Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004
June 9 - 11, 2004
Espoo, Finland

(John.Hansen@colorado.edu)

SPEECHFIND: Spoken Document Retrieval for a National Gallery of the Spoken Word

John H.L. Hansen[1],Rongqing Huang[1], Praful Mangalath[1], Bowen Zhou[1], Michael Seadle[2], John. R. Deller, Jr[3]

[1] Robust Speech Processing Group, Center for Spoken Language Research;
University of Colorado, Boulder, Colorado 80309-0594, USA
[2]Michigan State University, E308 Main Library, East Lansing, MI 48824, USA
[3]Michigan State University, Dept. Electrical & Computer Engineering, East Lansing, MI 48824, USA

ABSTRACT

In this study, we discuss a number of issues for audio stream phrase recognition for information retrieval for a new National Gallery of the Spoken Word (NGSW). NGSW is the first large-scale repository of its kind, consisting of speeches, news broadcasts, and recordings that are of historical content from the 20th Century. We propose a system diagram and discuss critical tasks associated with effective audio information retrieval that include: advanced audio segmentation, speech recognition model adaptation for acoustic background noise and speaker variability, and natural language processing for text query requests. A number of questions regarding copyright assessment, metadata construction, digital watermarking must also be addressed for a sustainable audio collection of this magnitude. Our experimental online system entitled "SpeechFind" is presented which allows for audio retrieval from a portion of the NGSW corpus. We discuss a number of research challenges to address the overall task of robust phrase searching in unrestricted audio corpora.

1. Overview

The problem of reliable speech recognition for spoken document/information retrieval is a challenging problem when data is recorded across different media, equipment, and time periods. In this paper, we address the problem of audio stream phrase recognition for a new National Gallery of the Spoken Word (NGSW)[1]. This is the first large-scale repository of its kind, consisting of speeches, news broadcasts, and recordings that are of significant historical content. The U.S. National Science Foundation recently established an initiative to provide better transition of library services to digital format. As part of this Phase-II Digital Libraries Initiative, researchers from Michigan State Univ. (MSU) and Univ. of Colorado Boulder have teamed to establish a fully searchable, online WWW database of spoken word collections that span the 20th Century. The database draws primarily from holdings of MSU's Vincent Voice Library that include +60,000 hours of recordings (from T. Edison's first cylinder disk recordings, to famous speeches such as man's first steps on the moon "One Small Step for Man", to American presidents over the past 100 years). In this partnership, MSU digitizes and houses the collection, as well as catalog, organize, and provide meta-tagging information. MSU is also responsible for several engineering challenges such as digital watermarking and effective compression strategies[6,7]. RSPG-CSLR is responsible for developing robust automatic speech recognition for transcript generation, and proto-type audio/metadata/transcript based user search engine, which is called ***SpeechFind***.

Spoken document retrieval focuses on employing text based search strategies from transcripts of audio materials. The transcripts in turn, have reverse index timing information that allows audio segments to be returned to the user to allow for user access. While automatic speech recognition (ASR) technology has advanced significantly, the ability to perform ASR for spoken document retrieval presents some unique challenges. First, while transcription of voice-mail, which only contains one speaker, or two-way telephone conversations which have two speakers, audio streams from NGSW encompass the widest range of audio materials available. The types of audio include: Speeches: (politicians, etc) which are read/prepared text; Interviews: (typically 2 speakers) which include question/answer spontaneous speech; Radio Broadcasts: which includes talk radio, music, call-in speakers, commercials, public radio (NPR); TV Broadcasts: news programs (60 Minutes TV show) with commercials; Meetings/Hearings: public formal inquiries (Watergate hearings, U.S. Supreme Court, etc.); Debates: presidential, formal and informal (Nixon-Kennedy, Clinton-Dole, etc.); Historical Recordings: NASA – walk on the moon, Nixon – "I'm not a crook", M.L. King – "I have a dream", etc. So, the audio content includes a diverse range of audio formats, recording media, and diverse time periods including names, places, topics, choice of vocabulary. Issues include: Do we transcribe commercials? Do we transcribe background acoustic noise/events? Do we identify speakers with the text? Do we identify where the speakers are speaking from (i.e., the environment/location)? How do we deal with errors in ASR (i.e., "dirty transcripts")? Since automatic transcription for such a diverse range of audio materials will lead to significant variability in word-error-rate (WER), SDR employing text based search of such transcripts will be an important research issue to consider.

