Phonetic Distance Based Confidence Measure

Wooil Kim, *Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—This letter presents a novel confidence measure for the purpose of improving user performance in Spoken Document Retrieval (SDR). The proposed confidence measure is based on the phonetic distance between subword models, employing an anti-model which is determined to be discriminative to a target model using offline training data. As an advancement from our previous work, the proposed method employs separate phonetic similarity knowledge for vowels and consonants, resulting in more reliable performance over diverse SDR recorded speech conditions. A transcript reliability estimator is also presented, with evaluation as an application of the proposed confidence measure. Analysis on a variety of corpora including background noise, frequency band-restrictions, and a range of real-life conditions, shows that the proposed confidence measure is more reliable in detecting corrupted speech due to acoustic conditions or an unarticulated speaking style, providing a higher correlation to word error rate (WER). The proposed confidence measure is effective in increasing transcript reliability estimation performance with a 16.21% relative improvement.**

***Index Terms*—Anti-model, confidence measure, phonetic distance, reliability estimation, spoken document retrieval.**

I. Introduction

As available audio information collections drastically increase, the need for effective information retrieval continues to expand, drawing remarkable attention to research on Spoken Document Retrieval (SDR) systems such as SpeechFind [1], [2] which serves as a platform for several programs across the United States for audio indexing and retrieval including the National Gallery of the Spoken Word (NGSW) [3] and the Collaborative Digitization Program (CDP) [4], [5]. Audio collections such as NGSW and CDP corpora generally include a diverse range of recording conditions which make Automatic Speech Recognition (ASR) for SDR extremely challenging. The motivation for the present study is to improve nonexpert user acceptance of ASR-based SDR systems, whose performance severely degrades due to adverse conditions in the audio corpus.

In this paper, we propose an advanced confidence measure which is based on the phonetic distance among subword models. The proposed confidence measure employs separate phonetic distance knowledge for vowels and consonants, which results in more reliable performance in various corrupted speech conditions compared to our earlier work [6]. By conducting a comprehensive analysis on both artificially generated acoustic conditions (i.e., background noise and channel distortion) and a wide range of real-life conditions, the proposed confidence measure will be shown to be more reliable and effective compared to other conventional confidence measures [7]–[10]. As an application of the proposed confidence measure, we present transcript reliability estimation with evaluations on the extended CDP corpus. The presented reliability estimator for text in the searched transcripts provides effective knowledge to an SDR user, who generally is not a speech/language technology expert.

Manuscript received July 03, 2009; revised September 16, 2009. First published October 16, 2009; current version published November 06, 2009. This work was supported by AFRL under a subcontract to RADC, Inc. (FA8750-05-C-0029). Approved for public release, distribution unlimited. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark Gales.

The authors are with the Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: wikim@utdallas.edu; john.hansen@utdallas.edu).

Digital Object Identifier 10.1109/LSP.2009.2034551

II. Phonetic Distance Based Confidence Measure

In this section, we present the confidence measure which employs a phonetic similarity between subword models. Our proposed confidence measure is an alternative to the anti-model based confidence measure generally obtained by calculating a log-likelihood ratio of a hypothesized (i.e., recognized) subword model and its corresponding anti-model [7]–[9]. In general, two approaches for obtaining the anti-model score are employed: i) union score and ii) max score. The union score approach compares the score of the recognized model against the average of the scores for all other models, while the max score approach compares the value to the maximum value among the remaining model scores. The model providing the max score is a competitive model to the hypothesized entry.

The union score approach can fail to discriminate scores of the recognized model and anti-model in some situations (e.g., due to noise-corruption or obscurely uttered speech), since its anti-model includes the competitive model. Both the union and max score approaches were originally designed to be applied to a subword sequence obtained by Viterbi decoding, so they can fail to estimate scores when Viterbi decoding fails due to severely corrupted speech. For the case where a frame-based subword (e.g., phone) recognizer framework is used without employing Viterbi decoding, the max score approach would have similar scores for a recognized model and its competitive model, which leads to difficulty in determining whether the input speech is clearly articulated or corrupted by other factors.

In our proposed confidence measure, we employ a phonetic distance that reflects the acoustic similarity of the subword models obtained by offline training [6]. Based on the obtained knowledge of the phonetic distance, a similar model group and dissimilar model group are determined for the input speech. If the incoming speech is clearly articulated and without background noise or channel distortion, the likelihood difference between the similar model to the input speech and the dissimilar entry is expected to become larger. Alternatively, input speech corrupted due to noise or obscure pronunciation would result in a failure to discriminate the two models, resulting in a smaller difference between the scores.

