SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word

John H. L. Hansen, *Senior Member, IEEE*, Rongqing Huang, *Student Member, IEEE*, Bowen Zhou, *Member, IEEE*, Michael Seadle, J. R. Deller, Jr, *Fellow, IEEE*, Aparna R. Gurijala, Mikko Kurimo, and Pongtep Angkititrakul, *Member, IEEE*

***Abstract*—Advances in formulating spoken document retrieval for a new National Gallery of the Spoken Word (NGSW) are addressed. NGSW is the first large-scale repository of its kind, consisting of speeches, news broadcasts, and recordings from the 20th century. After presenting an overview of the audio stream content of the NGSW, with sample audio files from U.S. Presidents from 1893 to the present, an overall system diagram is proposed with a discussion of critical tasks associated with effective audio information retrieval. These include advanced audio segmentation, speech recognition model adaptation for acoustic background noise and speaker variability, and information retrieval using natural language processing for text query requests that include document and query expansion. For segmentation, a new evaluation criterion entitled *fused error score* (FES) is proposed, followed by application of the CompSeg segmentation scheme on DARPA Hub4 Broadcast News (30.5% relative improvement in FES) and NGSW data. Transcript generation is demonstrated for a six-decade portion of the NGSW corpus. Novel model adaptation using structure maximum likelihood eigenspace mapping shows a relative 21.7% improvement. Issues regarding copyright assessment and metadata construction are also addressed for the purposes of a sustainable audio collection of this magnitude. Advanced parameter-embedded watermarking is proposed with evaluations showing robustness to correlated noise attacks. Our experimental online system entitled "SpeechFind" is presented, which allows for audio retrieval from a portion of the NGSW corpus. Finally, a number of research challenges such as language modeling and lexicon for changing time periods, speaker trait and identification tracking, as well as new directions, are discussed in order to address the overall task of robust phrase searching in unrestricted audio corpora.**

***Index Terms*—Accent classification, broadcast news, document expansion, environmental sniffing, fidelity, fused error score, information retrieval, language modeling, model adaptation, query expansion, robust speech recognition, robustness, security, speech segmentation, spoken document retrieval, watermarking.**

Manuscript received July 1, 2004; revised April 4, 2005. This work was supported by the National Science Foundation (NSF) Cooperative Agreement IIS-9817485. Any opinions, findings, and conclusions expressed are those of the authors and do not necessarily reflect the views of the NSF. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mazin Gilbert.

J. H. L. Hansen and R. Huang were with the Robust Speech Processing Group, Center for Spoken Language Research, University of Colorado Boulder, Boulder, CO 80302 USA. They are now with the Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX 75083 USA (e-mail: John.Hansen@utdallas.edu).

B. Zhou was with Robust Speech Processing Group, Center for Spoken Language Research, University of Colorado Boulder, Boulder, CO 80302 USA. He is now with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA.

M. Seadle is with the Main Library, Michigan State University, East Lansing, MI 48824 USA.

J. R. Deller, Jr. and A. R. Gurijala are with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824 USA.

M. Kurimo is with the Neural Networks Research Center, Helsinki University of Technology, Helsinki, Finland.

P. Angkititrakul was with the Robust Speech Processing Group, Center for Spoken Language Research, University of Colorado Boulder, Boulder, CO 80302 USA. He is now with Eliza Corp, Beverly, MA 01915 USA.

Digital Object Identifier 10.1109/TSA.2005.852088

I. Introduction

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 a number of issues associated with audio stream phrase recognition, copyright/watermarking, and audio content delivery 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 University (MSU) and the University of Colorado at Boulder (CU) have teamed to establish a fully searchable, online WWW database of spoken word collections that span the 20th century [5]. The database draws primarily from holdings of MSU's Vincent Voice Library (VVL) that include more than 60 000 hr of recordings (from Thomas 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 cataloging, organizing, and providing meta-tagging information. A networked client-server configuration has been established between MSU and CU to provide automatic transcript generation for seamless audio content delivery. MSU is also responsible for several engineering challenges such as digital watermarking and effective compression strategies [6], [7]. The Robust Speech Processing Group—Center for Spoken Language Research (RSPG-CSLR) (CU) 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*** [2].

In the field of robust speech recognition, there is a variety challenging problems that persist, such as reliable speech recognition across wireless communications channels, recognition of

speech across changing speaker conditions (emotion and stress [25]–[27], accent [28], [29]), or recognition of speech from unknown or changing acoustic environments. The ability to achieve effective performance in changing speaker conditions for large vocabulary continuous speech recognition (LVCSR) remains a challenge, as demonstrated in recent DARPA evaluations focused on Broadcast News (BN) versus previous results from the Wall Street Journal (WSJ) corpus. Although the problem of audio stream search is relatively new, it is related to a number of previous research problems. Systems developed for streaming video search based on audio [30] or closed-captioning can be effective but often assume either an associated text stream or a clean audio stream. Information retrieval via audio and audio mining have recently produced several commercial approaches [32], [33]; however, these methods generally focus on relatively clean single-speaker recording conditions. Alternative methods have considered ways to time-compress or modify speech in order to give human listeners the ability to more quickly skim through recorded audio data [34]. In general, keyword spotting systems can be used for topic or gisting[1] applications. However, for phrase search, the system must be able to recover from errors in both the user requested text-sequence and rank-ordered detected phrase sites within the stream. Phrase search focuses more on locating a single requested occurrence, whereas keyword/topic spotting systems assume a number of possible searched outcomes. Great strides have also been made in LVCSR for spoken document retrieval for BN in English [31], [35]–[39], German [40], [41], Italian [42], Korean [43], Japanese [44]–[47], Mandarin/Chinese [48]–[52], [57], Finnish [104], Portuguese [53], Arabic [54], and French [55]. The American English BN corpus reflects a wider range of acoustic environments than many large vocabulary corpora (e.g., WSJ, TIMIT). However, the recognition of speech in BN reflects a homogeneous data corpus (i.e., recordings from TV and radio news broadcasts from the 1990s, organized into seven classes from F0: clean, to FX: low fidelity with cross-talk). One natural solution to audio stream search is to perform forced transcription for the entire dataset and simply search the synchronized text stream. Whereas this may be a manageable task for BN (consisting of about 100 hr), the initial offering for NGSW will be 5000 hr (with a potential of 60 000 total hr), and it will not be possible to achieve accurate forced transcription since text data will generally not be available. Other studies have also considered web-based spoken document retrieval (SDR) [3], [4], [56]. Transcript generation of broadcast news can also be conducted in an effort to obtain near real-time closed-captioning [58]. Instead of generating exact transcripts, some studies have considered summarization and topic indexing [59]–[61] or, more specifically, topic detection and tracking [64]; others have considered lattice-based search [101]. Some of these ideas are related to speaker clustering [62], [63], which is needed to improve acoustic model adaptation for BN transcription generation. Language model adaptation [65] and multiple/alternative language modeling [66] have also been considered for SDR. Finally, cross and multilingual-based studies have also been performed for SDR [67], [68]. Advances represented by the cited BN and SDR studies notwithstanding, the NGSW database involves a level of complexity in terms of the range and extent of acoustic distortion, speaker variability, and audio quality that has not been approached in existing research. Probably the only corpus-based study that comes close to NGSW is one focused on Holocaust Survivors [69], consisting of a broad range of speakers in structured two-person interview formats.

In this paper, we introduce SpeechFind: an experimental online spoken document retrieval system for the NGSW. In Section II, we discuss the structure of the audio materials contained in the VVL including time periods, recording conditions, audio format, and acoustic conditions. Section III considers a brief discussion on copyright issues for NGSW. Section IV presents an overview of the SpeechFind system including transcript generation and text-based search. Next, Section V addresses transcript-generation based on i) unsupervised segmentation, ii) model adaptation, iii) LVCSR, and iv) text-based information retrieval. Section VI revisits copyright issues, with a treatment of digital watermarking strategies. Section VII considers additional audio stream tagging and language model concepts for next-generation SDR. Finally, Section VIII summarizes the main contributions and areas for future research.

II. Audio Corpus Structure of NGSW

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 for user access. Whereas automatic speech recognition (ASR) technology has advanced significantly, the ability to perform ASR for SDR presents some unique challenges. These include i) a diverse range of audio recording conditions, ii) the ability to search output text materials with variable levels of recognition (i.e., word-error-rate: WER) performance, and iii) decisions on what material/content should be extracted for transcript knowledge to be used for SDR (e.g., text content, speaker identification or tracking, environmental sniffing [93], [94], etc.). For some audio streams such as voice-mail, which normally contain only one speaker, or two-way telephone conversations with two speakers, transcription using ASR technology is possible since the task primarily focuses on detecting silence/speech activity and then engaging the recognizer appropriate for that speaker. However, audio streams from NGSW encompass one of the widest range of audio materials available today. Fig. 1 presents an overview of the types of audio files and recording structure seen in the audio. The types of audio include the following:

- ***Monologs:*** single speaker talking spontaneously or reading prepared/prompted text in clean conditions;
- ***Two-Way Conversations:*** telephone conversations between two subjects that are spontaneous and could contain periods with both talking;
- ***Speeches:*** audio data where a person (e.g., politician) is speaking to an audience–primarily one talker, but background audience noise could be present, and room echo or noise is possible; typically read/prepared text;

[1]Here, the word "gisting" refers to systems that identify the main topic or "gist" of the audio material.

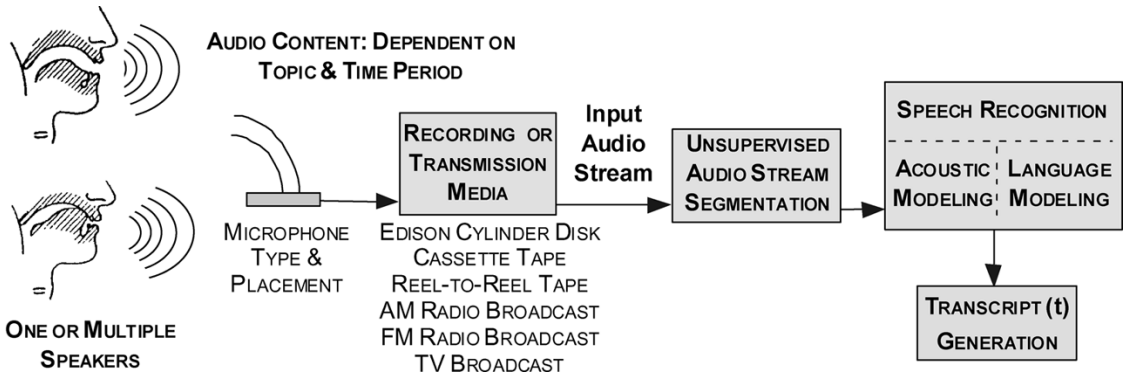


Fig. 1. Structure of i) NGSW audio recordings—speakers, microphone(s), recording media and ii) segmentation and classification, speech recognition, and transcript generation.

- ***Interviews/Debates:*** audio streams where a person is being interviewed by a TV or radio person. Debates could include a moderator and/or various audience participation (e.g., questions, applause, interruptions, etc.); typically two speakers with spontaneous speech in question/answer format;
- ***Radio/TV News Broadcasts:*** includes traditional news anchor with periods of both prompted read speech, talk radio, spontaneous speech, background music, call-in speakers, commercials, other background audio content (e.g., office noise such as typewriter, etc.). Audio content would come from TV or radio studio settings (e.g., public radio such as NPR or 60 Minutes TV show);
- ***Field News Broadcasts:*** audio content coming from news reporters in the field (e.g., emergency or war locations, city streets, etc.); contains a wide range of background noise content of unpredictable origin. Communication channels also impact frequency content of the audio;
- ***Recording Media/Transmission:*** audio properties can be transformed based on the type of recording equipment used (e.g., microphones, Edison cylinder disks, reel-to-reel tape, cassette tape, DAT, CD, etc.) or transmission (e.g., AM, FM , voice compression methods—CELP, MELP, ADPCM, etc.);
- ***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.

Therefore, NGSWaudio content includes a diverse range of audio formats, recording media, and diverse time periods including names, places, topics, and choice of vocabulary. The following issues arise for transcript generation for SDR: Do we transcribe commercials? Do we transcribe background acoustic noise/events? Do we identify speakers with the text? Do we identify from where the speakers are speaking (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 WER, SDR employing text-based search of such transcripts will be an important research issue to consider. For our initial system, we focus on transcript generation of individual speech and disable transcription production for music/commercials.

