SPEECHFIND FOR CDP: ADVANCES IN SPOKEN DOCUMENT RETRIEVAL FOR THE U. S. COLLABORATIVE DIGITIZATION PROGRAM

Wooil Kim and John H. L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas, USA
{wikim,John.Hansen}@utdallas.edu, http://crss.utdallas.edu

ABSTRACT

This paper presents our recent advances for SpeechFind, a CRSS-UTD designed spoken document retrieval system for the U.S. based Collaborative Digitization Program (CDP[1]). A proto-type of SpeechFind for the CDP is currently serving as the search engine for 1,300 hours of CDP audio content which contain a wide range of acoustic conditions, vocabulary and period selection, and topics. In an effort to determine the amount of user corrected transcripts needed to impact automatic speech recognition (ASR) and audio search, a web-based online interface for verification of ASR-generated transcripts was developed. The procedure for enhancing the transcription performance for SpeechFind is also presented. A selection of adaptation methods for language and acoustic models are employed depending on the acoustics of the corpora under test. Experimental results on the CDP corpus demonstrate that the employed model adaptation scheme using the verified transcripts is effective in improving recognition accuracy. Through a combination of feature/acoustic model enhancement and language model selection, up to 24.8% relative improvement in ASR was obtained. The SpeechFind system, employing automatic transcript generation, online CDP transcript correction, and our transcript reliability estimator, demonstrates a comprehensive support mechanism to ensure reliable transcription and search for U.S. libraries with limited speech technology experience.

Index Terms— SpeechFind, spoken document retrieval, CDP, NGSW, transcript verification, model enhancement

1. INTRODUCTION

As available online digital collections drastically increase, the need for automatic and efficient information retrieval continues to expand, placing demands on advances in technology including computational power and storage capacity. Recently, there has been growing interest in retrieving information, especially, online for multimedia data consisting of rich information such as audio, video and speech. Today, multimedia information collections include radio/television broadcast news, interviews, entertainment content, User Generated Content (UGC), and others. This increasing demand has drawn remarkable attention to research on Spoken Document Retrieval (SDR) [1, 2, 3, 4, 5].

SpeechFind is a SDR system serving as the platform for several programs across the United States for audio indexing and retrieval including the National Gallery of the Spoken Word (NGSW) and the Collaborative Digitization Program (CDP) [1, 6, 7]. The system consists of two main phases; (i) enrollment and (ii) online search retrieval. Our recent work on SpeechFind has included an effort to improve performance by addressing band-limited speech among a wide range of acoustic conditions [8].

This paper provides an overview of recent advances in the SpeechFind system and collaboration with the CDP. A prototype of SpeechFind for the CDP has been established to serve as the search engine for the CDP corpus which presently contains 1,300 hours of audio documents. An online system for verification of the ASR-generated transcripts has been developed to improve the speech recognition engine and evaluate overall transcript generation performance. To provide more reliable retrieval results for SpeechFind, we also developed a transcript reliability estimator for supplementary information for the user [9]. We also focus on our transcription improvement scheme which consists of speech/feature enhancement, language model selection, and acoustic model adaptation. Two different types of adaptation approaches (i.e., document-dependent and document-across) are employed based on the selective training set for each test utterance.

We review the SpeechFind system and recent collaboration with the CDP in Sec.2-3. In Sec.3, we discuss the structure of the audio materials from the CDP corpus. Sec.4 presents development of the transcript verification process including online web-interface. The transcription enhancement schemes are described in Sec.5. Representative experimental proce-

This work was funded by grants from RADC (A40104), the CDP, and University of Texas at Dallas under Project EMMITT.

[1]The CDP-Collaborative Digitization Program is a consortium of 29 U. S. libraries, museum, and archives brought together to establish best practices for presentation and access of historical audio materials from across the United States (http://cdpheritage.org).

dures and their results are presented and discussed in Sec.6. Finally, in Sec.7, we summarize and provide conclusions.

