

# SUMMARISATION OF SPOKEN AUDIO THROUGH INFORMATION EXTRACTION

*Robin Valenza*

*Tony Robinson*

*Marianne Hickey*

*Roger Tucker*

English Department  
Stanford University  
Stanford, CA 94305-2087  
USA  
rvalenza@leland.stanford.edu

Cambridge University  
Engineering Department  
Trumpington Street  
Cambridge, CB2 1PZ, U.K.  
ajr@eng.cam.ac.uk

Hewlett-Packard Laboratories  
Filton Rd, Stoke Gifford  
Bristol, BS12 6QZ, U.K.  
mh@hplb.hpl.hp.com  
rcft@hplb.hpl.hp.com

## ABSTRACT

Automatic summarisation of spoken audio is a fairly new research pursuit, in large part due to the relative novelty of technology for accurately decoding audio into text. Techniques that account for the peculiarities and potential ambiguities of decoded audio (high error rates, lack of syntactic boundaries) appear promising for culling summary information from audio for content-based browsing and skimming. This paper combines acoustic confidence measures with simple information retrieval and extraction techniques in order to obtain accurate, readable summaries of broadcast news programs. It also demonstrates how extracted summaries, full-text speech recogniser output and audio files can be usefully linked together through an audio-visual interface. The results suggest that information extraction based on statistical information can produce viable summaries of decoded audio.

## 1. APPLICATION CONTEXT

Managing this contemporary explosion of audio and video materials calls for intelligent strategies for indexing, summarising and otherwise condensing audio-visual (a/v) information so that it can later be accessed efficiently. The summarisation techniques presented in this paper are designed to facilitate the following applications:

- Rapid digestion of material contained in an audio file (“executive summaries”). In such applications, summaries should contain as much of the key information from the original audio as possible within a constrained length.
- Content-based browsing. Such summaries need to refer to, although not necessarily contain extensive detail about, all key information within a short compass. These summaries are linked to the corresponding portions of the full audio transcription so that more detail can be retrieved when further information is needed.
- Reduction of incorrect information from audio transcriptions for greater accuracy in information retrieval.

In order to take into account the inherent error rate of recognised speech and the lack of information about

syntactic boundaries in audio files, we approach the problem of summarising and retrieving audio information from a statistical, rather than an NLP, point of view. Textual information extraction techniques were adapted to the distinctive qualities of decoded audio. Proven statistical methods in information retrieval, such as those used in TREC-6 and 7 systems ([7], [3]), were combined with newer methods in calculating audio confidence measures to extract summary information from speech recogniser output. The most relevant portions of decoded audio were extracted and then combined to form a textual summary. The extracted portions of text were time-indexed into the original audio file so that 1. the original audio information could be retrieved and 2. the summary could be presented simultaneously as text and as audio using an a/v interface.

## 2. CORPORA

The training and test material consisted of segments of the first 95 files in the American Broadcast News (BN) corpus used in the TREC 6 Spoken Document Retrieval (SDR) experiment. The files were initially collected in 1996 by the Linguistic Data Consortium for the DARPA Hub-4 continuous speech recognition project.

The LDC manually labelled the BN files to separate the segments that contain the actual news broadcast versus those that contain advertisements or music without speech; each file contains between 9 and 36 minutes of spoken news.

The Abbot speaker-independent continuous speech recognition system was run only over the spoken portions of these audio files to generate the working corpus of decoded audio documents used in this study.<sup>1</sup>

Twenty-five additional BN files were decoded and incorporated for the information retrieval testing, giving a total of 110 files.

## 3. TYPES OF SUMMARIES

Three major types of summary “phrases” were generated:

<sup>1</sup>Speech recogniser performance was unreliable on advertisements, which usually contain many non-speech sounds.