To illustrate the range of NGSW recording conditions, three example spectrograms are shown in Fig. 2. The recordings are (a) **Thomas Edison**, "my work as an electrician" [talking about contributions of 19th century scientists; original Edison cylinder disk recording, 1908], (b) **Thomas Watson**, "as Bell was about to speak into the new instrument," [talking about the first telephone message from A. G. Bell on March 10, 1876; recorded in 1926], and (c) President **Bill Clinton**, "tonight I stand before you," [State of the Union Address on economic expansion, Jan. 19, 1999]. These examples indicate the wide range of distortions present in the speech corpus. Some of these include severe bandwidth restrictions (e.g., Edison style cylinder disks), poor audio from scratchy, used, or aging recording media, differences in microphone type and placement, reverberation for speeches from public figures, recordings from telephone, radio, or TV broadcasts, background noise including audience and multiple speakers or interviewers, a wide range of speaking styles and accents, etc.

As another example, we show in Fig. 3 a summary of U.S. Presidential speeches, consisting mostly of state-of-the-union or campaign speeches from 1893 to the present. For each presidential speech, we employed the NIST speech-to-noise ratio estimation scheme (STNR) to identify the mean speech and noise decibel values. As we see from this figure, the resulting digitized speech levels are typically near 80 dB, whereas background noise levels can vary significantly (42–78 dB). We obtained the STNR values for each presidential speech, which ranged between 4–37 dB. Clearly, the estimated STNR only has meaning if frequency content is consistent, but as we see in this figure, the estimated frequency bandwidth for early Edison cylinder disks is about 1–2.5 kHz, whereas recordings of today are closer to 7 kHz, with AM/FM radio bandwidths in the 5–10 kHz range (note that while the audio format is 44.1 kHz, 16 bit data, transcript generation uses a sample rate of 16 kHz; therefore, our maximum bandwidth from these recordings for speech content would have been 8 kHz). Recordings for Wilson and Hoover were extremely noisy, with background audience and echo distortion, as well as poor scratchy recording equipment. In addition, vocabulary selection varies significantly over the 110-year

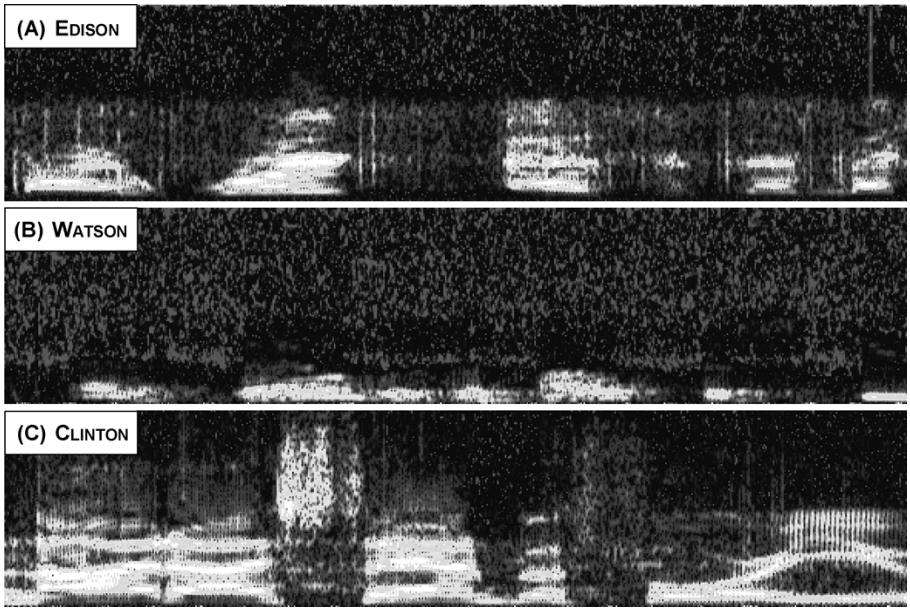


Fig. 2. Example audio stream (8 kHz) spectrograms from NGSW. (a) Thomas Edison, recorded in 1908. (b) Thomas Watson, recorded in 1926. (c) President William J. Clinton, recorded in 1999.

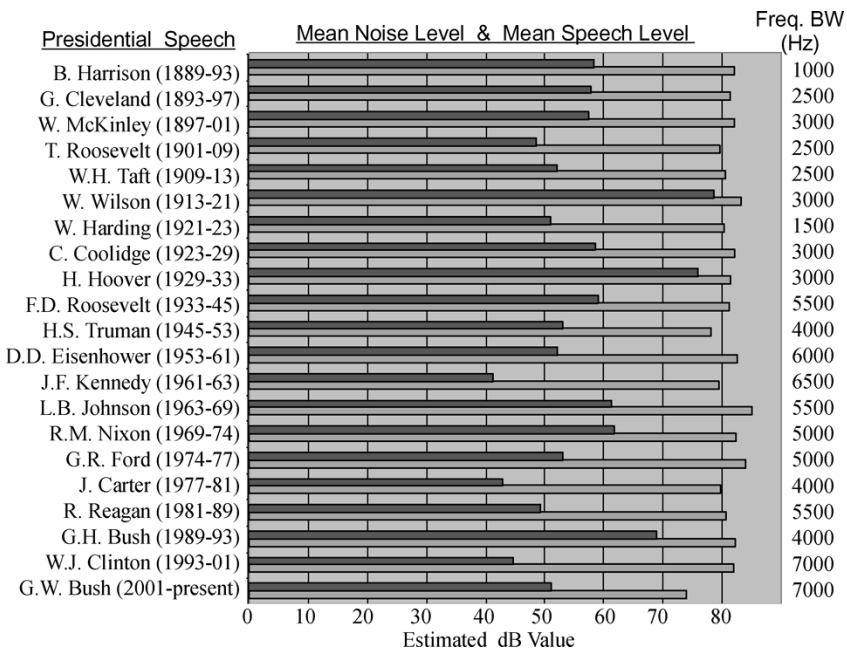


Fig. 3. Summary of presidential speeches (typically state-of-the-union addresses) from 1893 to present. Shown is each president, mean noise signal level (in decibels) (top bars in each pair), mean speech signal level (in decibels) (bottom bars in each pair); approximate frequency bandwidth (BW) of each recording, with an estimated speech-to-noise-ratio (STNR) varied from 4.5—37.25 dB.

period. Clearly, the ability to achieve reliable phrase recognition search for such data is an unparalleled challenge in speech recognition.

III. Copyright Issues in NGSW

When considering distribution of audio material via the WWW, one primary logistics issue concerns copyright ownership. Research on watermarking digital sound is integral to the creation of the NGSW. Most sound recordings have some form of copyright protection under existing law. The US copyright Law (Title 17 of the US Code) explicitly protects sound recordings made since 1978. Some famous speeches have been heavily litigated. An example is Martin Luther King's "I Have a Dream" speech [20].

Many rights holders are willing to make their sound recordings available for educational purposes, but they often require some form of technological protection to prevent legitimate educational copies from being used for unauthorized commercial purposes. The 1998 Digital Millennium Copyright Act (DMCA) introduced penalties for circumventing technological protections. Many in the academic community object to these penalties because they create a contradiction in U.S. law: many

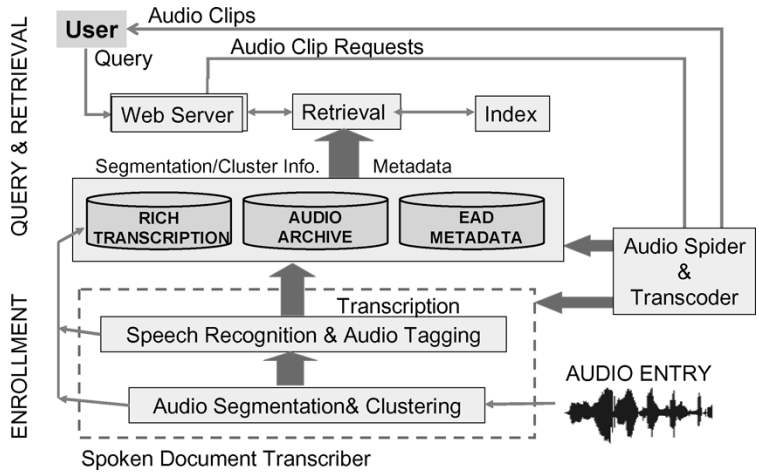


Fig. 4. Overview of SpeechFind system architecture (http://SpeechFind.colorado.edu).

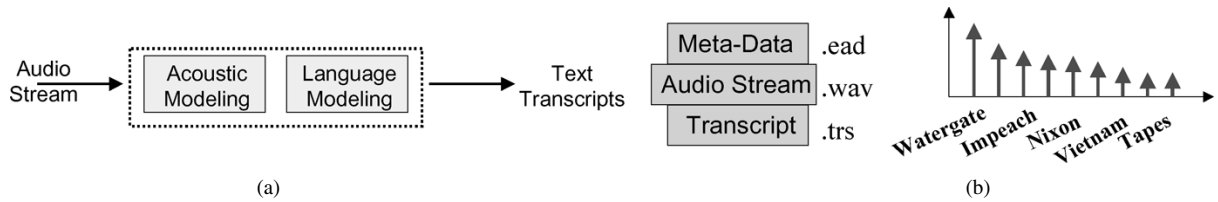


Fig. 5. (a) Automatic transcript generation (SDR). (b) Statistical information retrieval (SIR).

legal "fair uses" of technologically protected works can be exercised only through illegal circumvention. Audio watermarking is a desirable technological protection mechanism because it does not abrogate fair use rights.

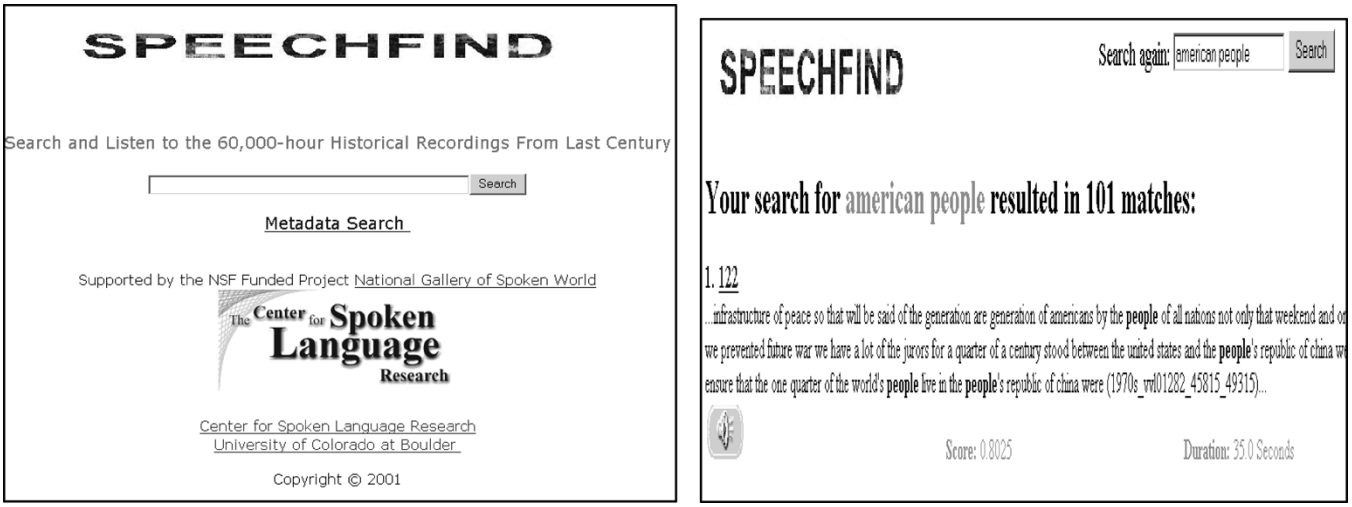
There is evidence that the courts consider audio watermarks to be a legitimate form of copyright protection. The Napster music file-sharing case, for example, mentions both the lack of watermarking on MP3 files and the intention to include it in the future [19]. Watermarking is therefore not a preventative.

Prevention is attractive to those who put significant capital toward the creation of audio works and who fear the loss of investment and future profits. However, prevention is fundamentally inconsistent with most US copyright law, which instead emphasizes mechanisms for redress once an infringement has occurred. Watermarking facilitates redress and represents a copyright protection technology that universities can use without being inconsistent with their interest in and commitment to sharing knowledge. Further treatment of copyright issues and fair use can be found in [7] and [21]–[24]. In Section VI, we consider advances made in digital watermarking for the NGSW project. For the present experimental online SDR system, digital watermarking is employed to both protect ownership as well as help ensure integrity of the audio content.