2. OVERVIEW OF SPEECHFIND

SpeechFind is a spoken document retrieval system developed to serve as the search engine for the National Gallery of the Spoken Word (NGSW) [1, 6]. The system includes the following modules: i) an audio spider and transcoder, ii) spoken documents transcriber, iii) transcription database, and iv) an online public accessible search engine. The audio spider and transcoder are responsible for automatically fetching available audio archives from a range of available servers and converting the incoming audio files into the designed audio formats for processing. This module also parses the metadata and extracts relevant information into a "rich" transcript database to guide future information retrieval.

The spoken document transcriber includes an 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 using a large vocabulary continuous speech recognition (LVCSR) engine.

The online search engine 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 the relevant transcribed documents that are ranked by relevance scores and associated with timing information, and provides the user with web links to access the corresponding audio clips.

The SpeechFind system is also currently serving as the search engine for the CDP audio corpus, which has been established via a collaboration between CRSS and the CDP program. Fig.1 shows the main page of SpeechFind specialized for the CDP corpus.

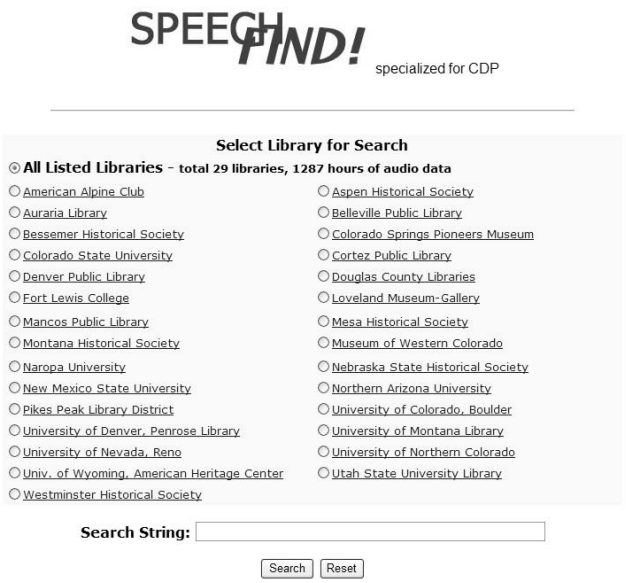

Fig. 1. *Main page of SpeechFind for the CDP corpus.* `http://SpeechFind.utdallas.edu/index_cdp.html`

3. STRUCTURE OF CDP AUDIO CORPUS

In this section, we discuss the structure of the CDP corpus. From the available limited metadata, it is known that the CDP audio files include interviews, discussions/debates, and lectures, each with 2-5 speakers participants. The recorded audio documents are spontaneously articulated with many overlapping speakers, and burst noise events such as clapping, laughing, etc. which make speech recognition challenging. The content of the speeches include speakers' personal experience and opinions on social issues such as Word War II, the Red Cross, civil rights, feminist activity, and other topics. The speakers are reported to be leaders in local communities including senators, professors, activity group leaders, etc. Recordings were conducted from the 1960s to 2000s and held at library offices, classrooms, homes, etc. Depending on the documents, there exists background noise which would occur due to recording media or transmission.

The audio corpus from a total 29 participants (libraries, societies, museums, etc.) are currently available on SpeechFind for search and retrieval, which contain approximately 1,300 hours and 1.2 TB of data as shown in Table 1. Speech data were automatically transcribed by our speech recognition engine for online document retrieval. Here, 5% of the total ASR-generated transcripts were verified by CDP participants via an online correction phase for performance improvement and evaluation. Table 2 shows details on the CDP corpus which has been verified for evaluation. Although verified transcripts make up about 5% of the entire CDP corpus, they are expected to represent the characteristics of the entire corpus because they were evenly selected from across each document. Perplexity, entropy and OOV (Out-Of-Vocabulary) in Tables reported here were obtained using CMU-Cambridge Statistical Language Modeling Toolkit [10]. OOV rates were calculated based on Broadcast News vocabulary consisting of 64K words which is employed for the current LVCSR engine. The total number of OOVs are 6,487 which mostly include name entities and some amount of miss-prints by proofreaders. The detected OOVs are used for updating the acoustic model and language model. We also plan to identify and correct the miss-prints in the transcripts to enhance the transcripts for model adaptation. STNRs were calculated using NIST Speech Quality Assurance software [11] and Fig. 2 shows the distribution of STNR on the CDP corpus from all libraries.