- **N-grams.** N-grams are units of N consecutive words taken from the decoded audio. These units are not marked by any syntactic or semantic boundaries. In the summaries generated for evaluation, N ranged from 1 to 200 words. When N=20, for example, we could have the 20-gram consisting of the following 20 consecutive words extracted from the May 15, 1996 broadcast of the PRN CNN affiliate station:

DEVOTE FULL TIME TO HIS QUEST FOR THE WHITE HOUSE [SIL] THIS DECISION IS SEEN BY SOME AS A POLITICAL GAMBLE

- **Utterances.** Utterances are consecutive segments of audio delimited by manually labelled boundaries marking speaker and/or content changes. The above 20-gram is contained within the full utterance:<sup>2</sup>

IN THE MEANTIME A BOLD AND STUNNING MOVE BY PRESUMPTIVE G. O. P. PRESIDENTIAL NOMINEE BOB DOLE [SIL] WHO ANNOUNCED HE IS RESIGNING FROM THE SENATE TO DEVOTE FULL TIME TO HIS QUEST FOR THE WHITE HOUSE [SIL] THIS DECISION IS SEEN BY SOME AS A POLITICAL GAMBLE [SIL] AN EFFORT TO RE ENERGIZE AND RE INVENT HIS LAGGING CAMPAIGN [SIL] THE THIRTY FIVE YEAR CONGRESSIONAL VETERAN SAID HE LEAVES CAPITOL HILL WITH [SIL] MIXED EMOTIONS

- **Key Words.** Key words are frequently occurring single word units (1-grams). They can be used as a very simple form of topic spotting.

Summaries were generated on an m-summary per minute basis, where m could be input on the command-line or through the interface. Although extracting material based on a criteria of certain quantity of output per minute of audio is somewhat artificial, it does guarantee that the extracted information will be spread throughout the audio program and not be concentrated in a single, high-scoring segment of audio.

## 4. ADAPTATION OF STATISTICAL MEASURES

### 4.1. Acoustic Confidence

A number of acoustic conditions can affect the quality of a speech recogniser's output, including: speech from speakers whose accents do not match the acoustic models well, speech from female or child speakers (whose speech speaker independent systems do not generally handle as accurately as speech from adult male speakers [3]), speech recorded through a poor microphone, speech containing many out of vocabulary (OOV) words, and speech recorded in the presence of background noise. Because these conditions can not always be avoided or even identified, given a set of acoustic observations recorded under a set of unspecified conditions, we would like to be able to determine the effect of these conditions on the recognition output. Even when the specific acoustic

<sup>2</sup>N-grams can also cross utterance boundaries; utterance boundaries were not considered in extracting n-grams.

qualities of the speech signal are unknown, we can analyse how well the acoustic input matches the acoustic models to give us a measure of how reliable the recogniser output is for that signal. We can then assign a confidence measure to the speech recogniser output that tells us how likely the output is to have been correct. Such confidence measures indicate numerically how well the acoustic input matches acoustic models.

In an HMM/ANN speech recogniser such as Abbot, local probability estimates for a single frame of speech are combined to produce an optimal (or nearly optimal Viterbi estimated) state sequence that expresses the posterior probability of a larger unit, such as a phone, phone sequence, or a word, given the set of acoustic frames. The estimated Viterbi aligned posterior probability of a word, W, given a series of acoustic observations, X = 1,2,...x, can be expressed

$$P(W|X) \simeq \underset{state-seq}{max} \prod_n P(q_n^k|x) \frac{P(q_n^k|W)}{P(q_k)} P(W) \quad (1)$$

Hybrid HMM/ANN systems like Abbot have been shown to produce useful estimates of the posterior phone probability given acoustic data [8]; these posterior phone probabilities can then be used in calculating acoustic confidence measures [6], [11].

Gethin Williams and Steve Renals at the University of Sheffield report using acoustic-based confidence measures derived from the posterior phone probabilities of the Abbot HMM/ANN speech recogniser to verify speech recognition hypotheses [9], [10]. They define the duration normalised posterior probability of a hypothesised word as the product of the posterior probability estimates of the constituent phones. Williams and Renals determined experimentally that they achieved the most discriminating measures by normalising all phones in a word by the duration of the entire hypothesised word, rather than normalising each phone constituent by its own duration [10]:

$$CM_{npost}(q_k) = \frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} \log(p(q_k|x^n)) \quad (2)$$

$$CM_{npost} = \frac{CM_{post}(q_k)}{n_e - n_s} \quad (3)$$

This confidence measure has the additional advantage of being straightforward to compute with values available directly from the decoder output.