IV. SpeechFind System Overview

Here, we present an overview of the SpeechFind system (see Fig. 4) and describe several key modules. The system is constructed in two phases: i) enrollment and ii) query and retrieval. In the enrollment phase, large audio sets are submitted for audio segmentation and transcription generation and metadata construction (EAD: extended archive descriptor). Once this phase is completed, the audio material is available through the online audio search engine (i.e., "query and retrieval" phase). The system includes the following modules: an audio spider and transcoder, spoken document transcriber, "rich" transcription database, and an online 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 16-kHz, 16-bit linear PCM raw audio data (note that in general, the transcoding process is done offline prior to being available for user retrieval). 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.

The spoken document transcriber includes two components, namely, the audio segmenter and transcriber. The audio segmenter partitions audio data into manageable small segments by detecting speaker, channel, and environmental 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. 5(a) shows that for the proposed SpeechFind system, transcript generation is first performed, which requires reliable acoustic and language models that are appropriate for the type of audio stream and time period. After transcript generation, Fig. 5(b) shows that three associated files are linked together, namely i) the audio



(a) (b)

Fig. 6. (a) Sample web page. (b) Output web page format. (http://SpeechFind.colorado.edu).

stream in (_.wav) format, ii) the transcript (_.trs) file[2] with time-indexed locations into the audio file, and iii) extended archive descriptor (_.ead) file that contains metadata. Each audio stream has a reverse index word histogram (with all stop words—"the, a, or, …" set aside) that is employed with the natural language processing text search engine. These integrated files form the statistical information retrieval (SIR) engine.

The online search engine (see Fig. 6) is responsible for information retrieval tasks, including a web-based user interface as the front-end and search and index engines at the back-end. 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 audio archive collection, due to both copyright and disk space issues. Several hundred hours of audio have been digitized by MSU, and a portion is accessible via SpeechFind [see Fig. 6(b)].

V. SpeechFind: Transcribing Audio Archives

As Fig. 4 illustrates, the enrollment phase for an audio stream first requires audio segmentation and clustering (see Section V-A). Having segmented the stream, speech recognition is performed for transcript generation (Section V-B). In Section V-C, we also consider advances in acoustic model adaptation to improve transcripts for non-native and native speakers. Finally, Section V-D considers the text-based information-retrieval (IR) search framework.

A. Spoken Archives Segmentation

The goal of audio segmentation and classification is to partition and label an audio stream into speech, music, commercials, environmental background noise, or other acoustic conditions. This preliminary stage is necessary for effective LVCSR, audio content analysis and understanding, audio information retrieval, audio transcription, audio clustering, and other audio recognition and indexing applications. Audio archive segmentation obtains manageable audio blocks for subsequent speech decoding, as well as allowing for location analysis of speaker(s), channel, and environmental change points 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 that are 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, traditional MFCCs may not be as effective for speaker segmentation. In this section, we consider segmentation for several features (e.g., PMVDR [10], SZCR, FBLC) and performance of the CompSeg segmentation scheme for the NGSW audio data.

1) Fused Error Score (FES): Alternative Evaluation Criterion: The goal of reliable audio stream segmentation is to measure the mismatch between hand/human segmentation and automatic segmentation. In ASR, an integrated measure such as WER incorporates substitutions, deletions, and insertions. Frame accuracy is generally used as a measure of segmentation performance; however, it may not be the best criterion for audio/speaker segmentation since frequent toggling between classes results in short audio segments that are not helpful for automatic transcription if model adaptation is used (i.e., longer homogenous segments are better than numerous short segments). Equal Error Rate (EER) is another popular evaluation criterion in segmentation. However, the miss rate can be more important than the false alarm rate, and the average time mismatch between experimental and actual break points is also important. Therefore, the proposed *FES* [15] combines the three evaluation criteria of false alarm rate, miss rate, and

[2]Note that the (_.trs) file format follows LDC Transcriber format. Adopting this format for transcript structure is critical for future advances in sharing digital audio content through library services.

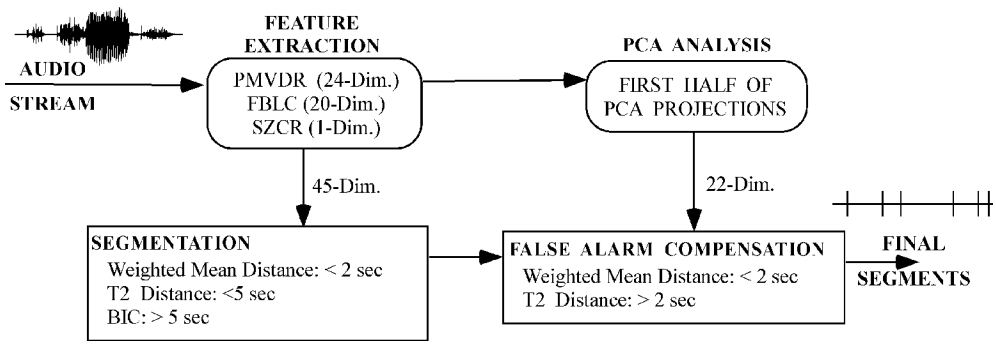


Fig. 7. Block diagram of CompSeg segmentation algorithm.

average mismatch in a manner similar in principle to WER and accuracy in ASR, as follows:

$$FusedErrorScore = (FalseAlarmRate\% + 2 \times MissRate\%) \times AvgMismatch_{\text{msec}} \quad (1)$$

where the average mismatch is in milliseconds, and the false alarm and miss rates are in percent. It is suggested that the FES integrates the three criteria that are important for assessing segmentation performance. It should be emphasized that this is a suggested integrated score, and we do not mean to imply that this is the best performance metric since other measures are clearly possible. The hope is that by suggesting the FES, other researchers would consider proposing alternative scoring strategies.

2) Compound Segmentation (CompSeg) and Weighted GMM (WGN) Classification: The proposed CompSeg algorithm uses a combination of three feature sets [PMVDR (24-Dim.: 12 static PMVDR, first 11 delta-PMVDRs, and static energy), SZCR (20-Dim.), FBLC (1-Dim.)], resulting in a 45-dimensional set. It applies a previously formulated T2-Mean distance measure for segments of duration shorter than 5 sec [15], BIC model selection [12] for longer duration segments, and, finally, a novel False Alarm Compensation post-processing routine. The block diagram of the CompSeg algorithm is shown in Fig. 7. The basis for using the Hotelling T^2-Statistic [16], [79] for speaker segmentation is the following: If two audio segments can be modeled by multivariate Gaussian distributions $\mathrm{N}(\mu_1, \Sigma_1)$ and $\mathrm{N}(\mu_2, \Sigma_2)$, and their covariances are equal but unknown, then the only difference is the mean values reflected in the T^2 distance as

$$T^2 = \frac{ab}{a+b}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) \quad (2)$$

where a, b are the numbers of frames within each of the audio segments, respectively. Using the T^2 statistic with BIC results in the T2-BIC segmentation scheme [16]: an improved version of original BIC [12] that performs segmentation 100 times faster with higher accuracy for short duration turns of less than 2 sec.

Under the equal covariance assumption, we can use more data to estimate the covariance and reduce the impact of insufficient data in the estimation. This is why the T^2 distance measure can detect the break point accurately. If the processing audio window is shorter than 2 sec, even a global covariance will suffer from insufficient estimation. We can then further assume the global covariance to be an identity matrix, in which case, we call this the Weighted Mean Distance. Therefore, the T^2 mean can be used to detect the break point in the short processing window (< 5 s) efficiently. As the window grows in duration, the covariance can be estimated more accurately, and we can then apply BIC to detect the break points directly, as in [12]. This summarizes the CompSeg segmentation scheme.

For audio classification, we employ a Weighted GMM Network (WGN), which uses the Variance of Spectrum Flux (VSF) or the Variance of Zero-Crosing Rate (VZCR) to weight a GMM network classification for speech/nonspeech classification [78].

3) Segmentation With Three Alternative Features: Having developed a new integrated evaluation criterion, we now turn to improved features for segmentation. We consider three features and compare them to traditional MFCCs (see [15] for further details).

PMVDR: High-Order Minimum Variance Distortionless Response (MVDR) provides 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 (ZCR) has previously been proposed for speaker classification [95]. We propose that a smoothed ZCR can be effective for segmentation [15] and computed using five sets of ZCR evenly spaced across an analysis window with no intermediate overlap. The SZCR is the mean of the five sets for this frame.

FBLC: Although it has been suggested that direct warping of the fast Fourier transform (FFT) power spectrum without filterbank processing can preserve most 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.

4) 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]. We first consider segmentation performance using the Hub4 '96 training data and the Hub4 '97 evaluation data. Table I shows that PMVDR can outperform MFCC on all levels (see [15] for more details). FBLCs have very small average mismatch, implying that they

TABLE I
SDR SEGMENTATION FEATURE PERFORMANCE. NOTE THAT '(x.x%)' REPRESENTS THE RELATIVE IMPROVEMENT IN FA: FALSE ALARM RATE, MIS: MISS DETECTION RATE, MMATCH: AVERAGE MISMATCH (IN MILLISECONDS), AND FES: FUSED ERROR SCORE

Feature	FA: False Alarm	MIS: Miss Rate	MMatch: mismatch in (msec)	FES: fused error score
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%)
MPVDR	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 II
SDR SEGMENTATION PERFORMANCE USING COMPSEG WITH IMPROVED FEATURES, AUDIO CLUSTERING, AND FALSE ALARM COMPENSATION WITH I) DARPA HUB4-97 BROADCAST NEWS DATA AND II) SAMPLE 1960S NGSW AUDIO MATERIALS. RELATIVE IMPROVEMENT OVER BASELINE IS SHOWN AS (XX.X%)

(i)

Algorithm	FA: False Alarm	MIS: Miss Rate	MMatch: mismatch in (msec)	FES: fused error score
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)

NGSW 1960's Data	Speaker Change	Speaker MMatch	Music & Sil Change	Music & Sil MMatch	False Alarm
	100%	129msec	26.9%	118msec	5.6%

are very sensitive to changes between speakers and environments. Because PMVDR does not apply filterbank processing, we combine PMVDR and FBLC together. In addition, 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 one SZCR (i.e., a 45-dimensional set). We normalize the features to zero mean and unit variance for improved discrimination ability.

5) NGSW and DARPA Hub4 Segmentation Evaluation: The DARPA Hub4 1997 Evaluation Data was used for segmentation performance assessment. The set contains 3 hr of Broadcast News data, with 584 break points, including 178 short segments (< 5 s). CompSeg uses PMVDR, SZCR, and FBLC features and applies T^2-Mean measure for segments of less than 5 sec and a novel False Alarm Compensation post-processing routine [15]. The improvement using these advances versus a baseline system employing MFCCs and traditional BIC [12] is shown in Table II(i). We also evaluate the CompSeg [15] algorithm with a portion of the NGSW corpus [1], using audio material from the 1960s. From Table II(ii), we see that CompSeg effectively detects not only speaker changes but music and long silence (> 2 s) segments as well.

B. Spoken Archives Transcription

For SpeechFind, all speech segments are decoded with a large vocabulary recognizer. We currently employ the CMU Sphinx3 for this task in this study. Using the Sphinx system, we employed acoustic models that contain 5270 GMMs, each of which has 32 mixture Gaussians. Acoustic models are built using a subset of the 200 hr of Broadcast News released by LDC in 1997 and 1998. The language model is composed of 64 K unigrams, 4.7 M bigrams, and 15 M trigrams. The average decoding speed is about 6.2 $\times$ real time on a P4-1.7 GHz Linux machine. In establishing the baseline experiments, no model adaptation schemes were applied at this stage, and first-pass decoding output is used as the automatic transcriptions, although a second-pass rescoring using a more complex language model might produce better results.

To evaluate recognition performance, 3.8 hr of sample audio data from the past six decades in NGSW is used as the test data. Table III 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, although the Out-Of-Vocabulary (OOV) rate does. Instead, the first three decades achieve better recognition accuracy, and the lowest WER is observed for corpora from the 1970s. This can be attributed to the lower average SNR for the recordings used from the 1980s and 1990s. For example, three long audio recordings from the 1990s that contain 2681 words have an average SNR near 12 dB, which produce WERs above 75%, whereas other recordings with a higher average SNR of 21 dB achieve WERs less than 25%. The average SNR of recordings from the 2000s is relatively high, whereas 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. This performance is sufficient for effective spoken document retrieval; however, it is clear that all possible methods for achieving robust speech recognition will need to be brought to bear to further reduce the WER as the diversity of the audio materials continues to expand.