2. SpeechFind System Overview

Here, we present an overview of the SpeechFind system (see Fig. 1) and briefly describe several key modules. The system includes the following modules: an audio spider and transcoder, spoken documents transcriber, "rich" transcription database, and an on-line public accessible search engine. As shown in the figure, the audio spider and transcoder are responsible for automatically fetching available audio archives from a range of available servers and transcoding the heterogeneous incoming audio files into uniform 16KHz, 16bit linear PCM raw audio data. In addition, for those audio documents with metadata labels, this module also parses the metadata and extracts relevant information into a "rich" transcript database for guiding information retrieval.

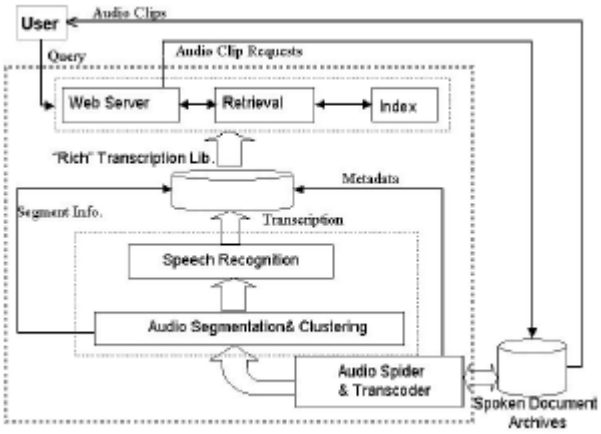

Fig. 1: Overview of SpeechFind system architecture

The Spoken document transcriber includes two components, namely the audio segmenter and transcriber. The audio segmenter partitions audio data into managable small segments by detecting speaker, channel and enviromental change points. The transcriber decodes every speech segment into text. If human transcripts are available for any of the audio documents, the segmenter is still applied to detect speaker, channel and environmental changes in a guided manner, with the decoder being reduced to a forced aligner for each speech segment to tag timing information for spoken words. Fig. 2 shows that for the proposed SpeechFind system, transcript generation is first performed which requires reliable acoustic and language models appropriate for the type of audio stream and time period. After transcript generation, Fig. 3 shows that three associated files are linked together, namely (i) the audio stream in (_.wav) format, (ii) the transcript (_.trs) file with time indexed locations into the audio file, and (iii) extended archive descriptor (_.ead) file that contains Meta-Data. Each audio stream has a reverse index word histogram (with all stop words – "the, a, or, etc." set aside) that is employed with the natural language processing text search engine. These integrated files form the statistical information retrieval (SIR) engine.

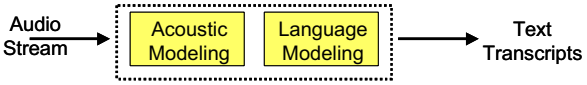

Fig. 2: Automatic Transcript Generation for SDR

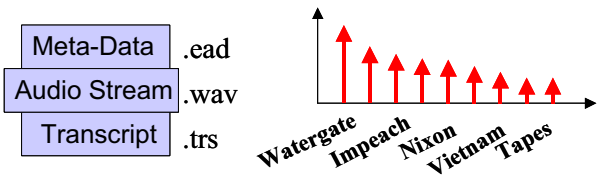

Fig. 3: Statistical Information Retrieval (SDR)

The on-line search engine is responsible for all information retrieval related tasks including a web-based user interface as the front-end, and search and index engines at the back-end. As the audio spider and transcoder, the indexer runs periodically and is activated in an event-driven manner. The web-based search engine responds to a user query by launching back-end retrieval commands, formatting the output with relevant transcribed documents that are ranked by relevance scores and associated with timing information, and provides the user with web based page links to access the corresponding audio clips. It should be noted that the local system does not store the entire collection of audio archives, due to both copyright and disk space issues. Several hundred hours of audio have been digitized by MSU, and a portion is processed and accessible via SpeechFind (see Fig. 4).