TABLE I
PHONETIC DISTANCE TABLE FOR VOWELS, SEMI-VOWELS AND GLIDES

	⟸ *similar*						*dissimilar* ⟹				
λ_i	$\lambda_{i,1}$	$\lambda_{i,2}$	$\lambda_{i,3}$	$\lambda_{i,4}$	$\lambda_{i,5}$	$\cdots$	$\lambda_{i,15}$	$\lambda_{i,16}$	$\lambda_{i,17}$	$\lambda_{i,18}$	$\lambda_{i,19}$
AA	AA	AW	AY	AO	AH	$\cdots$	IH	EY	UW	Y	IY
EH	EH	AE	EY	AY	IH	$\cdots$	Y	L	UW	IY	W
IH	IH	AH	EY	EH	UW	$\cdots$	AA	AO	OW	W	AW
OW	OW	OY	UH	AH	L	$\cdots$	UW	AE	EY	Y	IY
Y	Y	IY	IH	EY	UW	$\cdots$	AA	W	AO	OW	AW
L	L	OW	UH	AH	W	$\cdots$	EH	AE	EY	Y	IY
R	R	ER	UH	AO	AH	$\cdots$	W	AW	EY	IY	Y

TABLE II
PHONETIC DISTANCE TABLE FOR CONSONANTS

	⟸ *similar*						*dissimilar* ⟹				
λ_i	$\lambda_{i,1}$	$\lambda_{i,2}$	$\lambda_{i,3}$	$\lambda_{i,4}$	$\lambda_{i,5}$	$\cdots$	$\lambda_{i,16}$	$\lambda_{i,17}$	$\lambda_{i,18}$	$\lambda_{i,19}$	$\lambda_{i,20}$
B	B	G	D	T	P	$\cdots$	F	Z	S	ZH	SH
D	D	T	G	JH	B	$\cdots$	NG	F	S	ZH	SH
F	F	TH	T	K	CH	$\cdots$	B	ZH	N	M	NG
M	M	N	NG	G	D	$\cdots$	CH	F	ZH	S	SH
P	P	K	T	TH	B	$\cdots$	M	NG	SH	N	ZH
S	S	Z	T	CH	SH	$\cdots$	P	N	B	M	NG
T	T	D	K	G	HH	$\cdots$	N	SH	M	ZH	NG

Phonetic similarities between a particular subword model and the remaining models are identified using training data:

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

where $\mathbf{X}^{\{i\}}$ is a collection of training data labeled as model λ_i and $\lambda_{i,m}$ indicates the mth similar model among M subword models compared to the pivotal model λ_i. In our experiment, each subword model consists of a Gaussian mixture model (GMM) in the cepstral domain. The most similar model $\lambda_{i,1}$ is identical to the model itself (i.e., λ_i).

In this study, as an advanced aspect from our previous study [6], phonetic similarities were obtained separately for vowels and consonants to generate more reliable discrimination between similar and dissimilar models. In general, vowels are able to be acoustically represented more reliably compared to consonants, and their spectral difference between similar models and dissimilar models could still be large even with corrupted speech. Therefore, if the similar and dissimilar models are determined without separation of vowels and consonants, this spectral difference of vowel models would lead to a high score ratio in the corrupted condition, which results in the degradation of the confidence measure performance. Such a performance decrease was observed in our previous study [6], particularly for band-limited speech conditions where speech spectra are partially corrupted, so that vowels could be acoustically characterized well compared to consonants. Tables I and II show parts of the phonetic distance tables among context independent phones for vowels and consonants respectively, obtained by (1) using training data in this study. Here, we employed 19 vowels (including semi-vowels and glides) and 20 consonants. The identified models can be grouped into "similar group" $\boldsymbol{\lambda}_i^{sim}$ and "dissimilar group" $\boldsymbol{\lambda}_i^{dis}$ according to the phonetic distance to a pivotal model λ_i.

To calculate the proposed phonetic distance based confidence measure, the most similar group to the input speech X_t (i.e., cepstrum) is first determined based on a maximum likelihood decision:

$$\begin{aligned} i_{\max} &= \arg\max_i P(X_t|\boldsymbol{\lambda}_i^{sim}) \\ &= \arg\max_i \frac{1}{N_s}\sum_{n=1}^{N_s} P(X_t|\lambda_{i,n}) \end{aligned} \quad (2)$$

where the similar group $\boldsymbol{\lambda}_i^{sim}$ includes N_s of similar models (i.e., $\lambda_{i,1}$ to λ_{i,N_s}). In this study, five consecutive frames are used as an input to provide a cumulative score. The proposed confidence measure at time t is calculated using the log-likelihood ratio of the most similar group $\boldsymbol{\lambda}_{i_{\max}}^{sim}$ determined by (2) and the corresponding dissimilar group $\boldsymbol{\lambda}_{i_{\max}}^{dis}$ as follows:

$$p.cm_t = \log \frac{P(X_t|\boldsymbol{\lambda}_{i_{\max}}^{sim})}{P(X_t|\boldsymbol{\lambda}_{i_{\max}}^{dis})} \quad (3)$$