C. Model Adaptation for Automatic Transcription

From Section V-B, it is clear that advances in acoustic model adaptation would improve speech recognition performance. Currently, the most commonly used speaker adaptation algorithms include transformation-based techniques, the most common being Maximum Likelihood Linear Regression (MLLR) [82], which is achieved with affine transformations,

TABLE III
DESCRIPTION AND EVALUATION PERFORMANCE OF A SAMPLE PORTION OF THE NGSW AUDIO CORPUS (29 628 WORDS, 3.8 HR)

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

and Bayesian learning that includes Maximum *a Posterior* (MAP) [81] scoring, which combines adaptation data with some *a priori* knowledge of the model parameters. In addition, there are also several extensions to MAP and MLLR that have been extensively investigated in recent years. These include regression-based model prediction (RMP) [80], Structural MAP [84], block-diagonal MLLR [83], MAP Linear Regression (MAPLR) [86], [87], and Structural MAPLR [85], among others (refer to the review in [88] for more comparisons). For relatively small amounts of adaptation data, transformation-based schemes have demonstrated superior performance over MAP due to their global adaptation via transformation sharing. On the other hand, MAP adaptation is more desirable for its asymptotic convergence to maximum likelihood estimation when the amount of adaptation data continues to increase [81]. However, MLLR and MAP have not shown comparable improvements when only limited amounts of adaptation data are available (e.g., 5 sec of adaptation data). In this study, we consider a novel approach based on primary eignendirections called EigMap [89].

1) EigMap and SMLEM Model Adaptation: The basic idea of EigMap [89] is to maintain the between-class variances (i.e., the discrimination power) of the baseline Gaussian means along the first primary eigendirections in a test speaker's eigenspace. Given the primary eigendirections $\{\mathbf{e}_{\mathrm{y}1}, \mathbf{e}_{\mathrm{y}2}, \ldots \mathbf{e}_{\mathrm{y}p}\}$ of a test speaker's observation covariance matrix $\mathbf{B}_{\mathrm{y}}$, the adapted Gaussian means $\{\mathbf{y}_i \mid \mathrm{i} = 1, 2, \ldots, \mathrm{N}\}$ are expected to satisfy the following relationship:

$$\sum_{j=1}^{n} y_{ij} e_{ymj} = \sum_{j=1}^{n} x_{ij} e_{xmj}, \quad m = 1, \ldots, p. \tag{3}$$

For every component Gaussian x_i in the model Λ, all possible adapted means $\mathbf{y}_i$ that satisfies (3) form a $(\mathrm{n} - \mathrm{p})$-dimensional subplane $\Omega(\mathbf{x}_i)$ in the acoustic space that is given by

$$\Omega(\mathbf{x}_i) = \left\{ \mathbf{y}_i \mid \sum_{j=1}^{n} y_{ij} e_{ymj} = \sum_{j=1}^{n} x_{ij} e_{xmj}, \quad m = 1, \ldots, p \right\}. \tag{4}$$

Rapid model adaptation when only sparse observation data is available requires care, since not all Gaussians in the acoustic space will be represented in the adaptation data. A conservative approach is to minimize the shift from the well-trained baseline model parameters, given the constraint of no loss of discrimination power along the first dominant eigendirections in the test speaker eigenspace:

$$\mathbf{y}_i = \underset{\mathrm{y}_i \in \Omega(\mathrm{x}_i)}{\arg\min} (\mathbf{x}_i - \mathbf{y}_i)^T (\mathbf{x}_i - \mathbf{y}_i). \tag{5}$$

By substituting (4) into (5) and minimizing the objective function using the Lagrange Multiplier method, the adapted mean $\boldsymbol{y}_i$ can be obtained from $\boldsymbol{x}_i$ using a linear transformation $f : \mathbb{R}^n \rightarrow \mathbb{R}^n, \mathbf{y} = f(\mathbf{x}) = \Theta\mathbf{x}$, in which Θ is an $n \times n$ nonsingular matrix given by

$$\Theta = \mathbf{I}_n - \sum_{i=1}^{\mathrm{p}} (-1)^{(\mathrm{i}-1)} \mathrm{e}_{\mathrm{yi}}^T (\mathrm{e}_{\mathrm{yi}} - \mathrm{e}_{\mathrm{xi}}) \tag{6}$$

and where $\mathbf{I}_n$ is an $n \times n$ identity matrix. Considering the orthogonality between eigenvectors, one can show that $\Theta = \mathbf{E}_{\mathrm{y}}^{-1} \mathbf{E}_m$. After transforming the baseline model mean $\mathbf{x}$ into $\mathbf{y}$ using (6), the discrimination information is assumed to be mostly encapsulated in the first q dimensions where $p < q < n$; hence, the last $n - q$ dimensions of $\mathbf{y}$ can be discarded. In model space, this can be represented by setting the last $n - q$ rows of Θ to zeros

$$\overline{\Theta}_i = [\Theta_{q \times n}, 0_{(n-q) \times n}]^T \tag{7}$$

and the adapted Gaussian mean $\mathbf{y}$ is achieved through following transformation:

$$y_{q \times 1} = \overline{\Theta}_{q \times n} \mathrm{x}_{n \times 1}. \tag{8}$$

From the above equation, we note that the baseline model is not only adapted through the transformation but also compressed with reduced Gaussian dimensions of model mean, which further suggests that faster recognition speed can also be achieved using the adapted model due to reduced Gaussian computations.

A number of extensions to this EigMap model adaptation scheme have also been considered [89]. One such extension is SMLEM, which extends the core EigMap algorithm by imposing a further shift in the model space to maximize the adaptation data likelihood. To account for the adaptation data likelihood, the EigMap formulation can be extended by adding a linear bias $\mathbf{b}$ in the test speaker's eigenspace:

$$\mathbf{y} = \overline{\Theta}\mathrm{x} + \mathrm{E}_{\mathrm{y}}^{-1} \boldsymbol{b} \tag{9}$$

where $\mathbf{b}$ is derived in a manner that maximizes the adaptation data likelihood $P(\mathbf{O} \mid \Lambda)$ given the model Λ. Since only the Gaussian means are adapted, we ignore other model parameters in the auxiliary function using the EM algorithm. Further details concerning this extension can be found in [89].

2) Evaluations: The baseline speaker-independent acoustic model has 6275 context-dependent tied states, each having 16 mixture component Gaussians (i.e., in total, 100 400 diagonal mixture component Gaussians exist in the acoustic model). The baseline system uses a feature of 39 dimensions with 13 static cepstral coefficients plus delta and double-delta. The baseline speech recognition system used for experiments is the CMU

TABLE IV
WER IN PERCENT OF NON-NATIVE SPEAKERS (WSJ SPOKE3) WITH APPROXIMATELY 4.5 SEC OF UNSUPERVISED ADAPTATION DATA

	Baseline (%)	BD-MLLR (%)	EigMap (%)	SMLEM (%)
Avg. Across Speakers	20.7	17.4	16.9	16.2
Relative Improvement	---	15.9%	18.4%	21.7%

TABLE V
TRANSCRIPTION EVALUATION OF SAMPLE AUDIO TEST DATA FROM THE PAST SIX DECADES AFTER UNSUPERVISED MODEL ADAPTATION USING MLLR+EIGMAP

Decade	Substitution (%)	Deletion (%)	Insertion (%)	Avg. WER (%)	Relative Improvement
1950	24.5	3.8	4.7	33.1	14.2%
1960	20.2	3.5	5.6	29.3	20.2%
1970	14.5	5.5	1.9	22.0	14.1%
1980	30.3	13.5	2.9	46.6	22.5%
1990	30.4	7.9	3.8	42.1	12.3%
2000	27.7	16.2	11.7	55.6	5.9%

Sphinx-3.2 recognizer. The language model is a standard 5000-word back-off trigram. In offline multistream tree-structured Gaussian clustering, the 100 400 component Gaussians in the SIR models are grouped into 300 base classes, and hence, a binary tree with 599 nodes is used to represent the structural space for each stream.

The experimental results on the WSJ Spoke3 corpus (non-native speakers) are summarized in Table IV. On average, about 4.5 sec of adaptation data was used. Due to the mismatch between the model and test data, the averaged baseline model WER performance is as high as 20.7%. Table IV clearly shows that EigMap consistently improves the recognition for all non-native speakers, with a relative improvement of 18.4%, whereas BD-MLLR achieves a 15.9% relative improvement. By applying SMLEM to maximize adaptation data likelihood after EigMap, the overall relative performance is further improved to 21.7%.

Next, EigMap and SMLEM are applied to NGSW data. Since the NGSW task is a large-scale real-world task, an increased vocabulary size of 64 K is used. The acoustic models are cross-word triphones with 5270 decision tree-based tied states, each with 32 mixture Gaussians (i.e., a total of 168 640 mixture Gaussians are used). In offline multistream tree-structured Gaussian clustering, these component Gaussians in the baseline models are grouped into 500 base classes, and hence, a binary tree with 999 nodes is used to represent the structured eigenspace for each stream. Acoustic models are trained using a subset of the 200 hr of BN data from LDC (released in 1997 and 1998). The system uses a backoff trigram language model that is composed of 64 K unigrams, 4.7 M bigrams, and 15 M trigrams.

Unlike the WSJ corpus, NGSW contains continuous raw recordings with multiple speakers. As such, the T2-BIC segmentation scheme [16] previously discussed was used to first segment data into ideally homogenous segments of duration 35 sec or less.

Audio streams used in the evaluation from Table III are now considered for evaluating EigMap model adaptation. In Table V, we summarize transcription evaluations of sample audio data from the six-decade set after employing unsupervised adaptation using MLLR+EigMap. Compared with the baseline performance from Table III, it is clear that transcription performance for audio data from every decade has been uniformly improved. For some decades, the relative improvement is as high as 19.6%. For audio documents with extremely noisy segments and overlapping and indistinguishable speech such as the 1980s and 2000s, the deletion errors are still relatively high after adaptation. However, the effects of these errors for the overall SDR task can be greatly alleviated through a technique known as document expansion (see Section V-D).

TABLE VI
(I) DESCRIPTION OF SAMPLE NGSW HISTORICAL AUDIO DATA FOR EIGENSPACE MAPPING-BASED MODEL ADAPTATION. (II) EVALUATION OF EIGENSPACE MAPPING USING EIGMAP AND SMLEM MODEL ADAPTATION METHODS

Audio Doc	Baseline WER	MLLR	EigMap	SMLEM	MLLR+EigMap
1950	50.3%	44.1%	47.7%	45.3%	42.4%
1960	27.2%	24.8%	25.9%	24.6%	24.4%

Audio	Decade	# of words	Duration	OOV	SegSNR	# of Segs	Dur/seg
VVL00402	1950	631	351 sec.	2.06%	32.63 dB	33	10.6 sec.
VVL00114	1960	1476	687 sec.	3.18%	28.49 dB	46	15.6 sec.

We consider sample audio streams from the 1950s and 1960s in more detail in Table VI. The OOV rates for the recordings varied from 2–4%. The average length of the audio blocks after T2-BIC segmentation was 10.6 and 15.6 sec, respectively. Here, the SegSNR score from the NIST evaluation software reflects the level of background noise present in the audio data and helps explain why WERs were high (e.g., some background music, speaker variability, and acoustic noise). EigMap and SMLEM are both effective in reducing WER over the baseline system. MLLR achieves better performance for the 1950s than SMLEM, and SMLEM performs better than MLLR for the 1960s audio. Since the average segment duration is more than 10 sec, the amount of adaptation data is more than double what was available in Table IV using WSJ data. This increased amount of adaptation data allows for better estimation with MLLR. The performance is also better with WSJ data because that audio material is noise-free. Therefore, it is suggested that when noise is present, it may be necessary to consider both noise and speaker adaptation within the model adaptation process.

D. IR Over Automatic Transcripts and IR Advances

The current SpeechFind retrieval engine is a modified version of MG [17]. Here, the tfidf weighting scheme is replaced with Okapi weighting [90], and several query and document expansion technologies are incorporated. To ensure a sufficient number of documents from the perspective of IR, the

TABLE VII
DESCRIPTION OF DOCUMENT AND QUERY SETS

Number of Documents	956
Average Length of Documents	14 secs
Average # of Words per. Document	30 words
Number of Queries	25
Average Length of Queries	4.6 words
Number of Relevant Documents	324
Average Relevant Documents per. Query	13 docs

transcript from each recognition segment is treated as a single document. In our case, many historical spoken documents are typically longer than 30 min; therefore, the use of small segments as a search unit allows for a more specific user search. The SpeechFind web interface provides the user access to the detected speech segments and automatic transcripts and allows the user to preview/listen to any parts of the entire audio file containing the original detected segments.

Table VII describes the spoken document and query sets used in the evaluation. Here, 25 test queries were designed by an independent human researcher, based on human transcripts, and human relevance assessments were made based on the audio content of the corresponding segments. For indexing, stemming and case folding are performed, but no stop words are removed. As our document lengths are considerably shorter than corpora used in the IR literature, we must tune the Okapi parameters for our task. The baseline average precision after spoken transcripts and query normalization is 42.17%, with the best performance achieved when $k1 = 0.3$ and $b = 0.75$ for the Okapi weighting scheme (see Table VIII). In the following subsections, we report our retrieval performance with experiments on the transcribed spoken documents.