For this study, we appropriated this measure to assign an acoustic confidence measure to the single-best recogniser hypothesis for each word lattice.

For example, the following acoustic confidence measures were assigned to erroneously decoded words for which there was very audible noise from a weather helicopter present in the background during the acoustic frames:<sup>3</sup>

THE -2.1729

<sup>3</sup>What was really said was "[a tornado is on the] GROUND GO TO A BASEMENT..."

CONGO -1.9800  
A -2.1042  
TIME -1.4878  
MACHINE -1.7930

These acoustic confidence measures were considerably lower than those assigned to the correctly decoded words from a later portion of the same broadcast:

THEY -0.6940  
CAN -0.3401  
DESTROY -0.5927  
EVERYTHING -1.1479

demonstrating that acoustic confidence measures can help discriminate correct from incorrect recogniser output.

For the purpose of judging acoustic confidences of words, posterior phone probability, the probability of the phone given the observed acoustics, was calculated for each frame of speech. The frames corresponding to each word in the one-best hypothesis were added together and normalised by the frame length of that word. This score was then assigned as the acoustic confidence of that word.<sup>4</sup>

Used in combination with the inverse frequency scores of the type described in section 4.2, acoustic confidence measures were the basis of judgements to accept or reject words (or groups of words) for inclusion in summaries. Words were accepted if their scores were higher than the thresholds of the types described in section 4.3.

## 4.2. Inverse Frequency

Inverse frequency values were determined by the number of times a word occurs in a document divided by the number of times it occurred in the language model and were normalised by document length [7].

Although in theory the existing broadcast news corpus could be processed in bulk, to retain the ability to summarise single documents out of context, a language model that did not depend wholly on the vicissitudes of the current document set was considered attractive. In this work a language model generated from a 474,365 word frequency list from the Wall Street Journal was used.<sup>5</sup>

## 4.3. Combining the Measures

Within the document to be summarised, individual words were assigned scores derived from a weighted sum of their simple inverse frequencies and their phone-based acoustic confidence values:

$$w * \text{inverse frequency} + \alpha * \text{acoustic confidence} \quad (4)$$

The constant  $\alpha$  was generally set to 1. The values chosen for the inverse frequency weight depended on the

<sup>4</sup>The C++ program written to determine the posterior probabilities was based on code kindly provided by Gethin Williams.

<sup>5</sup>This language model was built by Gary Cook and James Christie in the Speech, Vision, and Robotics group at CUED. See [4] for more on the limitations of spoken document language models and reasons for using written-text-based models instead.

relative priority given to frequency versus accuracy. For example, summaries could be produced with relatively low error rates by setting  $w$  to a very low value and relying on acoustic confidence alone; however, achieving the lowest possible error rate (i.e. setting  $w$  to zero) might not always be the priority in summarisation – when the purpose is to include the most important information, a certain amount of inaccuracy might be acceptable.

Training the frequency weight objectively was not feasible because it depended on summarisation priorities (unique versus frequent information, precision versus accuracy, etc.), which are not precisely quantifiable. External, human evaluation of the summaries and error rates generated at different weights determined the range of useful values for  $\alpha$  and the inverse frequency weight.

N-grams and utterances were assigned scores based on normalised sums of their constituent words' scores. The highest-scoring n-grams or utterances were chosen for inclusion in summaries. User-specified thresholds (the value of  $n$ , the number of phrases per minute, the minimum phrase score, etc.) determined the number of phrases extracted for each summary.

In order to combine the best qualities of the coherence of longer units and the content-based focus of key words, summaries consisting of both a key word list and an n-gram or utterance list were generated for testing.

