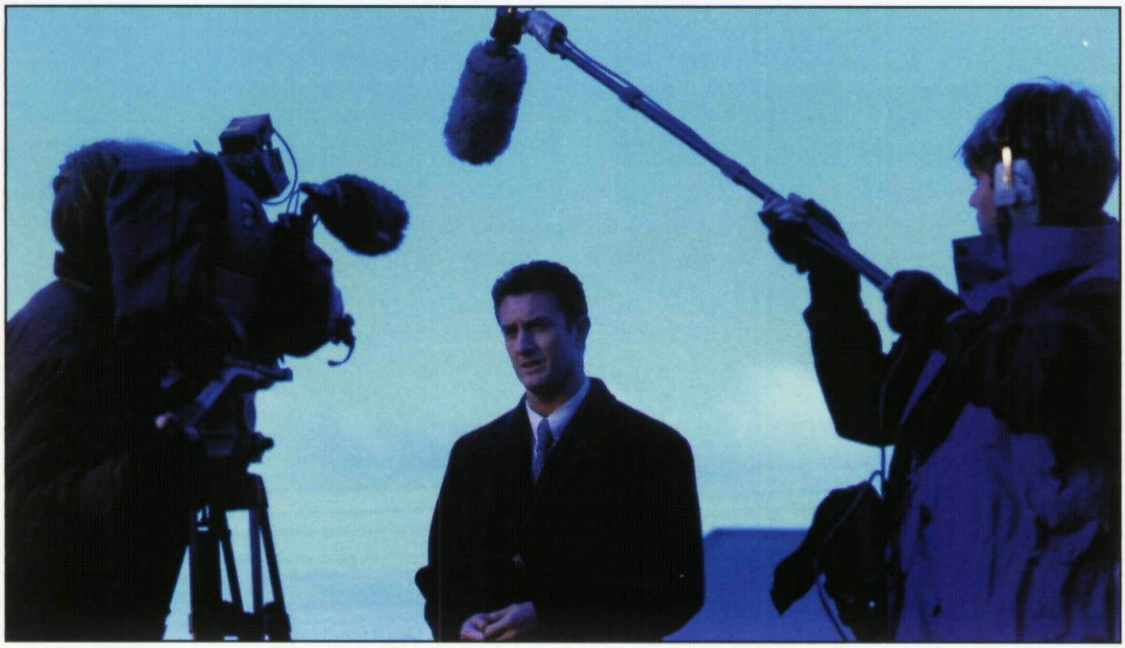
Transcribing speech and making it ready for browsing.

Integrated Technologies FOR Indexing Spoken Language

Speech is not valued today as an archival information source because it is impossible to efficiently locate information in large audio archives. By itself, speech does not permit content-based searches for information like those commonly employed for text documents over the Internet. But now, after more than a decade of steady advances in speech recognition, speaker identification, and language understanding, it is possible to begin building usable automatic content-based indexing tools for spoken language by integrating these emerging technologies. Once these tools become



TONY STONE IMAGES

powerful enough, we believe the original speech recordings will be valued and preserved for their informative nuance and full context.

Rough'n'Ready is the name of a prototype system under development at Bolt, Beranek and Newman (BBN) that aims to provide a practical means of access to information contained in spoken language from audio and video sources. It creates a *Rough* summarization of speech that is *Ready* for browsing. The summarization is a structural representation of the content in spoken language that is very powerful and flexible as an index for content-based information management. This summary, which is automatically produced by the system, includes extracted features such as the names of people, places, and organizations mentioned in the transcript as well as the identities and locations of the speakers in the recording. The system also breaks the continuous stream of words into passages that are thematically coherent. Each of these passages is automatically summarized with a short list of appropriate topic labels drawn from thousands of possibilities. Taken together, these capabilities effectively impose a document model upon spoken language, which permits it to be searched for content with the same ease as textual documents.

Each of the content-based features extracted by the system can be used as query terms to form selective searches over audio archives. For instance, the term *Clinton* can be specified as a person, a location, a speaker, a topic, or as an unqualified word as is common in Internet searches. In addition, the document structure created by the system can be exploited for information retrieval by using an entire passage as a relevance-feedback query. This is possible because the passages or stories that are bounded by Rough'n'Ready are topically coherent. Using a whole story as query-by-example frees the user from constructing complicated Boolean queries to find all similar stories in a large audio archive.

