



Advances in SpeechFind: Transcript Reliability Estimation Employing Confidence Measure based on Discriminative Sub-word Model for SDR

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 study presents our recent advances in our spoken document retrieval (SDR) system SpeechFind including our partnership with the Collaborative Digitization Program (CDP). A proto-type of SpeechFind for the CDP is currently serving as the search engine for 1,300 hours of the CDP audio content. These audio corpus of spoken document possess a wide range of conditions which make speech recognition challenging for reliable transcripts. In this paper, a reliability estimation method for the ASR-generated transcripts is proposed to provide more effective retrieval information for SpeechFind. The proposed estimator is based on Bayesian classification employing several confidence measures. We also propose a novel confidence measure for reliability estimation employing acoustically discriminative sub-word models. Experimental results on CDP material demonstrate that the proposed confidence measure is effective in improving the reliability estimator. By employing the proposed confidence measure based on discriminative model, 10.5% and 20.9% relative improvements were obtained in accuracy and critical error respectively.

Index Terms: SpeechFind, spoken document retrieval, reliability estimation, confidence measure, discriminative model.

1. Introduction

As available audio information collections drastically increase, the need for automatic and efficient information retrieval continues to increase. These advances in speech recognition and statistical information retrieval technology must include computational power and storage capacity. Recently, there has been growing interest in retrieving information, especially online for multimedia data consisting of broadcast news, entertainment, User Generated Content (UGC), video and other sources. The increasing demand has drawn remarkable attention to research on Spoken Document Retrieval (SDR).

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, 2, 3]. Audio collections such as NGSW and CDP corpora employing a SDR system generally includes a diverse range of conditions (e.g., background noise, channel distortion, recording media, speaking styles, accents, etc.) which make speech recognition for SDR intensely challenging.

This paper proposes to employ reliability estimation of ASR-generated transcripts in order to provide more reliable retrieval for the user in SDR. The obtained reliability degrees of

the searched transcripts will provide more useful guidance for the user, who generally is not technology oriented. The proposed reliability estimation method employs several confidence measures which are generally used for applications in speech recognition such as utterance verification and out-of-vocabulary rejection [4, 5, 6]. We also propose a novel confidence measure based on acoustic similarity for reliability estimation. The proposed confidence measure employs acoustically discriminative sub-word models to estimate the intelligibility of speech and eliminates the use of Viterbi decoding required by existing confidence measures.

We review the SpeechFind system and recent advances in SpeechFind in Sec.2 and 3. In Sec.3, we also discuss the CDP corpus which is used in our experiments. Sec.4 presents the confidence measures considered for reliability estimation and a confidence measure based on acoustic similarity is proposed in Sec.5. The proposed transcript reliability estimation method is presented in Sec.6. Representative experimental procedures and their results are presented and discussed in Sec.7. Finally, in Sec.8, we summarize and present conclusions.

2. Overview of SpeechFind

The SpeechFind [1] system includes the following modules: i) an audio spider and transcoder, ii) spoken document 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 depending on a large vocabulary continuous speech recognition 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.

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

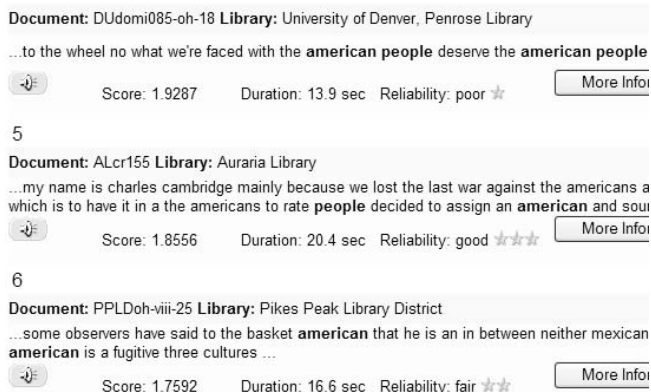


Figure 1: Example of retrieved audio segments with transcript reliability: http://SpeechFind.utdallas.edu/cdp_proto.html.