## 4.4. Evaluation

Summaries are inherently hard to evaluate because the quality of a summary depends both on the use for which it is intended and on a number of other, qualitative, human factors, such as how readable an individual finds a summary or what information an individual thinks should be included in a summary. Automatically generated summaries of textual material are often evaluated by subjective comparison to existing human-generated summaries of the same material [5]. However, without a corpus of existing summaries for spoken audio, this kind of evaluation was not feasible for this study. Since there was no single good way to evaluate the summaries generated, three different methods were used for evaluating different aspects of the summaries:

- The word error rate (WER) of the summaries was measured against human transcriptions of the audio. The WER for these summaries was also compared to that of the full recogniser output. The WER provided a measure of how accurate the summaries were, especially in comparison to the full recogniser output (i.e. the unsummarised transcription). See Table 1 for baseline and average values.
- A survey was completed by volunteers who evaluated several different types of automatically generated summaries. The survey indicated how well humans thought the extracted information summarised the full output.

1 utterance/min	10 gram/min	30 gram/min
11.3 %	6.15 %	9.78 %
20 10-grams/min	full recogniser	
3.42 %	25.1 %	

Table 1: The average word error rates for 95 BN files. Error rates were normalised by the number of words in the summary (or by the full text’s length in the case of the full recogniser output).

- An information retrieval test was run on the summaries to gain a measure of how well the summaries retained the key information from the full output.

As discussed in the succeeding sections, the results from all three tests were highly encouraging; these results suggest that information extraction using confidence measures is a promising method for audio summarisation.

#### 4.5. WER

The decline in WERs from the full audio transcription of summaries of all lengths generated (10-grams, 30-grams, utterances) is highly auspicious; it suggests that the selection heuristic is effective in extracting summary information that is less error-prone than the full recogniser output; a lower error rate implies a greater degree of readability. Both visual inspection and the survey results bear out these observations.

Figure 1 shows that the error rate normalised by summary length increases with the length of the n-gram. This indicates that the confidence measures are successful in picking out the most accurate words. The longer the required consecutive n-gram length, the more inaccurate words that will have to be included to get a consecutive unit of the desired length. The n-gram error rate approaches the global error rate of the file as n approaches the number of words spoken per minute, which averages about 200 wpm.

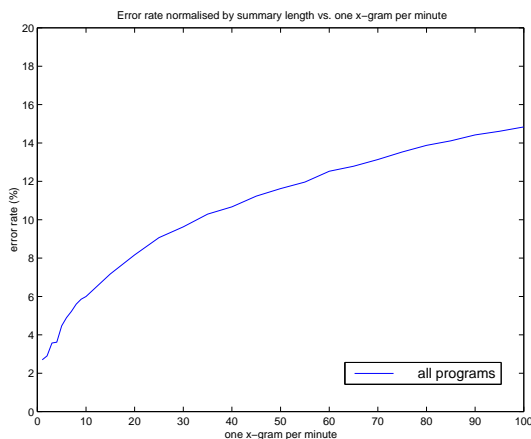


Figure 1: Average error rates for n-grams ranging from lengths 1 to 100 for 95 decoded BN audio files.

#### 4.6. Survey

Seven summaries of each of five broadcast news programs were presented to subjects for evaluation. The summaries were generated as follows:

- key word list (only)
- key word list + 1 10-gram per minute (low inverse frequency weight)
- key word list + 1 10-gram per minute (high inverse frequency weight)
- key word list + 1 30-gram per minute (low inverse frequency weight)
- key word list + 1 30-gram per minute (high inverse frequency weight)
- key word list + 1 utterance per minute (low inverse frequency weight)
- key word list + 1 utterance per minute (high inverse frequency weight)

The summaries were only labelled with the story number; no indication was given of how the summaries were generated. The order of the summaries for each story was chosen randomly. Subjects were also given the full recogniser output for purposes of comparison.

The universally preferred summary was one utterance per minute with a low frequency weight (= higher accuracy than high frequency rate), with one 30-gram per minute, low frequency rate, a close second. Subjects commented that a 10-gram per minute was not quite enough information to get the idea of a full minute of audio, but that utterances (averaging just over 30 words) and the 30-grams could give the gist of a minute of audio. (Several respondents also noted that they could scan a 30-gram or utterance in about the same amount of time as a 10-gram.)

#### 4.7. Information Retention

The information retrieval indexing and search software produced by the CUED HTK group for the 1998 TREC-7 conference was used to compare the IR performance of the overall decoded text to that of the summaries.