Table 1. Entire CDP corpus and verified parts for evaluation.

Entire Corpus	29 participants 1,286.5 hours (1.2 TB)
Verified Parts	70.6 hours 18,651 audio segments 5.5 % of entire

Table 2. Details on CDP corpus; verified parts for evaluation.

Total number of words	512,435
Total vocabulary size	20,003
OOV rate	2.34 %
Average STNR	23.65 dB
Perplexity	18.98
Entropy	4.23

4. TRANSCRIPT VERIFICATION PROCESS WITH CDP

We recently established a proto-type of the transcript verification process with CDP. An online web-interface was developed in order to improve the quality of the ASR-generated transcripts. The transcript verification process is as follows:

(1) Automatic Transcription: the audio documents delivered by CDP participants are automatically transcribed via our speech recognition engine. The original audio data with format of stereo, 44.1kHz, 24bit PCM are converted into single channel, 16kHz, 16bit PCM for processing. Every audio document with a length of approximately 15-40 min is segmented into small segments (15-30 sec) using our developed segmentation algorithm [12, 13]. Each segment is automatically transcribed by large vocabulary continuous speech recognition (LVCSR) engine currently employing SPHINX3.

(2) Online Verification: from each audio document, approximately 5% of the segments are selected in an approximate uniform manner across each file. The transcripts of the selected segments were uploaded to the online system where participants would log-in using their accounts for verification work. They would listen to audio clips and correct the uploaded transcripts via the online web-interface. The words newly appearing in the verified transcripts which are out of vocabulary employed by the ASR are automatically detected and stored in separate files for future processing. Several types of transcription conventions are allowed when correcting transcripts such as *(unknown)*, *(noise)*, *(clapping)* and *(laughter)*. Fig. 3 shows the online web-interface for CDP user transcript verification.

(3) Model Enhancement: the corrected transcripts are used for performance evaluation and model enhancement to improve the transcription generation. Model enhancement has been applied to a sub-set of the participants in the current system to assess performance, and will be applied to the entire audio corpus in the near future. Details on our model enhancement scheme using the verified transcripts will be presented in Sec.5. As shown in Table 1, 70.6 hours of speech segments have been verified via the online process, which is about 5% of the entire corpus.

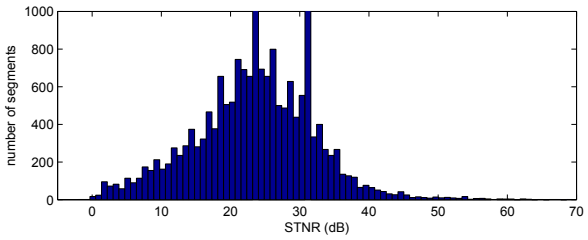

Fig. 2. *STNR histogram of CDP corpus.*

Fig. 3. *Online web-interface for transcript verification.*

5. TRANSCRIPT IMPROVEMENT VIA FEATURE/MODEL ENHANCEMENT

In order to obtain more reliable transcription of spoken documents for the SpeechFind system, we employ three different levels of enhancement schemes: 1) speech/feature, 2) linguistic information, and 3) acoustic model. Fig.4 shows illustration of our entire enhancement scheme to improve transcription performance in this study.

5.1. Speech/Feature Enhancement

The audio documents for SDR are likely to contain additive noise and channel distortions, which would come from back-

ground noise, band restriction, channel interference, and others due to recording media and conversion/transmission. In our previous study, we have proposed the effective reconstruction algorithms for band restricted speech [8, 14]. A large number of speech/feature enhancement algorithms including our studies can be selectively employed depending on given utterances and noise types. Enhanced speech and features have improved intelligibility, so they result in more effective acoustic model adaptation.