Francis Kubala, Sean Colbath, Daben Liu, Amit Srivastava, and John Makhoul

These powerful capabilities are possible today by leveraging state-of-the-art, but still imperfect, speech and language technologies in an integrated audio indexing system. Currently, the technologies integrated into Rough'n'Ready include large vocabulary speech recognition, speaker segmentation, speaker clustering, speaker identification, name spotting, topic classification, story segmentation, and information retrieval. The integration of such diverse technologies allows Rough'n'Ready to leverage its strengths in combination and produce a high-level structural summarization of the content of spoken language. All of the technologies used in the system employ statistical models whose parameters are estimated automatically from sim-

ply labeled training data, such as word transcriptions with marked names, and complete stories labeled with topics. Non-experts can produce these materials at predictable costs. The statistical models derived from this data by automatic learning algorithms are inherently robust in dealing with the extreme variability found in spoken language. They are also independent of the language used or the domain covered. In addition, these models can be scaled up to handle extremely large data sets with reasonable computational costs.

Audio Browsing with Rough'n'Ready

Elements of the structural summarization produced by Rough'n'Ready are shown in Figure 1. This is a screen shot of the audio browser positioned at the beginning of an episode of a television news program (ABC's "World News Tonight" from January 31, 1998). The episode is decomposed by the browser into three columns representing the speakers, the words spoken, and the topics under discussion.

In the leftmost column, the sequence of speakers is shown as they appear in the audio track of the episode. The boundaries between the speakers, which are automatically determined by the system, provide cues for the paragraph-like breaks in the transcription to the right. The speaker segments have been identified by gender and clustered over the episode to group together segments from the same speaker under the same label. One speaker, Elizabeth Vargas, has been identified by name using a speaker-specific acoustic model for identification.

The automatic transcription is shown in the middle column of the browser in Figure 1. The colored words in the transcription locate the names of people, places, and organizations that have been found by the system. Even though the transcript contains speech recognition errors, the augmented version shown here is easy to read and the gist of the story is apparent with a minimum of effort. It is important to recall that the raw output of the recognizer does not have these properties—it has no punctuation, capitalization, paragraphs, or highlighted content words. These important features of the transcript are derived by the system from additional integrated speech and language technologies—speaker segmentation, clustering, and identification, and name spotting.

In the rightmost column of the browser, a set of topic labels is shown that have been automatically selected by the system to describe the main themes of the first story in the news broadcast. These topic labels are drawn from a set of over 5,500 possible topics known to the system. They constitute a very high-level summary of the content of the underlying spoken language. Although it is not apparent from Figure 1, these topic labels apply to a specific span of words in the automatic transcription. There is a different set of topic labels specified for each story that is automatically detected in the episode. Boundaries between the stories are also located automatically by the system.