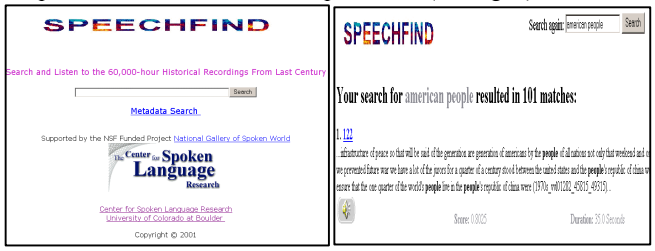

Fig. 4: (i) Sample Web Page & (i) Output
http://SpeechFind.colorado.edu

3. Transcribing Audio Archives

3.1 Spoken Archives Segmentation

Audio archive segmentation obtains manageable audio blocks for subsequent speech decoding, as well as allows for analysis of the location of speaker(s), channel and environmental change points as useful information to help track audio segments of interest.

The goals of effective audio/speaker segmentation[8,9] are different than those for ASR, and therefore features, processing methods and modeling concepts successful for ASR may not necessarily be appropriate for segmentation. Features used for speech recognition attempt to minimize the differences across speakers and acoustic environments (i.e., *Speaker Variance*), and maximize the differences across phoneme space (i.e., *Phoneme Variance*). However, in speaker segmentation for audio streams, we want to maximize speaker traits to produce segments that contain a single acoustic event or speaker, and therefore MFCCs may not be as effective for speaker segmentation. In the present study, we consider several novel features (e.g., PMVDR [10], SZCR, FBLC) and their combination.

3.1.1 FES: A New Evaluation Criterion

The goal of reliable segmentation in audio streams requires that we measure the mismatch between hand/human segmentation and automatic segmentation. We feel frame accuracy may not be the best criterion for audio/speaker segmentation since frequent toggling action between classes leads to short audio segments which are not helpful for automatic transcription if model adaptation is used for ASR. EER (Equal Error Rate) is another popular evaluation criterion. However, the miss rate can be more important than the false alarm rate. Also, the average mismatch between experimental and actual break points is an important norm. Therefore, the proposed *FES* – "fused error score" combines the three evaluation criteria of false alarm rate, miss rate, and average mismatch in a manner similar in principle to WER and accuracy in ASR, as follows:

$$Fused\ Error\ Score = (False\ Alarm\ Rate_{\%} + 2 * Miss\ Rate_{\%}) * Average\ Mismatch_{ms}$$

3.1.2 Segmentation with Three Novel Features

The segmentation scheme employed here is an iterative T^2-Statistic based Bayesian information criterion proposed in an earlier study[16] (called "T2-BIC" here). This approach[16] is an improved version of original BIC[12] which performs

segmentation 100 times faster which higher accuracy for short duration turns of less than 2 seconds.

Having developed a new integrated evaluation criterion, we now turn to improved features for segmentation. We consider three new features here, and compare them to traditional MFCCs.

PMVDR: High order MVDR(Minimum Variance Distortionless Response) models provide better upper envelope representations of the short-term speech spectrum than MFCCs[13]. A perceptual based MVDR feature was proposed in [10] which we consider for segmentation here (i.e., PMVDRs) that do not require an explicit filterbank analysis of the speech signal. We also apply a detailed Bark frequency warping for better results.

SZCR: A High Zero Crossing Rate Ratio has also been proposed for speaker classification. Here, we find the smoothed ZCR is more efficient, which is computed as: compute 5 sets of ZCR evenly spaced across the analysis window with no intermediate overlap; use the mean of the 5 sets as the feature of this frame.