5.2. Lexicon Update and Language Model Adaptation

Our baseline LVCSR engine for SpeechFind employs a lexicon (64K words) and language model obtained using the Broadcast News Corpus. Table 2 shows that the verified CDP corpus (5.5% of the entire corpus) contains a 2.34% OOV rate, which is expected to increase for the entire corpus. The lexicon for SpeechFind is updated using the OOVs appeared in the verified transcripts. As described in Sec.3, the CDP corpus contains a different audio structure in terms of speech types and vocabulary/period selections. Therefore, the language model also needs to be selectively updated or adapted depending on the audio document structure.

5.3. Acoustic Model Adaptation Using Selective Training Set

In our framework, the acoustic model adaptation has two types of targets: 1) document-dependent acoustic conditions and 2) document-across acoustic conditions. The document-dependent acoustics include the speaker dependent characteristics, time-varying/short-term background noise and channel interference, and others, which occur particularly in a given test utterance or adjacent audio segments. The document-across acoustics contain gender/age/accent dependent speech traits and the background noise/channel distortions observed broadly across the other audio documents.

For these two types of acoustic model adaptation, the training utterance sets are selected from the training database pool and constructed depending on a given test utterance. In our study, we use the Kullback-Leibler distance as a similarity metric between the given utterance and the training utterances. For document-dependent acoustic model adaptation, a small number of utterances (e.g., top five of the most similar utterances) are selected to represent more document specific acoustic condition and MLLR (Maximum Likelihood Linear Regression) adaptation is employed, which is known to be robust on the small sized adaptation database. A larger number of training utterances (e.g., totally 20-30 minutes) are selected for document-across acoustic conditions and MAP (Maximum *A Posteriori*) adaptation scheme is applied. To increase the adaptation effectiveness, unsupervised MLLR adaptation is also applied to each test utterance. Our preliminary experimental results on feature/model enhancement presented here using the verified transcripts will be discussed in the following Sec.6.

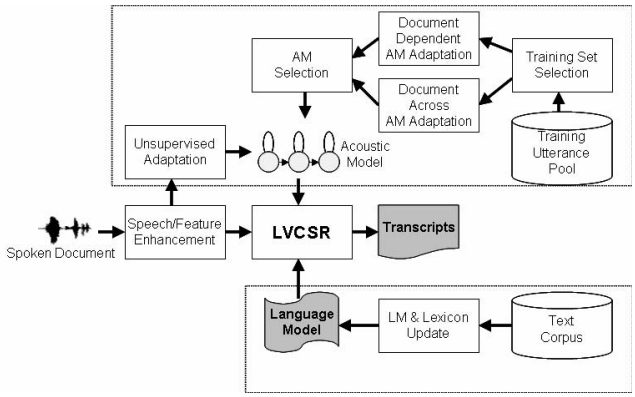

Fig. 4. *Feature/model enhancement scheme for transcription improvement.*

6. EXPERIMENTAL RESULTS

In this section, preliminary experimental results of our efforts to improve the transcription performance for SpeechFind utilizing the verified transcripts which were presented in Sec.5. To evaluate the performance, the database for test and adaptation were selected among the CDP audio documents which have the verified transcripts. Table 3 shows the configuration of the database used for acoustic model enhancement.

From each audio document, approximately 5% of the segments are selected as adaptation data with a uniform location distribution across the audio stream. The remaining segments are used for recognition testing. In our experiment, a single audio document per each library was selected for performance evaluation (i.e., a total of 532 segments and 2.0 hours). The training database pool for adaptation is constructed with all selected adaptation utterances (i.e., 5% from each document) resulting in a total of 3,216 segments and 12.3 hours. The selected set corresponding to each test utterance for the acoustic model adaptation is determined from the constructed utterance pool.