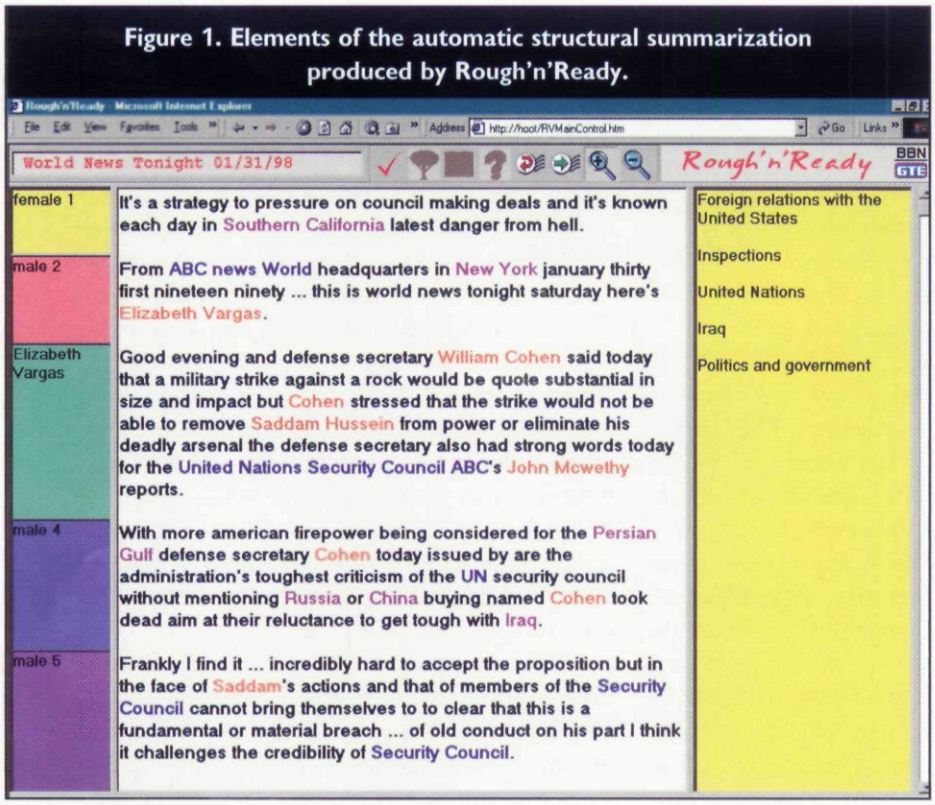


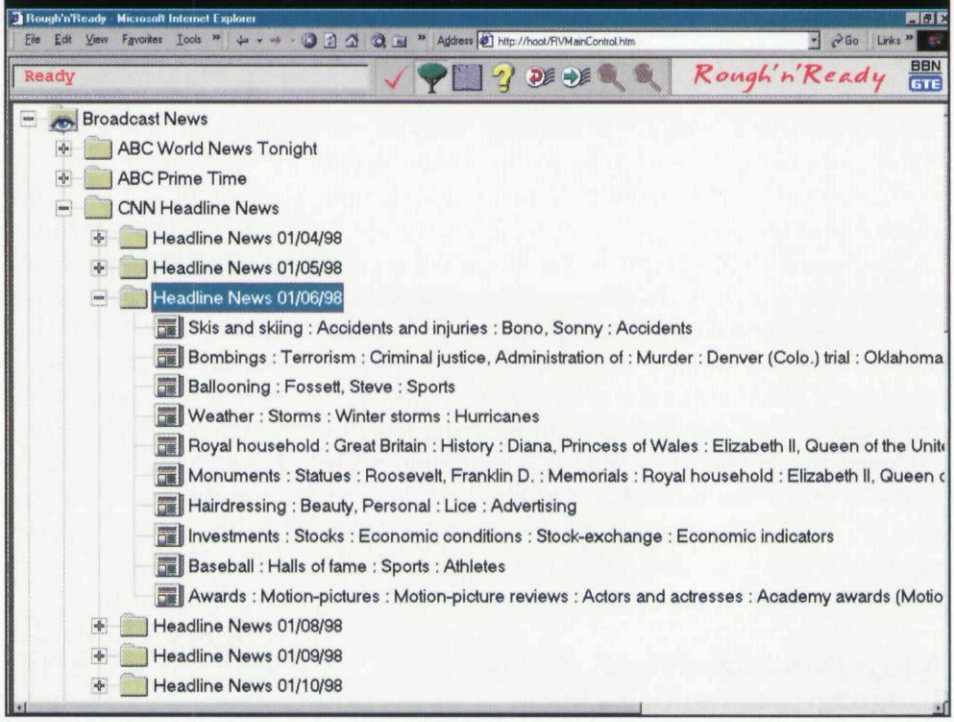
Figure 1. Elements of the automatic structural summarization produced by Rough'n'Ready.

Automatically breaking a continuous audio stream of spoken words into a sequence of bounded and labeled stories is a novel and powerful capability that enables Rough'n'Ready to effectively transform a large archive of audio recordings into a collection of document-like units. The effect of this fundamental capability is shown in Figure 2, which is another screen shot of the audio browser. In this view, an audio archive consisting of 150 hours of broadcast news is organized as a collection of episodes from various content producers. One particular

The summarization is a structural representation of the content in spoken language that is very powerful and flexible as an index for content-based information management.

episode ("CNN Headline News" from January 6, 1998) is expanded to show the sequence of stories detected by the system for this particular episode. Each story is represented by a short list of topic labels that were selected by the system to describe the themes of the story. The net effect of this representation is that a human can quickly understand the contents of a news broadcast from a small set of highly descriptive labels. These labels are very useful as query search terms because they generalize over the content of the story—that is, the words in the topic labels do not need to occur in the text of the story for retrieval to succeed.

Figure 2. A high-level view of an audio archive showing a news episode as a sequence of thematic stories.