1) Spoken Transcripts and Query Normalization: An inherent difference exists between transcribed spoken documents and typical text documents. Automatic transcriptions essentially decode acoustic recordings using the most probable in-vocabulary word sequences. On the other hand, 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 1960, but it is usually not included in either the dictionary or language models in most state-of-the-art speech recognizers. Hence, the audio phrase will appear as "nineteen sixty" in automatic spoken document transcripts. To address this issue, the spoken transcripts and queries are normalized in the SpeechFind system to bridge this gap. Through a predefined dictionary of mappings between "spoken words" and "simplified human notations," the automatic transcripts are filtered, which, for example, replace "N. B. C." with "NBC." Using an inverse of a similar dictionary, the queries are filtered as well (e.g., we change the query word "1st" to "first").

a) Query Expansion Using BRF: Query Expansion (QE) is an application that could be used to address the problem of missing query terms directly or missing term relations indirectly [91]. We first experiment with query expansion using Blind Relevance Feedback (BRF) on the test collection. Here, we consider explicitly adding new terms to the existing query Q. We suppose that the top R returned documents are related to the original query in the first round of retrieval; then, T expansion terms are chosen according to their Offer Weight (OW) [91] ranking in these R documents. It should be noted that stop words, which are defined in a list of 371 common words that appear in the R documents, are first excluded as expansion terms. We experiment with several pairs of R and T, and the results are summarized in Table VIII. The best result achieved is 44.72% when setting $\mathrm{R} = 10$ and $\mathrm{T} = 5$.

TABLE VIII
AVERAGE PRECISION FOR QUERY AND DOCUMENT EXPANSION

Doc Expan		Query Expan		
R	T	R	T	Avg. Precision(%)
–	–	5	1	42.22
–	–	9	5	43.82
–	–	10	3	44.22
–	–	10	5	**44.72**
–	–	10	6	42.76
2	10%	–	–	42.99
3	10%	–	–	**47.58**
3	15%	–	–	46.36
4	10%	–	–	43.94
3	10%	3	2	50.05
3	10%	4	2	**50.59**
3	10%	5	2	49.32
3	10%	6	3	48.51
3	10%	10	5	46.97
Baseline				**42.17**

b) Document Expansion Using PBRF: The idea behind Document Expansion (DE) [92] is that given a document, first identify other parallel documents related to those in hand, and bring "signal" words from the related documents into the present document. To expand spoken documents, we first run automatic transcription of the speech document as a query on a parallel collection, and then, the query documents are expanded using Parallel Blind Relevance Feedback (PBRF).

The effect of document expansion largely depends on the selection of the parallel text collection, which should be related to the spoken corpus. To construct a parallel text collection for our audio recordings, we fetch and parse related historical documents of the 20th century from the web [97]. We also include available human transcripts from NGSW audio data for the same period. The parallel collection contains about 150 K words.

In our experiments, we use the same scheme of BRF to expand automatic transcriptions (i.e., using the original spoken transcriptions as the query to search over the parallel collection; all stop words in the spoken documents are not included in the queries). The top R returned documents are assumed to be relevant, and then, the T expansion terms are chosen to expand the spoken document according to their ranking in terms of a weight scheme in these R documents. Since the transcribed audio segments have considerable length variations, we make T equal to some percentage of the number of terms in each original automatic audio transcription (which achieves better performance than picking a fixed number of terms for all spoken documents).

To rank the candidate terms, we propose the rtfrw weighting scheme:

$$\begin{aligned}\mathrm{rtfrw}(\mathrm{t}_i) &= \mathrm{rtf}(\mathrm{t}_i) \bullet \mathrm{rw}(\mathrm{t}_i) \\ &= \mathrm{rtf}(\mathrm{t}_i) \bullet \log\left(\frac{(\mathrm{r}+0.5)(\mathrm{N}-\mathrm{n}-\mathrm{R}+\mathrm{r}+0.5)}{(\mathrm{n}-\mathrm{r}+0.5)(\mathrm{R}-\mathrm{r}+0.5)}\right)\end{aligned} \quad (10)$$

where $rw(t_i)$ is the Relevance Weight defined in [91], r is the number of assumed relevant documents, where the term t_i

occurs, R is the the number of assumed relevant documents for a query, n is the number of documents in the collection where term t_i occurs, and N is the total number of documents in the collection; $\mathrm{rtf}(t_i)$ is the term frequency of term t_i in the R assumed relevant documents for a query. Candidate expanding terms are selected based on their ranking of rtfrw weight. Again, the stop words are excluded for expansion. Using the rtfrw weight achieves better results than using OW in our experiments. As shown in Table VIII, the best performance for using rtfrw weighting is 47.58%, whereas the best performance using OW is 43.76% {which is not shown in the table}.

After obtaining expanded spoken documents, the original queries can also be expanded based on expanded spoken document collections. The results of DE+QE (see Table VIII) clearly show that appropriate choices of R and T can be obtained, and the performance of DE using PBRF and QE using BRF are additive. The combination of DE+QE achieves an average precision of 50.59%, which is a relative 20% improvement from the baseline precision of 42.17%.

VI. Digital Speech Watermarking

As discussed in Section III, the NGSW employs SpeechFind to search a database containing rare and valuable audio material, the copyright to which is held by numerous individuals, estates, corporations, and other private and public interests. Accordingly, in accessing and distributing NGSW content, SpeechFind must integrate "speech protect" watermarking technologies.

The process of embedding a digital watermark [70] perturbs original signal content, ideally imperceptibly. Fidelity of the original content is adversely affected by increased perturbation, whereas the robustness [70] of the watermark to attack is generally improved by increased signal distortion. The first component of our strategy is the class of techniques called *parameter-embedded watermarking*. Parameter-embedded watermarking is effected through slight perturbations of parametric models of some deeply integrated dynamics of the speech. A second component is the deployment of innovative new methods for estimation of parametric models known as *set-membership filtering* (SMF). SMF provides the means to identify parametric watermarks that result in a rigorously quantified fidelity criterion. The third component of the strategy, which is also based on SMF, results in a set-solution of watermark candidates, each element of which adheres to the fidelity requirement. The properties of this set solution make it possible to develop multiwatermark strategies designed to guard against attacks.

A. Algorithmic Methods

1) Parameter-Embedded Watermarking: A general formulation of parameter-embedded watermarking is given in recent papers [6], [71]. Here, the signal to be watermarked (*coversignal*), say $\{y_n\}$, is assumed to follow the linear prediction (LP) model [72],

$$y_n = \sum_{i=1}^{M} a_i y_{n-i} + \xi_n = \mathbf{a}^T \mathbf{y}_n + \boldsymbol{\xi}_n \tag{11}$$

with coefficients $\{a_i\}_{i=1}^{M}$ and prediction residual $\{\xi_n\}$. For a frame to be watermarked, the LP parameters are perturbed by an independent (known) watermark vector by direct addition or by addition to the autocorrelation sequence for the frame.[3] The watermarked signal (*stegosignal*) is constructed by employing the perturbed LP coefficients, say $\{\overline{a}_i\}_{i=1}^{M}$, and the exact prediction residual in the FIR filter

$$\overline{y}_n = \sum_{i=1}^{M} \overline{a}_i y_{n-i} + \xi_n = \overline{\mathbf{a}}^T \mathbf{y}_n + \xi_n. \tag{12}$$

Parametric watermarking was found to be robust against a wide variety of attacks, as discussed in [6]. In particular, the technique is highly robust to additive white noise for two reasons: First, the solution has been shown to be asymptotically immune to additive white noise [73], and second, the parameter estimation process severely attenuates the noise energy in the parameter domain with respect to the signal domain [74].

An essential feature of any digital watermarking algorithm is security. Security refers to the ability of the technique to avert unauthorized detection, embedding, or removal. The following aspects of the parameter-embedding algorithm contribute to its security: Speech frames to be watermarked can be selected randomly, and the LP model order can vary across watermarked frames. In addition, a copy of the coversignal is required for watermark recovery. Because the prediction residual associated with the coversignal is used for reconstructing the stegosignal, the autocorrelation values of the stegosignal are different from the modified autocorrelation values derived from the perturbed LP coefficients and the prediction residual $\{\xi_n\}$. Hence, watermark recovery is rendered extremely difficult without a copy of the coversignal.

2) SMF-Based Fidelity Criterion: In [71], a general parameter-embedding problem was considered whose solution is subject to an ℓ_∞ fidelity constraint on the signal. This constraint can be generalized further to allow for more "local" fidelity considerations in time as the signal properties change. The SMF concept [75], [76] can be viewed as a reformulation of the broadly researched class of algorithms concerned with *set-membership identification* [77]. The SMF problem is used to design systems that are *affine-in-parameters* (but not necessarily in the data), subject to a bound on the absolute error between a desired sequence and a linearly filtered version of another sequence. The two sequences may be directly observed, or they may be nonlinear combinations of other sequences considered to be the system inputs and outputs.

Formally, the SMF problem is stated as follows: Given a sequence $\{x_m \in \mathbb{R}^M\}_{m=1}^{n}$ of observations, a "desired" sequence $\{z_m \in \mathbb{R}^M\}_{m=1}^{n}$ and a sequence of error "tolerances"$\{\gamma_m\}_{m=1}^{n}$, find the exact feasibility set $\mathcal{P}_n \subseteq \mathbb{R}^M$ of filters $\theta \in \mathbb{R}^M$ at time n

$$\mathcal{P}_n = \left\{\theta \middle| \left|z_m - \theta^T x_m\right| < \gamma_m, m \in [1, n]\right\}. \tag{13}$$

[3]Earlier work [6] suggested robustness benefits using autocorrelation perturbations, but further experimentation has shown that the relative robustness in LP or autocorrelation domains depends on the nature of embedded watermark vectors.

The solution uses a series of recursions that ultimately return a hyperellipsoidal *membership set*, say $\mathcal{E}_n \supset \mathcal{P}_n$, and the ellipsoid's center, say θ_n. The recursions execute an optimization strategy designed to tightly bound $\mathcal{P}_n$ by $\mathcal{E}_n$ in some sense. Accordingly, the broad class of algorithms employed in the SMF problem is often called the *optimal bounding ellipsoid* (OBE) algorithms (see the tutorial papers [76] and [77] for details).

The construction of a watermark set guaranteed to satisfy a fidelity criterion is readily solved as an SMF problem. Subtracting y_n from each side of (12) and then rearranging yields

$$\overline{y}_n - y_n = \sum_{i=1}^{M} \overline{a}_i y_{n-i} + \xi_n - y_n = \mathbf{a}^T \mathbf{y}_n - (y_n - \xi_n). \quad (14)$$

A fidelity criterion is prescribed in the form of a sequence of pointwise absolute bounds $\{\gamma_n\}_{n=1}^{N}$ on the coversignal perturbation: $|y_n - \overline{y}_n| < \gamma_n$ for each $n \in [1, N]$. Upon defining the sequence $z_n = y_n - \xi_n$, $n = 1, 2, \ldots, N$ (recall that $\{\xi_n\}$ is known), the search for the constrained watermark parameters is reduced to an SMF problem, as in (13). The result of applying the SMF estimation is the hyperellipsoidal set of watermark (perturbed model parameter) candidates $\mathcal{E}_{\mathcal{N}}$ guaranteed to tightly bound the exact set

$$\mathcal{P}_n = \left\{ \overline{\mathbf{a}} \in \mathbb{R}^M \middle| |z_n - \overline{\mathbf{a}}\mathbf{y}_n| < \gamma_n, n = [1, N] \right\}. \quad (15)$$

3) Watermark Recovery: For the watermark recovery algorithm, the LP model is used to parameterize long intervals of stationary or nonstationary speech. However, to understand the robustness aspects ξ of the watermarks, it is necessary to consider stationary segments of the coversignal and the stegosignal. That is, segments of $\{y_n\}$, $\{\xi_n\}$, and, hence, $\{\overline{y}_n\}$ are assumed to be partial realizations of wide-sense stationary and ergodic random processes.

Watermark recovery is effected through least-square-error (LSE) estimation of the perturbed parameters $\{\overline{a}_i\}_{i=1}^{M}$ in the following manner. Let us rewrite the stegosignal generation (12) as

$$d_n = \sum_{i=1}^{M} \overline{a}_i y_{n-i} = \overline{\mathbf{a}}^T \mathbf{y}_n \quad \text{with} \quad d_n = \overline{y}_n - \xi_n. \quad (16)$$

In principle, this system of equations taken over $n = 1, 2, \ldots, N$ is noise free and can be solved for $\overline{\mathbf{a}}$ using any subset of M equations. For generality, to smooth roundoff and to support further developments, we pose the problem as an attempt to compute the LSE linear estimator of the "desired" signal d_n given observations y_n.