3. Advances in SpeechFind & CDP Corpus

SpeechFind system has a proto-type currently serving as the search engine for the CDP audio corpus, which has been established via a collaboration between CRSS and CDP program (http://SpeechFind.utdallas.edu/index_cdp.html). The audio corpus include a total of 29 participants (libraries, societies, museums, etc.), currently available on SpeechFind for search and retrieval, totaling 1,300 hours and 150 GB of data. We also recently established a proto-type of the human verification process with CDP to improve the quality of the ASR-generated transcripts, which is conducted as an online web-based system. The speech database with the verified transcripts are used for performance evaluation and improving the SpeechFind system via model adaptation. The experiments in this paper also use the CDP corpus obtained by the transcript verification collaboration.

It was found that the audio files from CDP primarily include interviews, discussions/debates, and lectures, with 2-5 speakers within each audio stream. The speeches recorded are spontaneously articulated and there is much overlap of speakers, and burst noises such as clapping, laughing, etc. The audio content includes talks on personal life/experience and opinions on social issues such as Word War II, Red Cross, feminist activity, and other topics from local politics and history in the U.S. The speakers are reported to be leaders in local communities such as senators, president of university, professors, activity group leaders, etc. Recordings were conducted from the 1960s to 2000s and held at library offices, classrooms, homes. Depending on the documents, there exists background noise which would occur due to recording media or transmission.

Another proto-type version of the SpeechFind system has been developed employing transcript reliability estimation to provide more reliable retrieval. Fig. 1 shows an example display of the proto-type of SpeechFind where the proposed reliability estimation is applied. In the example, each retrieved audio segment appears with the reliability degree of its transcript (e.g., *good*, *fair*, *poor*, or *N/A*). Users would consider the transcripts with higher reliability degree as more reliable information. The motivation for transcript reliability estimation is that the users are not experienced users of speech recognition technology, and therefore their confidence in the system becomes eroded if they assume all transcripts returned are would error-free. Details on the proposed reliability estimation will be discussed in the following sections.

4. Confidence Measures for Reliability Estimation

To provide the users of spoken document retrieval system with more reliable results to their query, it is required to generate more reliable transcripts using the speech recognition. However, speech recognition performance significantly degrades when the operating condition is mismatched to the training condition. The audio corpus for this SDR system contains a diverse range of acoustic conditions such as background noise, recording media, speech styles, etc. In this study, we propose to employ reliability estimation for ASR-generated transcripts to provide more reliable information to the user. In this section, the confidence measures we consider for reliability estimation will be presented.

4.1. Signal-to-Noise Ratio: snr

SNR is generally considered to be inversely proportional to WER which represents reliability for speech recognition. SNR was calculated using the NIST Speech Quality Assurance software [7].

4.2. Acoustic and Language Model Scores: all.ph, lang

The scores obtained via the acoustic model such as GMM/HMM or language model are commonly used as a confidence measure [6]. Higher values result when the corresponding trained model matches the input test data conditions. In our work, we employed an all-phone model for the acoustic model score which consists of 44 context-independent (CI) phone and silence HMMs being fully connected to each other.

4.3. Active Word Count: awc

The number of all hypothesized words at a particular time represents the spectral stability of the speech. We count the number of hypothesized words appearing in the N-best lists at the end of each word in the best word sequence as a measure of confidence.

4.4. Word Density based Confidence Measure: wd.cm

Word density is obtained by calculating the probability of each hypothesized word i in the N-best word sequence as follows [5],

$$wd.cm_i = \frac{\sum_{r \in \{w_i|H\}} P(W_r|X)}{\sum_{l=1}^N P(W_l|X)} \quad (1)$$

where W_r indicates the r th hypothesis in the N-best list and $\{w_i|H\}$ denotes a set of the hypotheses which contain word w_i . H indicates the N-best alignment obtained from Viterbi decoding. We use the average value of the obtained densities of the best hypothesized words at the end of each word in the best word sequence.

4.5. Anti-Model Confidence Measures: anti.uni, anti.comp

The ratio of the score from the hypothesized acoustic model and anti-model is generally used as a confidence measure for recognition results. Here, we investigate the performance of the confidence measures employing anti-models which are generated using CI-phone. The anti-model based confidence measure is calculated using the log-likelihood ratio of a recognized model λ_n and a corresponding anti-model λ_n^{-1} as follows,

$$anti.cm = \frac{1}{N} \sum_{n=1}^N \log \frac{p(X_n|\lambda_n)}{p(X_n|\lambda_n^{-1})} \quad (2)$$

where N is the total number of recognized models considered for calculating confidence measure.