The first story in the expanded episode in Figure 2 is about the fatal skiing accident suffered by Sonny Bono. The three important themes for this story—skiing, accidents, and Sonny Bono—have all been automatically identified by the system. Just as importantly, the system rejected all of the other 5,500 topic labels for this story leaving only the concise list of four topic labels shown here to describe the story. Note that the system had never observed these topics together before in its training set, for Mr. Bono died only once. Nonetheless, it was able to select this very informative and parsimonious list of topics from a very large set of possibilities at the same time that it was segmenting the continuous word stream into a sequence of stories. To our knowledge, this powerful summarization capability has been demonstrated for the first time in the Rough'n'Ready audio indexing system.

The entire audio archive of broadcast news is automatically summarized in the same fashion as the expanded episode shown in Figure 2. This means the archive can be treated as a collection of textual documents that can be navigated and searched with the same ease that we associate with Internet search and retrieval operations. Every word of the transcript and all of the structural features extracted by the system are associated with a time offset within the episode, which allows the original audio or video segment to be retrieved from the archive on demand. The actual segment to be retrieved can be easily scoped by the user as a story, as one or more speaker segments, or as an arbitrary span of consecutive words in the transcription. This gives the user precise control over the segment to be retrieved.

In Figure 3, a summary panel is shown containing a list of all the topics found by Rough'n'Ready for the 150-hour broadcast news audio archive. The list is sorted alphabetically in this screen shot and contains the number of times each topic occurs in the archive.

Speech is always a rich source of information, and frequently—in telephone conversations, voice mail, and radio broadcasts, for example—the only source of information.

It is possible to summarize any extracted names or identified speakers in the same manner. Any term in the summary panel can be inserted into the query panel by selecting with the mouse. In Figure 3, the topic *Airplane accidents* is selected in the summary list and also appears in the query panel near the top of the screen. This query term is in green to signify that it is to be used as a topic term in the query. That is, it will match terms in the topic list for each story in the archive regardless of whether the words actually occur in the story. This provides a powerful generalization capability to the user.

The results from a search for all stories in the archive with the topic, *Airplane accidents*, are shown in Figure 4. The folder at the top contains 68 stories labeled with that topic. A new query can be scoped against this collection of stories to narrow it further or the user can view any one of them in more detail in the transcription view of Figure 1. Using topic labels as search terms is a fast and powerful way to discover the contents of a large audio archive. Rough'n'Ready allows the user to use any of its automatically extracted features as search terms.

Integrated Speech and Language Technologies in Rough'n'Ready

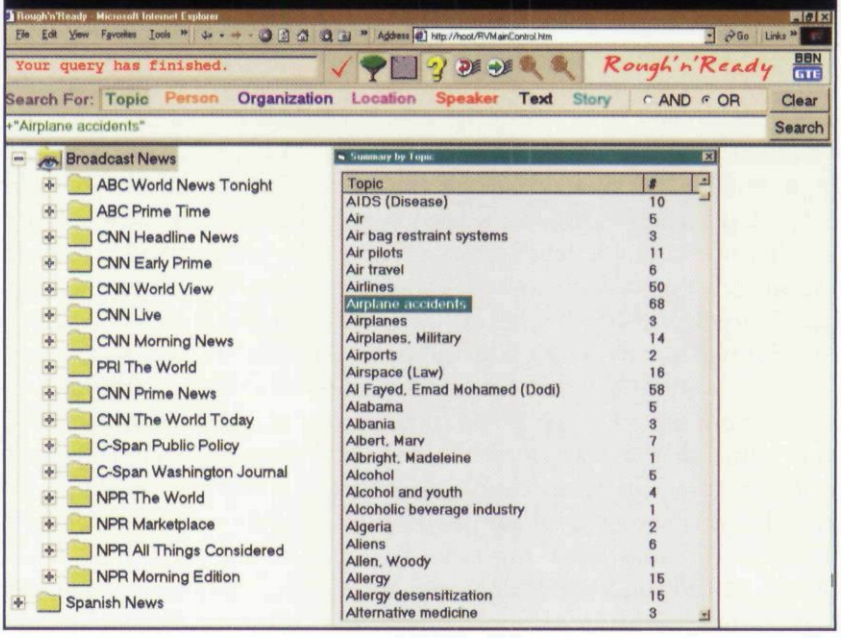
The architecture of the Rough'n'Ready system is illustrated in Figure 5. The overall system is composed of three subsystems—the Indexer, Server, and Browser. Features of the Browser have been described in the previous section. In this section, we will briefly describe each of the speech and language technologies that have been integrated into the Rough'n'Ready Indexer. We will discuss the Server only briefly in this article to describe the Information Retrieval subsystem.