The 49 IR queries from the TREC-6 Spoken Document Retrieval test were used because they were the most objective method available for IR evaluation of these summaries. These queries are less than ideal for measuring how much of the important information is retained because, although some of the queries target information locally important in the documents, other queries target details that are not necessarily important (or even related) to the broad topics covered in each audio document. However, despite their shortcomings, they do give some measure of information retention and are consequently included as an (albeit imperfect) evaluation method.

The TREC-6 SDR performance was evaluated based on two metrics, Expected Run Length, the mean rank at which the target document was found, when it was

Type of Document	ERL	MRR
full recogniser output	6.61	0.678
30-gram + key words	15.1	0.457
utterance + key words	14.8	0.486

Table 2: Results from the TREC-7 test.

found, over all queries:

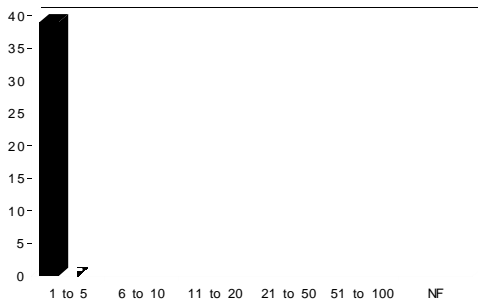
$$E = \frac{1}{N} \sum_{i=1}^N r_i \quad (5)$$

and Mean Reciprocal Rank, the mean of the reciprocal rank at which the target document was found, when it was found, over all queries:

$$E = \frac{1}{N} \sum_{i=1}^N \frac{1}{r_i} \quad (6)$$

The average performance degrades from the full-text to the summaries when queries that hinge on a single words (used once in the audio) being retained in the summary. Even so, each utterance and 30-gram summary retrieved 2/3 the number of stories in the top five that the full recogniser output did; that is, each summary format had 26 stories at ranks 1-5 compared to 39 for the full-text.

These results, shown in full in figure 2 are suggestive that in the majority of cases at least, key information is being retained in the summaries; in fact, in some instances precision increases from the full-text to the summaries.