The conventional set of normal equations is solved to produce the estimate of $\{\overline{a}_i\}_{i=1}^{M}$. This formulation admits a proof of asymptotic unbiasedness in the presence of an additive white noise attack and the introduction of prewhitening measures to likewise mitigate the effects of additive colored noise. Experiments with a variety of attack modes are described in [6], where it is reported that the parametric embedding is quite robust to additive noise, MP3 compression, jitter attack, cropping, and requantization. Filtering attacks are reported to be more challenging in that account. More recent efforts have concentrated on robustness to filtering with improved results [6], [71], [73]. Because of the correlation in the noise, however, the solution is biased asymptotically. Depending on the level of cross-correlation, the LSE estimation of the perturbed coefficients, and hence the watermark signal, may be affected. A remedy is to whiten the noise process, in which case, the effect of noise will be similar to the white noise attack.

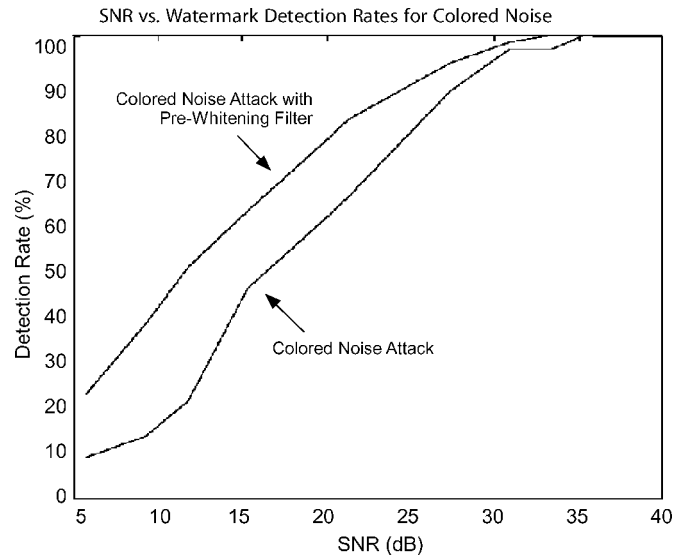


Fig. 8. Detection rates for correlated noise attack.

To demonstrate performance, we select a 3-sec coversignal of Thomas Edison's speech from the NGSW [1] [a portion is shown in Fig. 2(a)]. The signal was sampled at 44.1 kHz and partitioned into 33 frames of 3969 samples. A watermark was embedded into each of the frames. A fourth-order LP model was used for watermarking, and the fidelity constraint was set so that $|\overline{y}_n - y_n| < 0.25\,|y_n|$ for each n. Correlated noise of various SNRs was added to the stegosignals. Fig. 8 shows the watermark detection rates versus SNR, with and without the use of a prewhitening filter. Improved performance is observed when prewhitening is employed.

VII. Further SDR Advances: Integrating Accent Knowledge and Language Modeling

It is expected that advances in speaker and acoustic analysis will provide new directions for supplementing SDR. Here, we consider the application of automatic accent classification to detect whether an audio sequence is native (neutral) or non-native American English. We applied the framework developed in [98] to audio files uttered by former U.S. President R. M. Nixon (i.e., "Therefore, I shall resign the presidency effective noon tomorrow") and his Secretary of State H. Kissinger (i.e., "There is no other choice Mr. President"), who is a native speaker of German. These audio sequences were selected from a collection on the Nixon Watergate tapes, with similar topic/word content. We employ the proposed algorithm in [98], where the speech utterances were converted into a stream of feature vectors (12 MFCCs and log-energy), and then tokenized into a set of phone

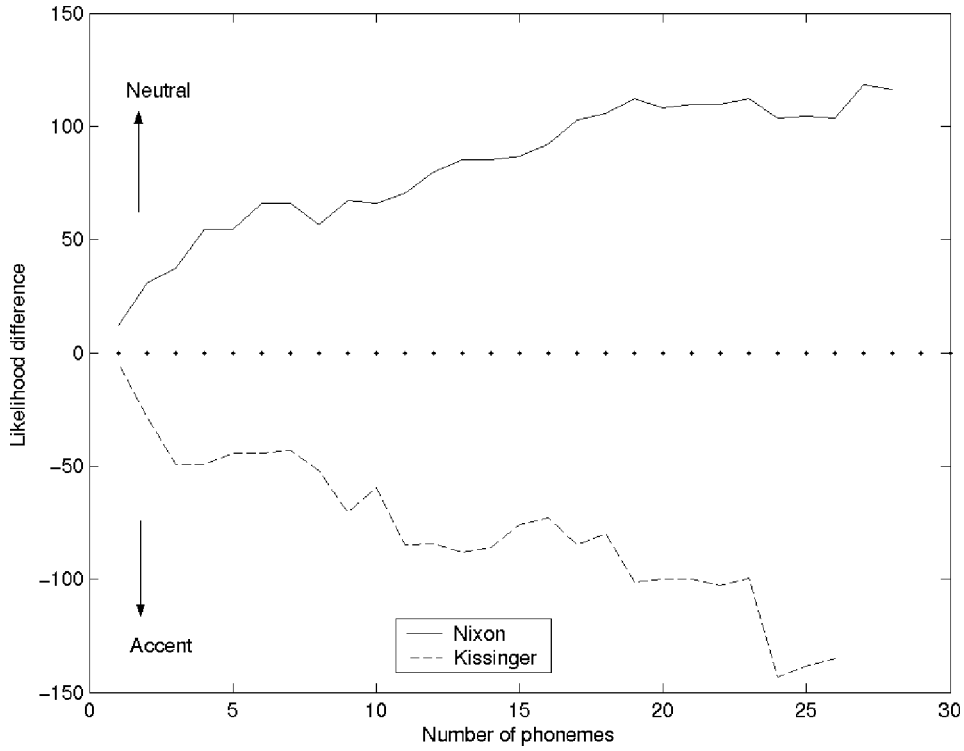


Fig. 9. Accumulated likelihood difference score for neutral/accent classification for Nixon and Kissinger audio sequences.

sequence using automatic forced alignment. During the detection stage, for each speech (phoneme) segment $\mathbf{X}$ and the corresponding phone-labeled s, the difference score between the neutral score and accent score is defined as

$$R(X, s) = L(X, \Phi_s^{(EN)}) - \min_{AC \in Q} L(X, \Phi_s^{(AC)}) \tag{17}$$

where $L(.)$ is a log-likelihood score generated from accent-dependent phone acoustic models, $\Phi_s^{(EN)}$ is a Stochastic Trajectory Model (STM) of phoneme s trained from Neutral American English speech, $\Phi_s^{(AC)}$ is an STM of phoneme s trained from accented speech, and Q is a set of prototype accents (i.e., $Q = \{\text{Chinese, Thai, Turkish}\}$; note that we deliberately do not include a German accent model since our focus is on open-set accent classification). Here, we used 35 context-independent phoneme STMs for each accent; each STM has two mixtures and five states obtained in a similar fashion to [99]. We suggest that the "min" argument should provide the smallest trajectory variation score from a set of accented models. Fig. 9 illustrates the plot of the accumulated summation of difference scores as the number of accumulated phonemes is increased, ignoring the score for which no phoneme acoustic models were trained. As the phoneme number increases, the accumulated difference score of the Nixon audio stream moves toward the more positive scale (neutral English scale), whereas the accumulated difference score of the Kissinger audio sequence moves toward the more negative scale (accent scale). Thus, this example shows that an accent classification system can be employed for tagging audio streams, even when that accent is not present in the model set (i.e., we used Chinese, Thai, and Turkish accent models). This suggests that future transcription-based strategies could be employed to enrich the search process by providing accent [98] or in-set speaker [100] recognition information.

Presently, SpeechFind employs whole-word search using our modified MG-based search engine from the resulting transcript outputs of a large vocabulary continuous speech recognizer. However, a number of alternative strategies are possible for text-based search using LVCSR. For BN stories, WERs can be low with much redundancy in the news stories, and therefore, search for key words over longer sequences is a reasonable approach. However, when searching for a specific string from an audio stream in the NGSW, it is expected that the particular text string may only occur once (e.g., the expression "I have a dream" occurs often in the speech by M. L. King concerning civil rights; however, the expression "One small step for man, one giant leap for mankind" is produced only once from the NASA transmissions from the moon). An effective approach to address this is to use output lattices of the word sets or subword units with weight pushing [101].

In addition to search methods, statistical language models play a key role in SDR for a corpus such as NGSW, which extends over 110 years. In general, for optimal retrieval performance, the index term representation should be based on words for which the queries and speech data can be easily mapped [66]. Proper names are typically used for queries, but they are difficult to recognize from speech because of pronunciation differences and the variety of names from such a time range. Practical issues often limit the applicable vocabulary for language models to be 65 k words, which increase OOV words. In [66], an experiment was performed using audio materials from the Chicago Roundtable discussions of the 1940s and the NGSW six-decade materials from Table III. A series of LMs

TABLE IX
SUMMARY OF EVALUATION OF SPEECH REFERENCE TRANSCRIPTS FROM SEVEN DECADES OF NGSW. [RESULTS INCLUDE OOV RATE, PERPLEXITY (PERPL.), AND TERM ERROR RATE (TER)]

Decade	Ref. # Words	Audio (mins)	SNR (dB)	BASELINE LM			Interpolated LM		
				OOV	Perpl.	TER	OOV	Perpl.	TER
1940	2068	14	10	0.5	258	59	0.6	177	55
1950	6241	52	34	1.5	325	42	1.5	280	33
1960	2142	17	20	2.2	384	33	2.5	343	29
1970	4434	35	21	0.8	132	34	0.8	151	24
1980	3330	27	21	0.9	177	42	0.8	194	35
1990	5951	47	14	1.7	280	82	1.8	285	78
2000	7530	50	28	0.9	237	75	0.9	285	75

was developed using standard back-off trigrams with interpolated Kneser–Ney smoothing [102], [103]. A 65 347-word vocabulary was selected based on the most common words in the BN and North American news texts (News) corpora. Speech decoding was performed using the SpeechFind framework using one LM at a time. LMs that were built include i) BN with Hub4 transcriptions using 168 M words, ii) News employs BN but also with News corpora totaling 724 M words, iii) Old which is an LM built using the Gutenburg archives from 1900–1920 texts and 1940s Chicago Roundtable texts (CTT), for a total of 5.7 M words, and iv) ALL, which is a language model using basically all the text from these sources, for a total of 750 M words. Evaluation of LM performance was assessed by computing the average of the inverse of the next word prediction probability (e.g., perplexity) on data that was left out of the LM construction. Perplexity evaluations were performed using the SRILM toolkit [103]. The perplexity of the four LMs ranged from 275 to 774 (see [66]), depending on which corpus of held-out text materials was used. We considered an *interpolated LM* by computing a new LM out of the two components using different interpolation weights. For this test, we used the smaller Old (using the Gutenburg text archives and CCT) and larger ALL language models from above. Table IX summarizes results, where the Baseline represents the original LM using BN data. The WER does not seem to be impacted much by LM improvements. For speech retrieval, however, the recovery of rare content words are more likely to influence statistical information retrieval than WER. Using the Term Error Rate (TER), which has been suggested to be a more effective performance measure for speech retrieval applications [39] and represents the difference between two word histograms (i.e., recognition result and correct transcription), we see that the large overall LM interpolated with the small old text LM improves perplexity and TER for the older time blocks, whereas for more recent time blocks, the modern BN (i.e., Baseline) model is better. This seems to suggest that focusing has an important effect for the speech transcripts as well as for the LM accuracy. The differences among the decades are due mostly to different recording methods and media as discussed in Section II. Further discussions concerning interpolated LMs and the NGSW are found in [66].

VIII. SUMMARY AND CONCLUSION

In this study, we have addressed a number of advances in establishing spoken document retrieval for a new National Gallery of the Spoken Word (NGSW). We first discussed an overview of the audio stream content of the NGSW with sample audio files. Next, we presented the SpeechFind system, which is an experimental online spoken document retrieval system for an historical archive with 60 000 hr of audio recordings from the last century. We introduced the SDR system architecture and focused on audio data transcription and information retrieval components. A number of issues regarding copyright assessment and metadata construction were discussed for the purposes of a sustainable audio collection of this magnitude. We considered a new segmentation performance criterion called the Fused Error Score (FES) and evaluated three features as alternatives to traditional MFCCs. 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 for a portion of the NGSW corpus and novel model adaptation using structure maximum likelihood eigenspace (SMLEM) mapping, which resulted in a relative 21.7% improvement over a BN-trained baseline speech recognition system. Information retrieval over automatic transcripts was considered by combining document expansion (DE) and query expansion (QE) using blind relevance feedback (BRF) and parallel blind relevance feedback (PBRF), which improves average returned document precision from a baseline of 42.17% to 50.59%. Advanced parameter-embedded watermarking and set-membership filtering-based fidelity criterion was proposed, with evaluations showing robustness to correlated noise attacks, as well as additive noise, MP3 conversion, jitter attack, and cropping.