The Indexer subsystem is shown on the left of the diagram in Figure 5 as a cascade of technologies that takes a single waveform as input and produces as output a compact structural summarization encoded as an XML file that is fed to the Server. The duration of the input waveform can be from minutes to hours long. The entire indexing process runs in four times real time on a 450MHz Pentium II processor. We will now briefly describe each of the speech and language technologies making up the Rough'n'Ready Indexer.

Speaker segmentation. The goal of speaker segmentation is to locate all the boundaries between speakers in the audio signal. This is a difficult problem in broadcast news because of the presence of background music and noise. Accurate detection of speaker boundaries provides the speech recognizer with input segments that are each from a single speaker, which enables speaker normalization and adaptation techniques to be used effectively on one speaker at a time. This is one example of the leverage gained through integration—speech recognition degrades without accurate knowledge of the location of speaker changes in the audio stream. Furthermore, speaker change boundaries break the continuous stream of words from the recognizer into paragraph-like units that are often homogeneous in topic.

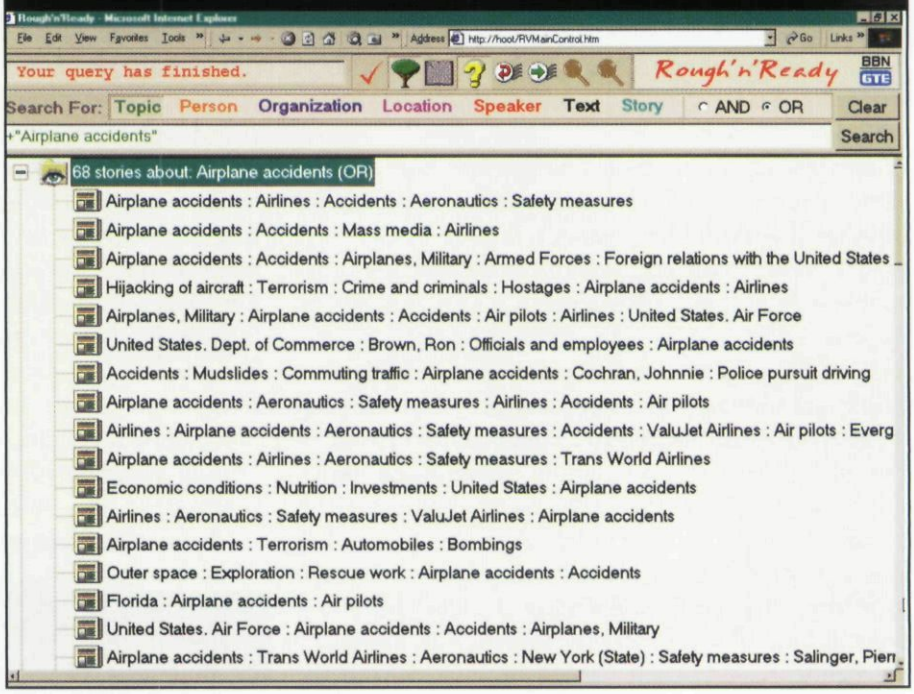
We have developed a novel two-stage approach to

Figure 3. A summary list of the topics found in an audio archive of broadcast news.



speaker change detection that is described in [1]. The first stage is a gender-independent phone-class recognition pass whose primary purpose is to label the input frames as speech or non-speech. We collapse the phoneme inventory to only three broad classes but we include five different models for typical non-speech phenomena (music, noise, laughter, breath, lip-smack) resulting in a small fast decoder that produces reliable speech/non-speech labels at each frame of the signal analysis. Locating non-speech frames reliably is important since 80% of the speaker boundaries in broadcast news occur within non-speech intervals. The system can confidently hypothesize a speaker boundary within a non-speech gap without risking splitting a word in two and thereby degrading the speech recognition that follows.

Figure 4. Returned results from a query for stories with the topic "Airplane accidents."



The second stage performs the actual speaker segmentation by hypothesizing a speaker change boundary at every phone boundary that was located in the first stage. The decision takes the form of a likelihood ratio test where the null hypothesis is that the adjacent segments are produced from the same underlying distribution. The distance metric used here is similar to the Bayesian Information Criterion introduced in [2]. The phone level time resolution in our approach permits the algorithm to run very quickly while maintaining the same accuracy as hypothesizing a boundary at every frame. In addition, accurate identification of the non-speech frames allows us to omit them from the distance calculation, which improves discrimination between the speech samples from two different speakers.