FBLC: Although, it has been suggested that direct warping of the FFT power spectrum without filterbank processing can preserve almost all the information in the short-term speech spectrum[10], we find that filterbank processing is more sensitive than other features in detecting speaker change. As such, the FBLC are the 20 Mel frequency FilterBank Log energy Coefficients.

3.1.3 Feature Evaluation

For our experiments, the evaluation data is drawn from broadcast news Hub4 1996 training data, Hub4 97 evaluation data and NGSW data[1]. Table 1 shows that PMVDR can outperform MFCC on all levels. FBLCs have very small average mismatch implying they are very sensitive to changes between speakers and environments. Because PMVDR does not apply filterbank processing, we combine PMVDR and FBLC together. Also, the SZCR encodes information directly from the waveform that we combine as well. We select the 24 features from PMVDR, all 20 features from FBLC, and 1 SZCR(i.e., a 45-dimensional set). We normalize the features to zero mean and unit variance for improved discrimination ability.

Feature	FA	MIS	MMatch	FES
MFCC	29.6%	25.0%	298.47	237.58
FBLC	29.8% (-0.7%)	25.3% (-1.2%)	266.80 (10.6%)	214.51 (9.7%)
PMVDR	25.9% (12.5%)	24.9% (0.4%)	284.29 (4.8%)	215.21 (9.4%)
Combine 45-D	23.8% (19.6%)	24.3% (2.8%)	265.06 (11.2%)	191.99 (19.2%)

Table 1: SDR Segmentation Feature Performance. Note: '(x.x%)' represents the relative improvement in FA: false alarm rate, MIS: miss detection rate, MMatch: average mismatch (msec), and FES: fused error score.

3.1.4 NGSW & DARPA Hub4 Segmentation Evaluation

The DARPA Hub4 1997 Evaluation Data was used for performance assessment. The set contains 3 hours of Broadcast News data, with 584 break points, including 178 short segments(<5s). CompSeg uses PMVDR, SZCR, FBLC features, applies T^2-Mean measure for segments less than 5 secs, and a novel False Alarm Compensation post-processing routine[15]. The improvement using these advances is shown in Table 2, where performance improves significantly on the Hub4 data. The baseline system uses MFCCs and traditional BIC[12] only.

We also evaluate CompSeg[15] algorithm with a portion of the NGSW corpus[1], using audio material from the 1960s. From Table 2, we see that CompSeg can detect not only the speaker changes, but also the music and long silence(>2s) segments.

(i.)

Algorithm	FA	MIS	MMatch	FES
Baseline	26.7%	26.9%	293.02	235.82
CompSeg	21.1% **(21.0%)**	20.6% **(23.4%)**	262.99 **(10.2%)**	163.84 **(30.5%)**

(ii.)

Speaker Change	Speaker MMatch	Music & Sil Change	Music & Sil MMatch	False Alarm
100%	129ms	100%	118ms	5.6%

Table 2: SDR Segmentation performance using a novel improved T^2-Mean+BIC segmentation scheme with improved features, audio clustering and false alarm compensation (the CompSeg scheme) with (i) DARPA Hub4-97 Broadcast News data, and (ii) sample NGSW audio materials.

4. Spoken Archives Transcription

For SpeechFind, all speech segments are decoded with a large vocabulary recognizer. We are currently using CMU Sphinx3 for this task, but are also using the CSLR Sonic recognizer[17]. The acoustic models contain 5270 GMMs, each of which has 32 mixture Gaussians. Acoustic models are built using a subset of the 200 hours of Broadcast News released by the LDC during 1997 and 1998. The language model is composed of 64K unigrams, 4.7M bigrams, and 15M trigrams. The average decoding speed is about 6.2x real time on a P4-1.7GHz Linux machine. In establishing the baseline experiments, no model adaptation schemes were applied at this stage, and the first pass decoding output is used as the automatic transcriptions, though a second pass re-scoring using a more complex language model might produce better results.