Table 4 shows the performance of baseline, spectral subtraction (SS) and MLLR adaptation. Spectral subtraction method is employed as a speech enhancement component among our transcription enhancement schemes in Fig.4. MLLR adaptation is applied to a given test utterance in an unsupervised manner where an initial decoding needs to precede using the baseline acoustic model. By applying spectral subtraction, relatively low SNR audio documents (i.e., DCL and UNC) show improvement in WER. We found that the test audio document from DCL and UNC include relatively high energy background noise which is considered due to the recording media. However, in cases of AL and MESA libraries which have audio that is relatively high in SNR and low WER as

Table 3. *Database set used for performance evaluation of transcription enhancement.*

Library	Test		Adaptation	
	# seg.	hour	# seg.	hour
AL	55	0.26	336	1.6
DCL	43	0.16	394	1.5
MESA	113	0.40	498	1.8
PPLD	109	0.35	977	3.7
UDPL	106	0.39	917	3.4
UNC	104	0.45	64	0.3
Total	**532**	**2.0**	**3,216**	**12.3**

Table 4. *Performance of transcription enhancement (WER, %).*

Library	STNR (dB)	Baseline	SS	Unsuper. MLLR
AL	39.0	41.1	44.6	38.9
DCL	18.6	74.9	71.6	69.8
MESA	22.3	51.9	55.7	49.8
PPLD	22.5	75.4	74.8	74.5
UDPL	20.8	59.1	57.8	56.2
UNC	17.2	75.1	68.8	71.0
Avg.	**23.4**	**63.0**	**62.0**	**60.1**

baseline, the performance decreases by applying spectral subtraction. Spectral subtraction used in our study employs a noise estimation algorithm based on minimum statistics [15], which is known to be robust to slowly changing background noise. Failure to correctly estimate the burst noise such as laughing, clapping, etc. would result in degraded recognition performance. Unsupervised MLLR shows consistent improvement for all audio documents.

The experimental results in Table 5 present the performance of the enhancement schemes for language and acoustic models. A separate language model (CDP-LM) was constructed using the verified CDP transcripts which has 70.6 hours of data made up of 512K words and a 20K-word vocabulary. In our framework, the language and acoustic model selection is conducted based on the scores of pilot utterances of each test document. In other words, each test document has a small representative set consisting of 10-12 segments and the models are selected based on scores of the pilot sets. As for language model selection, DCL, PPLD, and UNC show better performance when employing CDP-LM and the other library collections still perform better with the baseline Broadcast News language model.

The third column (AM Adapt) shows the performance of acoustic model adaptation using the selective training set. In this experiment, document-dependent acoustic model (DD-AM) adaptation is applied to AL, MESA, PPLD, and UDPL, where the five most similar training utterances are selected

Table 5. *Performance of transcription improvement via model enhancement (WER, %).*

Library	LM Select	AM Adapt	**Comb.**	**Relative Impr.(%)**
AL	41.1	41.3	**38.9**	**5.4**
DCL	71.3	61.1	**56.3**	**24.8**
MESA	51.9	49.6	**49.3**	**5.0**
PPLD	71.2	72.7	**67.6**	**10.3**
UDPL	59.1	57.5	**56.3**	**4.7**
UNC	75.4	71.9	**66.8**	**11.5**
Avg.	**62.0**	**59.8**	**56.6**	**10.2**

to apply MLLR adaptation for each test utterance. The selected training set for DD-AM adaptation mostly includes the adjacent utterances segmented from the same document and are considered to represent the document-dependent acoustic conditions such as speaker characteristics and existing noisy conditions in a given test utterance. DCL and UNC show better performance with the case of document-across acoustic model (DA-AM) adaptation, where we select 100 similar training utterance for MAP adaptation. It was found that the background noise and channel distortions in the test documents of DCL and UNC are also observed across the other documents of the same library, which are considered due to the recording/transmission media/environment.