This approach has performed better than any competing approach we have tried in the past. We define a boundary error as any hypothesized boundary that is more than 100ms away from a true boundary, which is marked by hand in the test sample. 100ms is the duration of a single phoneme and, as such, is a very conservative tolerance for measuring speaker change detection error. Using this stringent measure of accuracy, our speaker segmentation algorithm detects 70% of the true speaker boundaries in broadcast news.

Speech recognition. The transcription in Rough'n'Ready is created by the BYBLOS large vocabulary speech recognition system that has been intensively developed over many years at BBN. This is a continuous-density hidden Markov model (HMM) system that has been competitively tested in annual formal evaluations for the past 12 years [4]. In Rough'n'Ready, we are currently running BYBLOS at three times real time using a 60,000-word dictionary. The acoustic models used are speaker- and gender-independent and do not include unsupervised adaptation at this time. The recognition accuracy for this configuration is given in Table 1 on the standard Hub4 broadcast news test sets. These results are only about 20% worse than those achieved in the 1998 Hub4 evaluation by systems running at the rate of several hundred times real time.

Speaker clustering. The goal of speaker clustering is to identify all segments from the same speaker in an episode and assign them a unique label. It is a form of unsupervised speaker identification that is useful for News on Demand systems since it allows skimming audio by speaker identity. One recent approach to the problem is described in [3]. The problem is difficult in broadcast news because of the extreme variability of the signal and also because the true number of speakers can vary so widely (anywhere from 1 to about 100). We have found an acceptable solution to this problem using an agglomerative clustering approach that is described in [5]. The total number of clusters

produced is controlled by a penalty that is a function of the number of clusters hypothesized. A positive bias is applied for segments that are adjacent to each other in time. This algorithm has proved effective over a very wide range of news broadcasts. It performs well regardless of the true numbers of speakers in the episode.

Speaker identification. The first step is to identify every speaker cluster by gender. This is a form of supervised speaker identification. A Gaussian mixture model (GMM) for each gender is estimated from a large sample of training data that has been partitioned by gender.

In addition to gender, the system can identify a specific target speaker if given one minute of speech from the speaker. Again, a GMM is estimated from the given data and this is used to identify segments of speech from the target speaker. The general approach is described fully in [6]. Any number of target models can be constructed and used simultaneously in the system to identify the speakers. To make their labeling decisions, the set of target models compete with a speaker-independent *cohort* model that is estimated from the speech of hundreds of speakers. Each of the target speaker models is adapted from the cohort model. This results in robust target speaker models that are normalized to a common base. Since every Gaussian in the speaker-specific models is tied to a matching Gaussian in the cohort model, the system is able to identify beforehand the 5 most important Gaussians in the mixture for the given speaker and cohort. This approximation allows the algorithm to run very fast while sacrificing almost nothing in accuracy.

Name spotting. The objective of name spotting in Rough'n'Ready is to extract important terms from the speech and collect them in a database. Currently, the system locates names of persons, places, and organizations. Most of the previous work in this area has considered only text sources of written language and has concentrated on the design of rule driven machines to locate the names. Extraction from automatic transcriptions of spoken language is more difficult than written text due to the absence of capitalization, punctuation, and sentences as well as the presence of recognition errors. These have a significant degrading effect on the performance of rule-driven systems.

To overcome these problems, we have developed an HMM-based learning name extraction system called IdentiFinder, which is described in [7]. The model employs a statistical bigram language model, part-of-speech tagging, and a set of features based on the word orthography. Taken together, these features allow the model to generalize from the learning examples to extract new names from independent test data. The parameters of the system are automatically estimated from a corpus of labeled data. The costs for bootstrapping such a learning system to a new domain are modest and predictable since non-expert annotators easily learn the labeling task. In contrast, rule design is a high art requiring years of experience.

IdentiFinder's name spotting accuracy on written text sources is competitive with state-of-the-art rule based systems, which are able to detect more than 90% of the names in newswire texts. Since IdentiFinder is a learned statistical model, it degrades very little in the absence of case, punctuation, and sentence boundaries—conditions that are typical of speech recognition output. In the presence of speech recognition errors, it degrades in a predictable fashion that is a linear function of the word error rate (75% name spotting accuracy at 20% recognition word error).