To evaluate recognition performance, 3.8 hours of sample audio data from the past 6 decades in NGSW is used as the test data. Table 3 provides a summary of the audio statistics along with WER averaged for each decade. Here we note that average WER does not increase as we move back in time, though the Out-Of-Vocabulary (OOV) rate does. Instead, the first 3 decades achieves better recognition accuracy, and the lowest WER is observed for corpora from the 1970's. This can be attributed to the lower average SNR for the recordings used from the 1980s and 1990s. For example, three long audio recordings of 1990's that contain 2681 words have an average SNR near 12dB, which produce WERs above 75%, while other recordings with a higher average SNR of 21dB achieve WERs less than 25%. The average SNR of recordings from the 2000s is relatively high, while the audio files are from news conferences regarding the hand counting of votes for the U.S. President in Florida. As a result, this portion becomes transcribed primarily as noise by the recognizer, and as much as 35% of the overall WER is from deletions. It is clear that all possible methods for achieving robust speech recognition will need to be brought to bear to successfully address this problem.

Decade	# of Doc	Audio Length (Min)	# Words	OOV(%)	Avg. SNR (dB)	Avg. WER (%)
1950	4	52	6241	1.42	26.63	38.6
1960	2	17	2142	1.52	21.34	36.7
1970	2	35	4434	0.81	20.87	25.6
1980	3	27	3330	0.63	17.97	60.1
1990	4	47	5951	1.28	14.79	48.0
2000	3	50	7530	0.78	26.81	59.1

Table 3: Description & Evaluation performance of a sample portion of the NGSW audio corpus (29,628 words, 3.8hrs)

5. IR Over Automatic Transcripts & IR Advances

The current SpeechFind retrieval engine is a modified version of MG[16]. Here, the *tfidf* weighting scheme is replaced with Okapi weighting, and several query and document expansion technologies are incorporated. In addition, the SpeechFind web interface provides the user access to the detected speech segments and automatic transcripts, and allows the user to preview and listen to any other portions, or the entire audio file that contains the original detected segments.

Automatic transcriptions essentially decode acoustic recordings using the most probable in-vocabulary word sequences, while text documents and queries written by humans tend to use a simplified notation. For example, "1960" could be widely used in human-written documents to indicate the year, but is not included in either the dictionary or language models in most state-of-the-art speech recognizers. To address this, spoken transcripts and queries are normalized in SpeechFind system to bridge this gap.

The SpeechFind system also includes *Query Expansion* (QE) to address the problem of missing query terms directly using Blind Relevance Feedback (BRF), and *Document Expansion* (DE) which allows other parallel documents related to those in hand to be identified by bringing in "signal" words from related parallel documents using Parallel Blind Relevance Feedback (PBRF). Combining both DE and QE improves overall precision from 42.17% to 50.59%.

Finally, the basic SpeechFind system allows the user to query the audio archive with any word/phrase of interest. Clearly, it is useful to associate audio streams into common semantic associations (i.e., a semantic tree structure with audio stream associations like: inventions, war, civil rights, environment, etc.). Our recent work has focused on Latent Semantic Analysis (LSA)[18] for extracting and representing the contextual-usage meaning of words. In our case, we are using NGSW transcripts to organize audio streams based on specialized domains such as time periods and topic content from the transcripts. One example is:

[Speaker=Herbert Hoover][Topic=Discourse-on-war-strategies][Recording=TV-Broadcast+high-squealing-in-background][Log=(1950s_vvl00218_6000_9500)]
[Survey global Military Situation] "..first survey of a global military situation..."

Here, content within "[...]" will be displayed with a similarity score to allow the user to modify their search for semantically similar audio content.