SpeechFind was established as an experimental platform to perform SDR from historical archives. In the future, the system will be improved in a number of ways. First, the quality of automatic speech transcripts can be boosted by improving the baseline modeling through retraining of time-specific acoustic models. Recent work suggests that period specific language modeling can also help improve transcript generation [66]. Future work will include further integration of the model adaptation technologies, especially adapting the acoustic models for varied background noises and speakers, and adjusting the language models when topics and decades change dramatically. Moreover, richer information such as accent, stress, emotion, and speaker identification contained in spoken segments could also be extracted and used to guide retrieval tasks. An example was shown for accent classification using audio sequences

from Nixon and Kissinger. Further progress in SDR could benefit from improving IR performance. In our task, reliable document categorization could be achieved with the help of metadata associated with some spoken documents, (i.e., so-called EAD extended archive descriptor files used in library archive services), which narrows a search and, hence, improves the retrieval precision. Presently, SpeechFind tracks and displays EAD metadata, but research has yet to be performed to determine the weight balance given to information contained within EAD versus that obtained from transcript IR. In addition, a statistical retrieval framework incorporating the uncertainty of automatic transcripts is another interesting research topic, which would help improve user search performance.

REFERENCES

[1] . [Online]. Available: http://www.ngsw.org

[2] . [Online]. Available: (Original website) http://speechfind.colorado.edu/; http://speechfind.utdallas.edu/

[3] B. Zhou and J. H. L. Hansen, "SPEECHFIND: An experimental on-line spoken document retrieval system for historical audio archives," in *Proc. Int. Conf. Spoken Language Process.*, vol. 3, Denver, CO, Sep. 2002, pp. 1969–1972.

[4] J. H. L. Hansen, B. Zhou, M. Akbacak, R. Sarikaya, and B. Pellom, "Audio stream phrase recognition for a National Gallery of the Spoken Word: 'One small step'," in *Proc. Int. Conf. Spoken Lang. Process.*, vol. 3, Beijing, China, Oct. 2000, pp. 1089–1092.

[5] J. H. L. Hansen, J. Deller, and M. Seadle, "Engineering challenges in the creation of a National Gallery of the Spoken Word: Transcript-free search of audio archives," in *Proc. IEEE ACM Joint Conf. Digital Libraries*, Roanoke, VA, Jun. 2001, pp. 235–236.

[6] A. Gurijala, J. R. Deller Jr., M. S. Seadle, and J. H. L. Hansen, "Speech watermarking through parametric modeling," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, Sep. 2002, pp. 621–624.

[7] M. S. Seadle, J. R. Deller Jr., and A. Gurijala, "Why watermark? The copyright need for an engineering solution," in *Proc. Second ACM/IEEE Joint Conf. Digital Libraries*, Portland, OR, Jun. 2002.

[8] A. Adami, S. Kajarekar, and H. Hermansky, "A new speaker change detection method for two-speaker segmentation," in *Proc. ICASSP*, 2002.

[9] L. Lu and H. Zhang, "Speaker change detection and tracking in real-time news broadcasting analysis," in *Proc. ACM Multimedia*, Paris, France, Dec. 2002.

[10] U. Yapanel and J. H. L. Hansen, "A new perspective on feature extraction for robust in-vehicle speech recognition," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1281–1284.

[11] M. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recog. Workshop*, Chantilly, VA, 1997, pp. 97–99.

[12] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. Broadcast News Trans. Under. Workshop*, 1998.

[13] S. Dharanipragada and B. Rao, "MVDR-based feature extraction for robust speech recognition," in *ICASSP*, Salt Lake City, UT, 2001.

[14] T. Hain, S. Johnson, A. Tuerk, P. Woodland, and S. Young, "Segment generation and clustering in the HTK: Broadcast news transcription system," in *DARPA Broadcast News Workshop*, Herndon, VA, 1998.

[15] R. Huang and J. H. L. Hansen, "Unsupervised audio segmentation and classification for robust spoken document retrieval," in *Proc. IEEE ICASSP*, vol. 1, Montreal, QC, Canada, May 2004, pp. 741–744.

[16] B. Zhou and J. H. L. Hansen, "Efficient audio stream segmentation via T2 statistic based Bayesian information criterion (T2-BIC)," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, Jul. 2005.

[17] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*. San Francisco, CA: Morgan Kaufmann, 1999.

[18] T. Landauer, P. Foltz, and D. Laham, "Introduction to latent semantic analysis," *Discourse Processes*, vol. 25, pp. 259–284, 1998.

[19] "N. District Court Calif.," A&M Records, Inc. v. Napster, Inc., 99-5183, 2000.

[20] "11th Circuit Court of Appeals," Estate of Martin Luther King v. CBS, 98-9079, 1999.

[21] M. Seadle, "Copyright in the networked world: New rules for images," *Library Hi Tech.*, vol. 20, no. 2, 2002.

[22] ——, "Whose rules? Intellectual property, culture, and indigenous communities," *D-Lib Mag.*, vol. 8, no. 3, Mar. 2002.

[23] ——, "Copyright in the networked world: Multimedia fair use," *Library Hi Tech.*, vol. 19, no. 4, 2001.

[24] ——, "Spoken words, unspoken meanings: A DLI2 project ethnography," *D-Lib Mag.*, Nov. 2000.

[25] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun., Special Issue on Speech Under Stress*, vol. 20, no. 2, pp. 151–170, Nov. 1996.

[26] R. Sarikaya and J. H. L. Hansen, "High resolution speech feature parameterization for monophone based stressed speech recognition," *IEEE Signal Process. Lett.*, vol. 7, no. 7, pp. 182–185, Jul. 2000.

[27] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 429–442, Jul. 2000.

[28] L. M. Arslan and J. H. L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *J. Acoust. Soc. Amer.*, vol. 102, no. 1, pp. 28–40, Jul. 1997.

[29] P. Angkititrakul and J. H. L. Hansen, "Advances in phone-based modeling for automatic accent classification," IEEE Trans. Speech Audio Proc., to be published.

[30] J. Foote *et al.*, "Talker-Independent keyword spotting for information retrieval," in *Proc. Eurospeech*, vol. 3, 1995, pp. 2145–2149.

[31] P. C. Woodland *et al.*, "Experiments in broadcast news transcription," in *Proc. IEEE ICASSP*, Seattle, WA, 1998, pp. 909–912.

[32] . [Online]. Available: http://speechbot.research.compaq.com/

[33] . [Online]. Available: http://www.dragonsys.com/news/pr/audiomine.html

[34] B. Arons, "A system for interactively skimming recorded speech," *ACM Trans. Computer-Human Interaction*, vol. 4, no. 1, pp. 3–38, 1997.

[35] V. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 288–298, Mar. 2001.

[36] C. Hori, S. Furui, R. Malkin, Y. Hua, and A. Waibel, "Automatic speech summarization applied to English broadcast news speech," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, Orlando, FL, May 2002, pp. 9–12.

[37] S. Wegmann, Z. P. Zhan, and L. Gillick, "Progress in broadcast news transcription at dragon systems," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, Phoenix, AZ, Mar. 1999, pp. 33–36.

[38] S. S. Chen, E. M. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanevsky, and P. Olsen, "Recent improvements to IBM's speech recognition system for automatic transcription of broadcast news," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, Phoenix, AZ, 1999, pp. 37–40.

[39] S. E. Johnson, P. Jourlin, G. L. Moore, K. S. Jones, and P. C. Woodland, "The Cambridge University spoken document retrieval system," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, Phoenix, AZ, Mar. 1999, pp. 49–52.

[40] R. Hecht, J. Riedler, and G. Backfried, "German broadcast news transcription," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, Sep. 2002, pp. 1753–1756.

[41] W. Macherey and H. Ney, "Toward automatic corpus preparation for a German broadcast news transcription system," in *Proc. Int. Conf. Spoken Lang. Process.*, vol. 1, Denver, CO, May 2002, pp. 733–736.

[42] F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani, "A baseline for the transcription of Italian broadcast news," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 3, Istanbul, Turkey, Jun. 2000, pp. 1667–1670.

[43] Y.-H. Park, D.-H. Ahn, and M. Chung, "Morpheme-based lexical modeling for Korean broadcast news transcription," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1129–1132.

[44] A. Kobayashi, F. J. Och, and H. Ney, "Named entity extraction from Japanese broadcast news," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1125–1128.

[45] M. Afify and O. Siohan, "Sequential estimation with optimal forgetting for robust speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 1, pp. 19–26, Jan. 2004.

[46] L. Nguyen, X. Guo, R. Schwartz, and J. Makhoul, "Japanese broadcast news transcription," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, Sep. 2002, pp. 1749–1752.

[47] H. Nishizaki and S. Nakagawa, "Comparing isolately spoken keywords with spontaneously spoken queries for Japanese spoken document retrieval," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, Sep. 2002, pp. 1505–1508.

[48] B. Chen, H.-M. Wang, and L.-S. Lee, "Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese," *IEEE Trans. Speech Audio Proc.*, vol. 10, no. 5, pp. 303–314, Jul. 2002.

[49] J.-T. Chien and C.-H. Huang, "Bayesian learning of speech duration models," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 558–567, Nov. 2003.

[50] P.-C. Chang, S.-P. Liao, and L.-S. Lee, "Improved Chinese broadcast news transcription by language modeling with temporally consistent training corpora and iterative phrase extraction," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 421–424.

[51] Y. Liu and P. Fung, "State-dependent phonetic tied mixtures with pronunciation modeling for spontaneous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 351–364, Jul. 2004.

[52] B. Chen, H.-M. Wang, and L.-S. Lee, "Retrieval of broadcast news speech in Mandarin Chinese collected in taiwan using syllable-level statistical characteristics," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 3, Istanbul, Turkey, Jun. 2000, pp. 1771–1774.

[53] H. Meinedo and J. Neto, "Audio segmentation, classification and clustering in a broadcast news task," in *Proc. IEEE Inter. Conf. Acoust. Speech, Signal Process.*, vol. 2, Hong Kong, Apr. 2003, pp. 5–8.

[54] J. Billa, M. Noamany, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, and F. Kubala, "Audio indexing of Arabic broadcast news," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, Orlando, FL, May 2002, pp. 5–8.

[55] C. Barras, L. Lamel, and J.-L. Gauvain, "Automatic transcription of compressed broadcast audio," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, Salt Lake City, UT, May 2001, pp. 265–268.

[56] A. Fujii and K. Itou, "Building a test collection for speech-driven web retrieval," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1153–1156.

[57] W.-K. Lo, H. M. Meng, and P. C. Ching, "Multi-scale and multi-model integration for improved performance in Chinese spoken document retrieval," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, Sep. 2002, pp. 1513–1516.

[58] M. Saraclar, M. Riley, E. Bocchieri, and V. Go, "Toward automatic closed captioning :Low latency real time broadcast news transcription," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, Sep. 2002, pp. 1741–1744.

[59] S. R. Maskey and J. Hirschberg, "Automatic summarization of broadcast news using structural features," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1173–1176.

[60] C. Hori and S. Furui, "Automatic speech summarization based on word significance and linguistic likelihood," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 3, Istanbul, Turkey, Jun. 2000, pp. 1579–1582.

[61] C. Neukirchen, D. Willett, and G. Rigoll, "Experiments in topic indexing of broadcast news using neural networks," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 2, Phoenix, AZ, Mar. 1999, pp. 1093–1096.

[62] Y. Moh, P. Nguyen, and J.-C. Junqua, "Toward domain independent speaker clustering," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 2, Hong Kong, Apr. 2003, pp. 85–88.

[63] K. Mori and S. Nakagawa, "Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, Salt Lake City, UT, May 2001, pp. 413–416.

[64] F. Walls, H. Jin, S. Sista, and R. Schwartz, "Probabilistic models for topic detection and tracking," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, Phoenix, AZ, Mar. 1999, pp. 521–524.

[65] C. Langzhou, J.-L. Gauvain, L. Lamel, and G. Adda, "Unsupervised language model adaptation for broadcast news," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, Hong Kong, Apr. 2003, pp. 220–223.