Topic classification. We have developed a novel approach to topic classification that selects a set of topics appropriate to a given document. It has been shown to work well even when the set of possible topics numbers in the thousands. Both of these attributes are in marked contrast to nearly all of the previous work in topic classification, which assigned a single topic to a document from a set of a few hundred possibilities at most. In addition, our method includes an explicit model of general language that absorbs common words that do not contribute toward a specific topic.

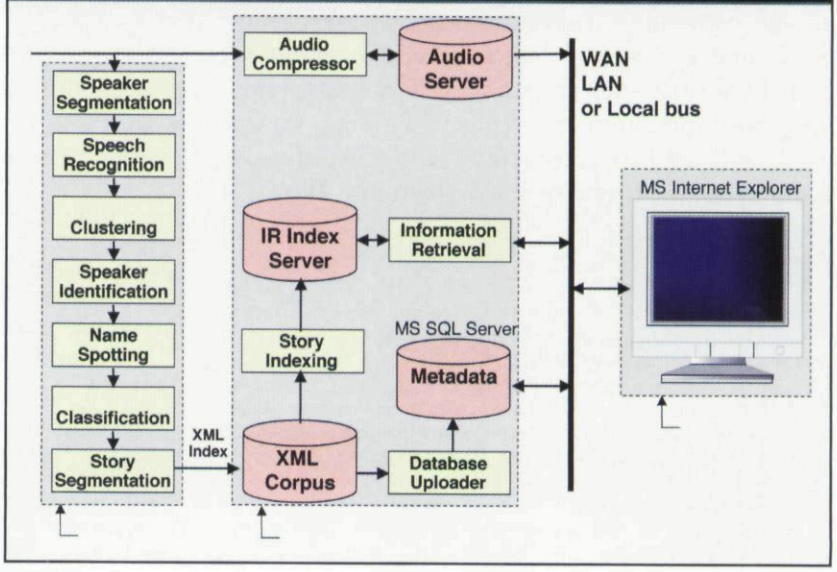
Table 1. BYBLOS speech recognition word error rates on the Hub4 evaluation test sets. The recognizer was run at three times real time.

HUB4 TEST SET	WORD ERROR RATE
1996	29.1
1997	21.5
1998 Set 1	21.4
1998 Set 2	18.8

Our topic classification subsystem, called OnTopic, is a probabilistic HMM whose parameters are estimated from training samples of documents with given topic labels. The model allows each word in the document to contribute different amounts to each of the topics assigned to the document. The algorithm is described in detail in [8]. The output from OnTopic is a rank ordered list of all possible topics and scores for the given document. In Rough'n'Ready, we apply OnTopic to the continuous stream of words in the transcription to overlapping passages of two hundred words span. The step size

between successive passages is four words. Hence, we obtain a ranked list of all 5,500 topics known to the system at four-word intervals throughout the episode. These lists are automatically pruned based on their scores to preserve only the top scoring (i.e., the most relevant) topics for the passage. This sequence of topic vectors is used as the input to the succeeding Story Segmentation stage.

Figure 5. Distinguished architecture of the Rough'n'Ready audio indexing system.



Story segmentation. Story segmentation is the most recent technology to be added to Rough'n'Ready. It is the vital step that turns the continuous stream of spoken words into document-like units with a coherent set of topic labels assigned to each story. The approach, which is described in [9], begins by locating regions of topic stability by observing the relative persistence of the topics surviving in the pruned rank lists produced by OnTopic at four-word intervals. The story boundaries are then located more precisely between these stable regions using the locations of the topic support words as evidence. Topic support words are those words in the document or passage that contribute to the score of a surviving topic after the ranked lists are pruned. We have observed a marked and predictable divergence in the associations of support words across true story boundaries. We exploit this effect to automatically locate the story boundaries occurring between stable topic regions identified in the transcription. We also constrain the boundary decision to prefer a nearby speaker boundary and to avoid splitting names. As shown earlier in Figure 2, this approach is very effective in summarizing an entire news broadcast as a small number of stories, each labeled with a small set of informative topics.