We consider two approaches to construct the anti-model score for the confidence measure: (i) union score, and (ii) max score. For the union score, we test each phone model, and compare that score against the avg. of the scores for all other models. This union approach is written as follows, where we assess the difference between the present model and avg. union score,

$$\lambda_i^{-1} = \bigcup_{j \neq i} \lambda_j. \quad (3)$$

The second confidence measure approach is to again use the output Viterbi decoded score, and find the anti-model score as the max of the remaining models. This is written as follows, where in effect the confidence measure is assessed as the difference between the top two models,

$$\lambda_i^{-1} = \arg \max_{j \neq i} P(X|\lambda_j). \quad (4)$$

The confidence measure performance for two anti-model approaches are presented in Sec.7.

5. Confidence Measure based on Acoustically Discriminative Sub-word Model: dis.cm

Next, a novel confidence measure is proposed based on acoustically similar and dissimilar sub-word models to the incoming speech. If the incoming speech is clearly articulated without background noise, the likelihood difference of the similar model to input speech and dissimilar one obviously becomes larger. Basically the corrupted speech due to noise or obscurely pronounced fails to discriminate the two models. Based on this motivation utilizing the acoustical dissimilarity, the similar sub-word models to the input speech should be correctly determined. However, the speech recognizer for SDR must address a wide range of acoustic conditions and often fails to generate the sub-word sequences which need to be acoustically similar to the aligned speech segments.

The proposed method does not rely on the sub-word sequence generated by Viterbi decoding which conventional confidence measures employ. First, acoustic similarities between a particular sub-word model and the other remaining models are identified depending on training data as follows,

$$P(\mathbf{X}^{\{i\}}|\lambda_{i,1}) \geq P(\mathbf{X}^{\{i\}}|\lambda_{i,2}) \geq \dots \geq P(\mathbf{X}^{\{i\}}|\lambda_{i,M}) \quad (5)$$

where $\mathbf{X}^{\{i\}}$ is a collection of training data labeled as model λ_i and $\lambda_{i,m}$ indicates the m th similar model among M sub-word models to the pivotal model λ_i . In most cases, the most similar model $\lambda_{i,1}$ is identical to the model itself λ_i . Table 1 shows a part of the acoustic similarity relationship among CI-phones identified for our work. The identified models are grouped into *similar group* λ_i^{sim} and *dissimilar group* λ_i^{dis} according to the similarity to a particular model λ_i .

The best *similar group* to the incoming speech X_t is found based on a maximum likelihood decision as follows,

$$i_{\max} = \arg \max_i P(X_t|\lambda_i^{sim}) = \arg \max_i \frac{1}{N_s} \sum_{n=1}^{N_s} P(X_t|\lambda_{i,n}) \quad (6)$$

where N_s indicates the number of models in the *similar group* λ_i^{sim} . The proposed discriminative model based confidence measure at time t is calculated using log-likelihood ratio of *similar group* and *dissimilar group* as follows,

Table 1: Acoustic similarity between CI-phones.

λ_i	\Leftarrow similar				$dissimilar \Rightarrow$		
	$\lambda_{i,1}$	$\lambda_{i,2}$	$\lambda_{i,3}$	\dots	$\lambda_{i,43}$	$\lambda_{i,44}$	$\lambda_{i,45}$
AA	AA	AW	AO	\dots	T	TS	SIL
AE	AE	AW	AY	\dots	CH	TS	SIL
\vdots	\vdots	\vdots	\vdots	\dots	\vdots	\vdots	\vdots
P	P	F	V	\dots	SH	TS	SIL
R	R	ER	EH	\dots	CH	TS	SIL
S	S	Z	T	\dots	ng	AW	SIL
\vdots	\vdots	\vdots	\vdots	\dots	\vdots	\vdots	\vdots
ng	ng	axn	OO	\dots	T	S	SIL
xl	xl	OW	AO	\dots	S	T	SIL