6. Summary and Conclusion

In this paper, we have presented the SpeechFind system, an experimental on-line spoken document retrieval system for a historical archive with 60,000 hours of audio recordings from the last century. We introduced the system architecture, with focus on audio data transcription and information retrieval components. We considered a new segmentation performance criterion called the Fused Error Score (FES) and evaluated three novel features as alternatives to traditional MFCC based ASR features. We saw that a combined feature set improves segmentation performance by 19.2% over traditional MFCC based BIC. We also evaluated these advances using a recently developed CompSeg segmentation method using DARPA Hub4 and NGSW audio corpora. Next, we considered transcript generation and IR engine using DE and QE for NGSW materials. We combining DE and QE using BRF and PBRF improves average returned document precision from a baseline of 42.17% to 50.59%. Finally, we discussed recent work using LSA analysis for semantic clustering of audio streams for improved user search and retrieval.

Acknowledgements

This work was supported by NSF Cooperative Agreement No. IIS-9817485. Any opinions, findings, and conclusions expressed are those of the authors and do not necessarily reflect the views of NSF.

REFERENCES

[1] http://www.ngsw.org
[2] http://speechfind.colorado.edu/
[3] B. Zhou, J.H.L. Hansen, ICSLP-02, pp. 1969-1972, CO 2002.
[4] J.H.L. Hansen, B. Zhou, M. Akbacak, R. Sarikaya, B. Pellom, "Audio Stream Phrase Recognition for a National Gallery of the Spoken Word: 'One Small Step'," ICSLP-2000, pp. 1089-1092, China, Oct. 2000.
[5] J.H.L. Hansen, J. Deller, M. Seadle, "Engineering Challenges in the Creation of a National Gallery of the Spoken Word: Transcript-Free Search of Audio Archives," IEEE & ACM JCDL-2001: Joint Conf. Digital Libraries, pp. 235-236, Roanoke, VA, June 24-28, 2001.
[6] A. Gurijala, J.R. Deller, Jr., M.S.Seadle, J.H.L.Hansen, "Watermarking through parametric modeling," *ICSLP-02,* pp. 621-624, Sept. 2002.
[7] M.S. Seadle, J.R. Deller, Jr., and A. Gurijala, "Why watermark? The copyright need for an engineering solution," *Proc. 2d ACM/IEEE Joint Conf. Digital Libraries*, Portland, June 2002.
[8] A.Adami, S.Kajarekar, H.Hermansky, "A New Speaker Change Detection Method for Two-Speaker Segmentation," ICASSP-02.
[9] L.Lu, H.Zhang, "Speaker Change Detection and Tracking in Real-Time News Broadcasting Analysis," ACM Multimedia, France, Dec. 2002
[10] U.Yapanel, J.H.L.Hansen, "A New perspective on Feature Extraction for Robust In-Vehicle Speech Recognition," Eurospeech-03, 2003
[11] M.Siegler, U.Jain, B.Raj, R.M.Stern, ``Automatic Segmentation, Classification and Clustering of Broadcast News Audio," DARPA Speech Recog. Workshop, Virginia, USA, pp. 97-99, 1997.
[12] S. Chen, P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via The Bayesian Information Criterion," Proc. Broadcast News Trans. & Under. Workshop, 1998.
[13] S.Dharanipragada, B.Rao, "MVDR-Based Feature Extraction for Robust Speech Recognition," ICASSP-01, Utah, 2001
[14] T.Hain, S.Johnson, A.Tuerk, P.Woodland, S.Young, "Segment Generation and Clustering in the HTK: Broadcast News Transcription System," DARPA Broadcast News Workshop, 1998
[15] R. Huang, J.H.L. Hansen, "Unsupervised Audio Segmentation and Classification for Robust Spoken Document Retrieval," ICASSP-2004.
[16] B. Zhou, J.H.L. Hansen, "Efficient Audio Stream Segmentation Via T2 Statistic Based BIC," IEEE Trans. Speech & Audio Proc., 2004.
[17] I.H. Witten, A. Moffat, T.C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*, Morgan Kaufmann Pub., 1999.
[18] T. Landauer, P. Foltz, D. Laham, Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284, 1998.