah, 1995.
- [10] Woodland, P.C., Gales, M.J.F., Pye, D., Young, S.J. "The development of the 1996 HTK Broadcast News Transcription System", DARPA Speech Recognition Workshop, pp. 73-78, Chantilly, Virginia, 1997.
- [11] Woodland, P.C., Hain, T., Johnson, S.E., Niesler, T.R., Tuerk, A., Whittaker, E.W.D., Young, S.J. "The 1997 HTK Broadcast News Transcription System", 1998
- [12] Cook, G., Christie, J., Clarkson, P., Hochberg, M., Logan, B., Robinson, A., Seymour, C., "Real Time Recognition of Broadcast Radio Speech", in Proceedings of ICASSP, 1996.
- [13] Schauble, P., & Wechsler, M. "First Experience with a System for Content Based Retrieval of Information from Speech Recordings", in Proceedings of the IJCAI Workshop on Intelligent Multimedia Information Retrieval. 1995.
- [14] Brown, G., Foote, J., Jones, G., Spark K. J., Young, S. "Automatic Content-Based Retrieval of Broadcast News" in Proceedings of ACM International Conference on Multimedia, p. 35-43, San Francisco, 1995.
- [15] Wechsler, M., Munteanu, E., Schauble, P., "New Techniques for Open-Vocabulary Spoken Document Retrieval", in Proceedings of SIGIR 1998.
- [16] Voorhees, E., Garofolo, J., Spark K., J "The TREC-6 Spoken Document Retrieval Track", in Proceedings of TREC6, 1997.
- [17] Dharanipragada, S., Franz, M., Roukos, S. "Audio-Indexing For Broadcast News", in Proceedings of TREC6, 1997.
- [18] Singhal, A., Choi, J., Hindle, D., Pereira, F., "AT&T at TREC-6: SDR Track", in Proceedings of TREC6, 1997.
- [19] Schauble, P., Mateev, B., Munteanu, E., Sheridan, P., Wechsler, M., "ETH TREC-6: Routing, Chinese, Cross-Language and Spoken Document Retrieval", in Proceedings of TREC6, 1997.
- [20] Smeaton, A.F., Morony, M., Quinn, G., Scaife, R., "Taiscéalaí: Information Retrieval from an Archive of Spoken Radio News", 2nd European Digital Libraries Conference, 1998.
- [21] Kelledy, F. "Query Space Reduction in Information Retrieval", Ph.D. Thesis, Dublin City University, 1997.
- [22] Quinn, G. "Content Based Retrieval from an Archive of Spoken Radio News", M.Sc. Thesis, Dublin City University, 1998.
- [23] Ng, K., Zue, V. "Subword Unit Representation for Spoken Document Retrieval". in Proceedings of ESCA Eurospeech Conference, 1997.