$$dis.cm_t = \log \frac{P(X_t|\lambda_{i_{\max}}^{sim})}{P(X_t|\lambda_{i_{\max}}^{dis})} \quad (7)$$

where $P(X_t|\lambda_{i_{\max}}^{dis}) = \frac{1}{N_d} \sum_{n=1}^{N_d} P(X_t|\lambda_{i_{\max},M-N_d+n})$ and N_d indicates the number of elements in the *dissimilar group* $\lambda_{i_{\max}}^{dis}$. Here, $dis.cm_t$ is obtained every δ frame, so that scores are accumulatively calculated for $X_t = [x_{t-\delta+1}, x_{t-\delta+2}, \dots, x_t]$ and the average value of $dis.cm_t$ over an utterance is obtained as the confidence measure.

In the proposed method, the closest group of sub-word models is determined over the input speech duration and then the likelihood ratio to the acoustically farthest model group is obtained for use of confidence measure. The obtained score ratio represents the intelligibility of the utterance, since the selected *similar group* and *dissimilar group* are acoustically discriminative to each other. The proposed method utilizes the acoustically dissimilar model to obtain the likelihood difference of the determined model based on input speech, while conventional approaches consider the hypothesized model generated by Viterbi decoding.

6. Reliability Estimation based on Bayesian Classification

Reliability estimation is designed employing the confidence measures presented in Sec.4 and 5. The estimator is based on Bayesian classification consisting of a Gaussian mixture model and prior information. Several combinations of the presented confidence measures are employed as feature vectors and compared for performance. We classify the ASR generated-transcripts into three categories (e.g., *good*, *fair* and *poor*) according to word error rate (WER). The model parameters for each class are obtained via training data which are labeled as three different groups by WER.

To evaluate the reliability estimator, we propose an evaluation criterion namely *critical error rate* as follows,

$$100 \times \frac{N_{poor|good} + N_{good|fair} + N_{good|poor} + N_{fair|poor}}{N_{total}} \quad (8)$$

where $N_{A|B}$ indicates the number of set $\{decided\ as\ A|input \in B\}$ and N_{total} is the number of total inputs. The proposed criterion represents a combination of false alarms ($N_{good|fair}$, $N_{good|poor}$, and $N_{fair|poor}$) and significant miss ($N_{poor|good}$) which would critically mislead users who are depending on the retrieved transcripts. The proposed criterion is used together with decision accuracy for performance evaluation of the reliability estimator. The designed reliability estimator is used to classify actual transcripts obtained by ASR for SDR. The classified transcripts will be helpful for users to search and retrieve audio clips on SpeechFind, because they will now have a sep-

Table 2: Correlation between confidence measures and WER.

snr	all.ph	lang	awc
-0.216	0.200	0.211	0.335
wd.cm	anti.uni	anti.comp	dis.cm
-0.339	-0.311	-0.383	-0.377

arate measurement of the transcription performance which improves the user’s acceptance of the search engine.

7. Experimental Results

The confidence measures considered in our experiments are:

- **snr**: signal-to-noise ratio
- **all.ph**: all-phone model: fully connected CI-phone HMMs
- **lang**: language model score
- **awc**: active word count
- **wd.cm**: word density based confidence measure
- **anti.uni**: union anti-model confidence measure
- **anti.comp**: competitive anti-model confidence measure
- **dis.cm**: proposed confidence measure based on discriminative sub-word models.

We first evaluate the correlation relationships between the presented confidence measures and WER. Table 2 shows the resulting correlation coefficients. From the results, the anti-model-based confidence measure employing the competitive phone model (*anti.comp*) shows the highest correlation to WER. The proposed discriminative model based confidence measure (*dis.cm*) also shows the comparable correlation degree. The proposed confidence measure no longer requires a decoding process compared to anti-model based measures. When compared to the score of all-phone model (*all.ph*) which has similar number of computational operations, the proposed measure has significant advantage.

Employing the presented confidence measures, the proposed transcript reliability estimation method was evaluated on the CDP database. Table 3 shows the database used for the performance evaluation of the proposed reliability estimator. The acoustic model for each class (*good*, *fair* and *poor*) consists of 8-component Gaussian mixture with diagonal covariance. The prior probability of each model has the same value 1/3.