The last two columns in Table 5 show the combination of all enhancement components employed in our study. The experimental results show relatively high improvement for DCL, PPLD and UNC which have relatively low performance in baseline WER. For the case of DCL, a 24.8% relative improvement in WER is seen as significant. By employing our transcription enhancement scheme, we obtained a 10.2% average relative improvement, which shows our enhancement scheme is effective in improving the transcription performance.

We should note that the process for obtaining 1,300 hours of transcripts was achieved with very limited human, computer, and transcription overhead, offering an effective means for libraries and archives to obtain audio search support with low technology expertise.

7. CONCLUSIONS

In this paper, we presented our recent advances in SpeechFind and our collaboration with the CDP. A proto-type of SpeechFind for the CDP was established serving as the search engine for about 1,300 hours of the CDP audio content. The web-based online interface for verification of the ASR-generated transcripts has been developed for use in improving and evaluating the speech recognition engine. We developed the transcription improvement scheme consisting of feature/speech enhancement and language/acoustic model enhancement. Experimental results on the CDP corpus demonstrate that the

model enhancement schemes using the verified transcripts is effective in improving recognition accuracy. Through combining feature/model enhancement schemes, up to 24.8% relative improvement was obtained on relatively low SNR audio documents. The framework and results here help suggest an effective process to provide transcription support to libraries with limited ASR/search expertise.

8. REFERENCES

[1] Hansen, J. H. L., Huang, R., Zhou, B., Seadle, M., Deller, J. R. Jr., Gurijala, A. R., Kurimo, M., Angkititrakul, P., "SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word," *IEEE Trans. SAP*, 13(5):712-730, 2005.

[2] Wegmann, S., Zhan, P., and Gillick, L., "Progress in Broadcast News Transcription at Dragon Systems," *ICASSP-99*, pp.33-36, March 1999.

[3] Chen, S.S., Eide, E.M., Gales, M.J.F., Gopinath, R.A., Kanevsky, D., and Olsen, P., "Recent Improvements To IBM's Speech Recognition System For Automatic Transcription Of Broadcast News," *ICASSP-99*, pp.37-40, March 1999.

[4] Johnson, S.E., Jourlin, P., Moore, G.L., Jones, K.S., and Woodland, P.C., "The Cambridge University Spoken Document Retrieval System," *ICASSP-99*, pp.49-52, March 1999.

[5] Ramabhadran, B., Huang, J., and Picheny, M., "Towards Automatic Transcription of Large Spoken Archives - English ASR for the MALACH Project," *ICASSP2003*, 2003.

[6] http://www.ngsw.org.

[7] http://cdpheritage.org.

[8] Kim, W., Hansen, J. H. L., "Missing-Feature Reconstruction for Band-Limited Speech Recognition in Spoken Document Retrieval," *Interspeech2006*, pp.2306-2309, Sept. 2006.

[9] Kim, W., Hansen, J. H. L., "Advances in SpeechFind: Transcript Reliability Estimation Employing Confidence Measure based on Discriminative Sub-word Model for SDR," *accepted for Interspeech2007*, 2007.

[10] http://www.speech.cs.cmu.edu/SLM/toolkit.html

[11] NIST SPeech Quality Assurance (SPQA) package version 2.3, http://www.nist.gov/speech.

[12] Zhou, B., Hansen, J. H. L., "Efficient Audio Stream Segmentation via T2 Statistics Based Bayesian Information Criterion (T2-BIC)," *IEEE Trans. on SAP*, 13(4), 2005.

[13] Huang, R., Hansen, J. H. L., "Advances in Unsupervised Audio Classification and Segmentation for the Broadcast News and NGSW corpora," *IEEE Trans. on ASLP*, 14(3):907-919, 2006.

[14] Morales, N., Hansen, J. H. L., "Blind Feature Compensation for Time-Variant Band-Limited Speech Recognition," *IEEE Signal Proc. Letters*, 14(1):70-73, 2007.

[15] Martin, R., "Spectral Subtraction Based on Minimum Statistics," *EUSIPCO-94*, pp.1182-1185, 1994.