[66] M. Kurimo, B. Zhou, R. Huang, and J. H. L. Hansen, "Language modeling structures in audio transcription for retrieval of historical speeches," in *Proc. 12th Eur. Signal Process. Conf.*, Vienna, Austria, Sep. 6–10, 2004, pp. 557–560.

[67] H.-M. Wang, H. Meng, P. Schone, B. Chen, and W.-K. Lo, "Multi-scale-audio indexing for translingual spoken document retrieval," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, Salt Lake City, UT, May 2001, pp. 605–608.

[68] J. Navratil, "Spoken language recognition-a step toward multilinguality in speech processing," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 678–685, Sep. 2001.

[69] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 420–435, Jul. 2004.

[70] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*. San Diego, CA: Academic, 2002.

[71] A. Gurijala and J. R. Deller Jr., "Speech watermarking by parametric embedding with an ℓ_∞ fidelity criterion," in *Proc. Interspeech/Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 2933–2936.

[72] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, Second ed. Piscataway, NJ: IEEE, 2000, ch. 5.

[73] A. Gurijala and J. R. Deller Jr., "Speech watermarking with objective fidelity and robustness criterion," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Nov. 2003.

[74] ——, "Speech watermarking through parametric modeling," , submitted for publication.

[75] S. Gollamudi, S. Nagaraj, S. Kapoor, and Y. F. Huang, "SMART: A toolbox for set-membership filtering," in *Proc. Eur. Conf. Circuit Theory Design*, Budapest, Hungary, 1997.

[76] S. Nagaraj, S. Gollamudi, S. Kapoor, and Y. F. Huang, "BEACON: An adaptive set-membership filtering technique with sparse updates," *IEEE Trans. Signal Process.*, vol. 47, no. 11, pp. 2928–2941, Nov. 1999.

[77] J. R. Deller Jr. and H. F. Huang, "Set-membership identification and filtering in signal processing," *Circuits, Syst., Signal Process., Special Issue on Signal Process. Applications*, Feb. 2002.

[78] R. Huang and J. H. L. Hansen, "High-level feature weighted GMM network for audio stream classification," in *Proc. Int. Conf. Spoken Language Process.*, Jeju Island, Korea, Oct. 2004.

[79] T. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958.

[80] S. M. Ahadi and P. C. Woodland, "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models," *Comput. Speech Language*, vol. 11, pp. 187–206, 1997.

[81] J. L. Gauvain and C. H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[82] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Language*, vol. 9, pp. 171–185, 1995.

[83] L. R. Neumeyer, A. Sankar, and V. V. Digalakis, "A comparative study of speaker adaptation techniques," in *Proc. Eurospeech*, Madrid, Spain, 1995, pp. 1127–1130.

[84] K. Shinoda and C. H. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding*, Santa Barbara, CA, 1997, pp. 381–388.

[85] O. Siohan, T. A. Myrvoll, and C. H. Lee, "Structural maximum *a posteriori* linear regression for fast HMM adaptation," *Comput. Speech Language*, vol. 16, no. 1, pp. 5–24, Jan. 2002.

[86] C. Chesta, O. Siohan, and C. H. Lee, "Maximum a posterior linear regression for hidden Markov model adaptation," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 203–206.

[87] W. Chou, "Maximum a posterior linear regression with elliptically symmetric matrix priors," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 1–4.

[88] P. C. Woodland, "Speaker adaptation: Techniques and challenges," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding*, Keystone, CO, 1999, pp. 85–90.

[89] B. Zhou and J. H. L. Hansen, "Rapid discriminative acoustic modeling based on eigenspace mapping for fast speaker adaptation," IEEE Trans. Speech Audio Process., to be published.

[90] S. E. Robertson and S. Walker, "Okapi/Keenbow at TREC-8," in *Proc. TREC-8*, 1999.

[91] S. E. Robertson and K. S. Jones, ""Simple, Proven Approaches to Text Retrieval," Tech. Rep.," Cambridge Univ., Cambridge, U.K., 1997.

[92] A. Singhal and F. Pereira, "Document expansion for speech retrieval," in *Proc. 22nd ACM SIGIR Conf.*, Berkeley, CA, Aug. 1999.

[93] M. Akbacak and J. H. L. Hansen, "Environmental sniffing: Noise knowledge estimation for robust speech systems," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 2, Hong Kong, Apr. 2003, pp. 113–116.

[94] ——, "ENVIRONMENTAL SNIFFING: Robust digit recognition for an in-vehicle environment," in *Proc. INTERSPEECH/Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 2177–2180.

[95] L. Lu, H. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech Audio Proc.*, vol. 10, no. 7, pp. 504–516, Oct. 2002.

[96] B. Zhou, "Audio Parsing and Rapid Speaker Adaptation in Speech Recognition for Spoken Document Retrieval," Ph.D. dissertation, Robust Speech Processing Group, Center for Spoken Language Research, Univ. Colorado, Boulder, CO, 2003.

[97] . [Online]. Available: http://www.ukans.edu/carrie/docs/am-docs_index.html

[98] P. Angkititrakul and J. H. L. Hansen, "Advances in phone-based modeling for automatic accent classification," IEEE Trans. Speech Audio Proc., to be published.
[99] Y. Gong, "Stochastic trajectory modeling and sentences searching for continuous speech recognition," *IEEE Trans. Speech. Audio Proc.*, vol. 5, no. 1, pp. 33–44, Jan. 1997.
[100] P. Angkititrakul and J. H. L. Hansen, "Discriminative in-set/out-of-set speaker recognition," IEEE Trans. Speech Audio Processing, submitted for publication.
[101] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in *Proc. HLT-NAACL*, Boston, MA, May 2004, pp. 129–136.
[102] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in stochastic language modeling," *Comput. Speech Language*, vol. 8, pp. 1–38, 1994.
[103] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc. Int. Conf. Spoken Language Process.*, Denver, CO, Sep. 2002, pp. 901–904.
[104] M. Kurimo, V. Turunen, and I. Ekman, "Speech transcription and spoken document retrieval in Finnish in machine learning for multimodal interaction," in *Revised Selected Papers MLMI 2004 Workshop*, vol. 3361, Lecture Notes in Computer Science, 2005, pp. 253–262.

John H. L. Hansen (S'81–M'82–SM'93) received the B.S.E.E. degree from Rutgers University, New Brunswick, N.J. in 1982, and the M.S. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1983 and 1988.

He joined the University of Texas at Dallas (UTD) in the fall of 2005, where he is Professor and Department Chairman in Electrical Engineering, and holds an Endowed Chair in Telecommunications. At UTD, he established the Center for Robust Speech Systems (CRSS), which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor with the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor with the Department of Electrical and Computer Engineering, University of Colorado, Boulder, from 1998 to 2005, where he co-founded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. He has supervised 33 Ph.D./M.S./M.A. thesis candidates. He is author/co-author of 195 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals* (New York: IEEE Press, 2000), and lead author of the report "The Impact of Speech Under "Stress" on Military Speech Technology" (NATO RTO-TR-10, 2000).

Dr. Hansen is serving as IEEE Signal Processing Society Distinguished Lecturer for 2005 and has served as Technical Advisor to the U.S. Delegate for NATO (IST/TG-01), Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 1992 to 1999, and Associate Editor for IEEE SIGNAL PROCESSING LETTERS from 1998 to 2000. He also served as guest editor of the October 1994 Special Issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He received the 2005 University of Colorado Teacher Recognition Award as voted by the student body. He also organized and served as General Chair of the International Conference on Spoken Language Processing, September 16–20, 2002.

Rongqing Huang (S'01) was born in China on November 12, 1979. He received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, in 2002 and the M.S. degree in electrical engineering from th University of Colorado, Boulder, in 2004, from which he expects to receive the Ph.D. degree in electrical engineering in 2006.

From 2000 to 2002, he was a research assistant with the iFlyTek Speech Lab, University of Science and Technology of China. From 2002 to 2005, he was with the Robust Speech Processing Group, the Center for Spoken Language Research, University of Colorado, Boulder, where he was a Ph.D. research assistant with the Department of Electrical and Computer Engineering. Currently, he a research assistant with Motorola Research Labs, Schaumburg, IL, and is also a research staff with the Human Language Technology Research Institute, University of Texas, Dallas. His research interests include speech recognition and synthesis, machine learning and data mining, and digital signal processing and communication.

Bowen Zhou (M'04) received the B.E. degree from the University of Science and Technology of China, Hefei, in 1996, the M.E. degree from the Chinese Academy of Sciences, Beijing, China, in 1999, and the Ph.D. degree from the University of Colorado, Boulder in 2003, all in electrical engineering.

He is currently a Research Staff Member with the Department of Human Language Technologies, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, which he joined in March 2003. Prior to that, he was a Research Assistant with the Center for Spoken Language Research, University of Colorado at Boulder. He was an invited speaker for IBM User Interface Technology Student Symposium in November 2002. His current research interest includes automatic speech recognition, statistical speech-to-speech translation, spoken language understanding, spoken information retrieval, and machine learning.

Dr. Zhou has actively served as a technical reviewer for the IEEE TTRANSACTIONS ON SPEECH AND PROCESSING, the IEEE SIGNAL PROCESSING LETTERS, and *Speech Communication*.

Michael Seadle received the Ph.D. degree in history from the University of Chicago, Chicago, IL, and a library degree from the University of Michigan, Ann Arbor.

He has written over 80 articles, chapters, and books on a wide range of subjects including German history, computing management, and digitization. He has over a decade of experience as a computer professional. At present, he is editor of the peer-reviewed academic journal *Library Hi Tech*. He has also served as Principal Investigator (PI) for a Library of Congress/Ameritech award for text digitization, PI for two Institute of Museum and Library Services grants, co-PI for an NSF award; and Co-PI for a Digital Library Initiative award for sound digitization. He works as Assistant Director for Information Technology at Michigan State University Libraries, East Lansing.

J. R. Deller, Jr. (SM'89–F'98) received the B.S. degree in electrical engineering, *Summa Cum Laude*, in 1974 from the Ohio State University, Columbus, and the M.S. degree in biomedical engineering in 1975, the M.S. degree in electrical and computer engineering in 1976, and the Ph.D. degree in biomedical engineering in 1979, from the University of Michigan, Ann Arbor.

He is a Professor of electrical and computer engineering at Michigan State University, East Lansing, where he directs the Speech Processing Laboratory. His research interests include statistical signal processing with application to speech processing, communications technologies, digital libraries, and biomedicine. He has co-authored two textbooks, is co-authoring a third, and has contributed chapters to numerous research volumes.

Dr. Deller received the IEEE Millennium Medal for contributions in signal processing research and education, the IEEE Signal Processing Best Paper Award in 1998, and the IEEE Signal Processing Society's 1997 Meritorious Service Award for his six-year service as Editor-in-Chief of the IEEE SIGNAL PROCESSING MAGAZINE.

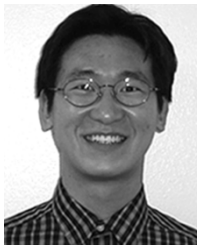
Aparna R. Gurijala received the M.S. degree in electrical engineering from Michigan State University (MSU), East Lansing, in 2001, where she is currently working toward the Ph.D. degree.

She served as a research assistant on the National Gallery of Spoken Word (NGSW) project from 2000 to 2004 and has been involved in the development of speech watermarking algorithms. Her research interests lie in the areas of multimedia security and signal processing, with a particular focus on speech watermarking and adaptive signal processing.

Ms. Gurijala is a member of the IEEE Signal Processing Society.

Mikko Kurimo received the D.Sc.(Ph.D.) degree in computer science from Helsinki University of Technology, Espoo, Finland, in 1997.

From 2001 to 2003, he was a Professor of computer science with the Laboratory of Computer and Information Science, Helsinki University of Technology. He is currently Academy Research Fellow with the Finnish Academy and Director of the speech recognition research group with the Neural Networks Research Center (Laboratory of Computer and Information Science), Helsinki University of Technology. His main research areas are speech recognition, machine learning, and spoken document retrieval.



Pongtep Angkititrakul (M'05) was born in Khonkaen, Thailand. He received the B.S.E.E. degree from Chulalongkorn University, Bangkok, Thailand, in 1996 and the M.S. degree in electrical engineering from University of Colorado, Boulder, in 1999, from which he also received the Ph.D. degree in electrical engineering.

He has been a research assistant with the Robust Speech Processing Group—Center for Spoken Language Research (CSLR), University of Colorado, since 2000. In Dececember 2004, he joined Eliza Corporation, Beverly, MA. His research interests are in the general area of automatic speech/speaker recognition, statistical signal processing, pattern recognition, data mining, and speech processing.