Table 4 shows the performance of the proposed reliability classifiers employing several combinations of confidence measures as the feature vector. Our baseline system employs five confidence measures as feature vector which contains *snr*, *all.ph*, *lang*, *awc* and *wd.cm*. From the results, by adding the anti-model confidence measures (*anti.uni* and *anti.comp*) and the proposed discriminative model based confidence measure (*dis.cm*) to baseline, improved performance was obtained. Considering both accuracy and critical error, adding *anti.comp* to baseline showed the best performance, which is consistent with the degree of correlation from Table 2.

From the results, we can see that replacing *all.ph* score with the proposed discriminative model confidence measure *dis.cm* results in better performance rather than employing them together. This is because the proposed *dis.cm* has the lowest correlation to *all.ph* score among all confidence measures while it shows remarkably higher correlation with WER compared to *all.ph* score. Replacing *all.ph* score with *dis.cm* considerably increased the performance in all cases compared to baseline. By adding anti-model based confidence measures (*anti.uni* or *anti.comp*), consistent improvement was also obtained. Including *anti.comp* showed significant improvement in both accu-

Table 3: Database for evaluation of reliability estimator.

Class	Criteria	Train		Test	
		# seg.	min.	# seg.	min.
<i>good</i>	WER \leq 45%	706	163	107	14
<i>fair</i>	45% < WER \leq 75%	790	182	193	25
<i>poor</i>	WER \geq 75%	924	213	137	18
Total		2,420	557	437	57

Table 4: Performance of transcription reliability estimator (%).

Baseline	Accuracy	Critical Error
snr+all.ph+lang+awc+wd.cm	48.74	18.76
+anti.uni	52.17	16.02
+anti.comp	54.46	16.48
+dis.cm	50.11	16.25
Replacing all.ph with dis.cm	Accuracy	Critical Error
snr+lang+awc+wd.cm+dis.cm	52.40	16.48
+anti.uni	55.38	13.73
+anti.comp	60.18(10.5%)	13.04(20.9%)

racy and critical error with 10.5% and 20.9% respectively in relative improvements compared to *baseline+anti.comp*. The results demonstrate that the proposed confidence measure, motivated from acoustic dissimilarity, is significantly effective in improving performance of the reliability estimator by employing acoustically discriminative models to identify intelligibility of the utterance.

8. Conclusions

In this study, we presented our recent advances in SpeechFind including collaboration with CDP. We proposed a reliability estimation method of the ASR-generated transcripts to provide more reliable retrieval results for SpeechFind. We also proposed a novel confidence measure for reliability estimation, which is based on acoustically discriminative sub-word models. The proto-type version of SpeechFind was developed with the proposed reliability estimator applied. Experimental results on CDP database demonstrate that the proposed confidence measure is effective in improving the reliability estimator performance, improving user acceptance of the SDR system.

9. Acknowledgments

We wish to acknowledge the outstanding collaborations with CDP, in particular Leigh Grinstead, CDP Projects Coordinator, Jill Koelling, CDP Executive Director, and Nancy Allen, CDP Projects Chair for their support and extensive discussions which helped shape this project. We also thank the CDP for their financial support as well.

10. References

- [1] Hansen, J. H. L., Huang, R., Zhou, B., Seadle, M., Deller, J. R. Jr., Gurijala, A. R., Kurimo, M., Angkittrakul, P., “SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word,” *IEEE Trans. SAP*, 13(5):712-730, 2005.
- [2] <http://www.ngsw.org>.
- [3] <http://cdpheritage.org>.
- [4] Sukkar, R. A., Lee, C. H., “Vocabulary independent Discriminative Utterance Verification for Nonkeyword Rejection in Sub-word based Speech Recognition,” *IEEE Trans. SAP*, 4(6):420-429, 1996.
- [5] Wessel, F., Schluter, R., Macherey, K., Ney, H., “Confidence Measures for Large Vocabulary Continuous Speech Recognition,” *IEEE Trans. SAP*, 9(3):288-298, 2001.
- [6] Jiang, H., “Confidence Measures for Speech Recognition: A Survey,” *Speech Communication*, vol.45, pp.455-470, 2005.
- [7] NIST SPeech Quality Assurance (SPQA) package version 2.3, <http://www.nist.gov/speech>.