Information retrieval. Information indexing and retrieval take place on the Rough'n'Ready Server. Whenever a new episode is processed by the Rough'n'Ready Indexer, a new retrieval index is generated over the entire archive of indexed stories. The Rough'n'Ready Browser gives the user a powerful query-by-example capability whereby an entire news story is submitted to the Golden Retriever search engine as a query to find all similar stories in a large audio archive. This provides an effective means for a user to find related passages once a single example of interest has been found. This capability would not be possible on audio sources like broadcast news without the topic classification and story segmentation capabilities described previously.

Golden Retriever is a probabilistic HMM-based IR system described in [10]. The model computes the probability that a document is relevant, given a query, and ranks all documents in the collection based on this measure. It uses unsupervised relevance feedback to enlarge the given query and improve performance. This technology was tested for the first time the 1998 TREC7 evaluation and ranked among the leaders in the ad hoc query track.

Future Directions

To date, Rough'n'Ready has focused completely upon spoken language because speech is always a rich source of information, and frequently—in telephone conversations, voice mail, and radio broadcasts for example—it is the only source of information. It is clear, however, that a great deal of information is conveyed by means other than speech in a medium such as broadcast news. Content cues can be extracted directly from the video or onscreen text. In many cases, close-captioned text is available. There is a great deal of research activity today that is directed toward incorporation of all of these multiple media cues into comprehensive media management systems. Several of these experimental systems—such as CMU's Informedia, SRI's MAESTRO, and MITRE's Broadcast News Navigator—are described elsewhere in this special section. Rough'n'Ready's rich content summary of spoken language can form an important component within such multiple media systems or it can be used in a standalone manner.

One important but undeveloped area for information management systems is overall evaluation. Objective evaluations have been conducted for years on the component technologies in Rough'n'Ready, but formal testing of the integrated system has not yet been done. In order to continue effective development of information management systems, we will need well-defined, repeatable, and cost-effective designs for end-to-end system evaluations. C

References

1. Bikel, D., Miller, S., Schwartz, R., Weischedel, R. Nymble: A high-performance learning name finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Association for Computational Linguistics, 1997, 194–201.
2. Chen, S. and Gopalakrishnan, P. Speaker, environment, and channel change detection and clustering via the Bayesian information criterion. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, 1998
3. Hain, T. et al. Segment generation and clustering in the HTK broadcast news transcription system. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop* (Lansdowne, VA, Feb. 1998).
4. Herb, G., and Schmidt, M. Text-independent speaker identification. *IEEE Signal Processing Magazine* (Oct. 1994), 18–32.
5. Imai, T., Schwartz, R., Kubala, F., and Nguyen, L. Improved topic discrimination of broadcast news using a model of multiple simultaneous topics. In *Proceedings of ICASSP97*, (Munich, Germany, Apr. 1997), 727–730.
6. Jin, H., Kubala, F., and Schwartz, R. Automatic speaker clustering. In *Proceedings of the DARPA Speech Recognition Workshop*, February 1997, pp. 108–111.
7. Kubala, F. et al. The 1997 BYBLOS system applied to broadcast news transcription. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, (Lansdowne, VA, Feb. 1998).
8. Liu, D. and Kubala, F. Fast speaker change detection for broadcast news transcription and indexing. In *Proceedings of Eurospeech99* (Budapest, Hungary, Sept. 1999), 1031–1034.
9. Miller, D., Leek, T., and Schwartz, R. BBN at TREC7: Using Hidden Markov Models for information retrieval. In *Proceedings of the Seventh Text REtrieval Conference* (TREC 7), NIST Special Publication 500-242, p. 133.
10. Srivastava, A. Story segmentation. Master's Thesis. Northeastern University, Boston, Mass., 1999.

Francis Kubala (fkubala@bbn.com) is a division scientist at BBN Technologies, GTE Technologies.
Sean Colbath (scolbath@bbn.com) is a scientist at BBN Technologies, GTE Technologies.
Daben Liu (dliu@bbn.com) is a staff scientist at BBN Technologies, GTE Technologies.
Amit Srivastava (asrivast@bbn.com) is a staff scientist at BBN Technologies, GTE Technologies.
John Makhoul (makhoul@bbn.com) is the chief scientist at BBN Technologies, GTE Technologies.

Each of the core speech and language technologies in Rough'n'Ready has been supported by DARPA and other agencies of the U.S. government. They have benefited greatly by formal competitive technology evaluations sponsored by DARPA over many years. The Rough'n'Ready system itself was developed with support from the Intelligent Collaboration and Visualization program at DARPA and monitored by the Air Force Rome Laboratory.