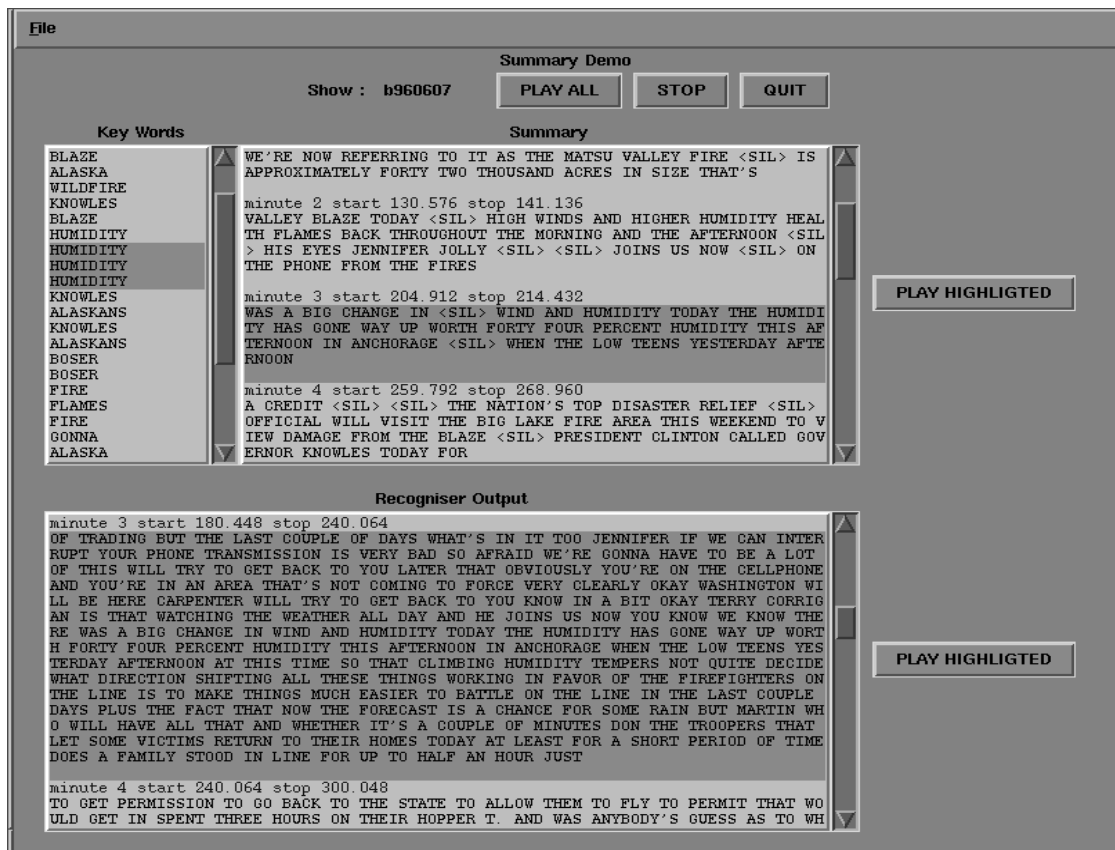


Figure 3: Screen shot of the Tcl/Tk interface. Key words appear in the upper left-hand window, summaries in the upper right-hand window, and full recogniser output in the bottom window.

might be used as both audio and visual summaries of audio. What it is difficult to tell from a printed paper is how much easier it is to browse summaries visually when relevant sections of text can be highlighted and scrolled appropriately, and when summaries, key word lists, full-text output, and audio are dynamically time-aligned as they are highlighted.

The results from this study suggest that statistical techniques that account for both accuracy and frequency of words can be used to produce viable summaries of decoded audio. There is still much work to be done to find the ideal format for summaries and ways to access them, but this work has made inroads into the rather new field of decoded audio summarisation and also points in very promising directions.

## 7. REFERENCES

- [1] Blum, T., D. Keislaer, J. Wheaton, E. Wold. "Audio Databases with Content-Based Retrieval." *Intelligent Multimedia Information Retrieval*, M. Maybury, ed. London: AAAI/MIT, 1997.
- [2] Brown, M.G. J.T. Foote, G.J. Jones, K. Sparck Jones, and S. Young. "Video mail retrieval using voice: An overview of the Cambridge/Olivetti retrieval system." *Proceedings of the ACM Multimedia '94 Conference Workshop on Multimedia Database Management Systems*. San Francisco: 1994.
- [3] Johnson, S. "Reducing Word Error Rates of Found Speech - X Program for Evaluating Recogniser Transcriptions." Technical Report CUED/F-INFENG/TR 330, Cambridge Univ. 1998
- [4] Jones, G., J. Foote, K. Sparck Jones, and S. Young. "The Video Mail Retrieval Project: Experiences in Retrieving Spoken Documents." *Intelligent Multimedia Information Retrieval*, M. Maybury, ed. London: AAAI/MIT, 1997.
- [5] Marsh, E., H. Hamburger, and R. Grishman. "A Production Rule System for Message Summarization." *Readings in Information Retrieval*, K. Sparck Jones, ed. San Francisco: Morgan Kaufmann, 1997.
- [6] Rivlin, Z, M. Cohen, V. Abrash, and T. Chung. "A Phone-Dependent Confidence Measure for Utterance Rejection." *Proceedings of the IEEE*, 1996.
- [7] Robertson, S. and K. Sparck Jones. "Simple, Proven Approaches to Text Retrieval." Technical Note, Cambridge University Computer Laboratory, 1997.
- [8] Robinson, A.J., M. Hochberg, and S. Renals. "The Use of Recurrent Networks in Continuous Speech Recognition." *Advanced topics in Automatic Speaker Recognition*, C. Lee and F.K. Soom, eds. Kluwer 1996.
- [9] Williams, G. and Steve Renals. "Confidence Measures for Hybrid HMM/ANN Speech Recognition." *Proc. Eurospeech 1997*.
- [10] Williams, G. "A Study of the Use and Evaluation of Confidence Measures in Automatic Speech Recognition." Technical report, University of Sheffield University, Dept. of Computer Science, 1998.
- [11] Young, S.R. and W. Ward. "Recognition confidence measures for spontaneous spoken dialogue." *Proc. Eurospeech 1993*.