where $P(X_t|\boldsymbol{\lambda}_{i_{\max}}^{dis}) = 1/N_d \sum_{n=1}^{N_d} P(X_t|\lambda_{i_{\max},M-N_d+n})$ and N_d indicates the number of models in the dissimilar group $\boldsymbol{\lambda}_{i_{\max}}^{dis}$. In this study, we use 10 for both N_s and N_d considering the number of vowels and consonants. The $p.cm_t$ is obtained every 5 frames and averaged over the entire input duration for a final confidence measure. In this method, pairs of a target model (i.e., similar group) and its anti-model (i.e., dissimilar group) are determined using training data offline, resulting in acoustically discriminative entries for each other. Therefore, the phonetic distance based confidence measure is expected to be more robust than other conventional anti-model based measures, which can fail to differentiate scores from the recognized model and anti-model depending only on input speech. It is also noted that the proposed method does not require a subword sequence or segmentation obtained by Viterbi decoding.

III. APPLICATION: TRANSCRIPT RELIABILITY ESTIMATION

In order to achieve more effective spoken document retrieval, a more consistent and accurate collection of transcripts, and in particular keywords, are needed for search. However, most SDR systems based on speech recognition significantly degrade due to the diverse range of acoustic conditions in the audio such as background noise, recording media, speech styles, etc. For SDR systems such as SpeechFind for the NGSW [1] where audio material consist of recordings from the past 110 years, nonexperts in speech technology lose confidence in the system if they believe all transcripts are flawless. Therefore, the user would be helped if an estimate of transcript accuracy could also be provided through the system.

A transcript reliability estimator employing a confidence measure has been proposed in our earlier work [6]. In this study, the reliability estimator is based on Bayesian classification consisting of a Gaussian mixture model and prior information. Various confidence measures, including the proposed phonetic distance based measure are employed as feature vectors. We classify the ASR generated-transcripts into three categories (e.g., *good*, *fair* and *poor*) according to word-error-rate (WER). The designed reliability estimator can be used to classify the actual transcripts generated by the ASR engine for SDR, which improves the user's acceptance of the search engine by allowing them to have a separate measurement of transcript performance.

IV. EXPERIMENTAL RESULTS

A. Analysis on Artificial Acoustic Conditions: TIMIT Corpus

In this section, we present analysis of the proposed phonetic distance based confidence measure, with a comparison to other conventional confidence measures on a variety of spoken document conditions. We first evaluate the confidence measure using the 8 kHz sampled TIMIT database, which consists of phonetically balanced sentences with full-band speech (0–4 kHz).

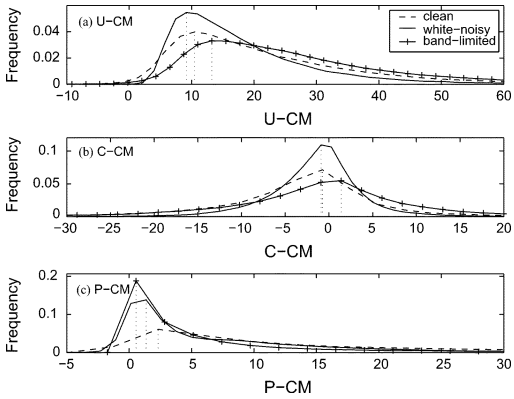

Fig. 1. Pdf estimates (normalized histogram) of CMs for vowels.

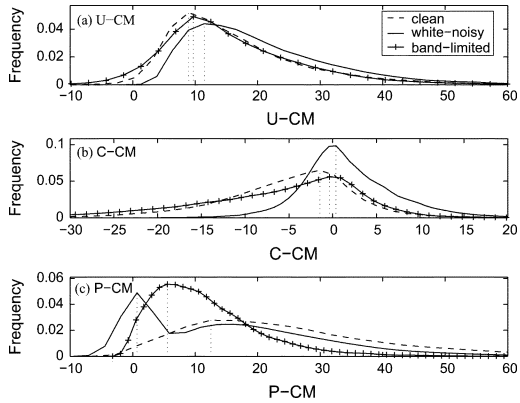

Fig. 2. Pdf estimates (normalized histogram) of CMs for consonants.

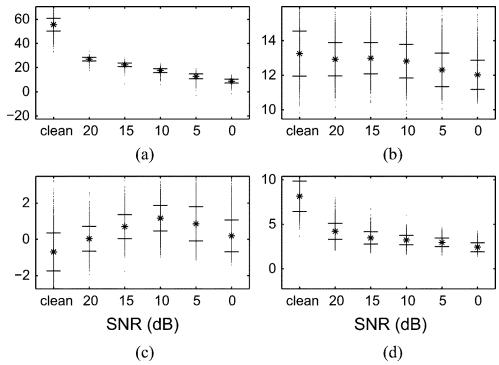

Fig. 3. Distributions of CMs in white noise corrupted speech conditions (a) SNR (b) U-CM (c) C-CM (d) P-CM.

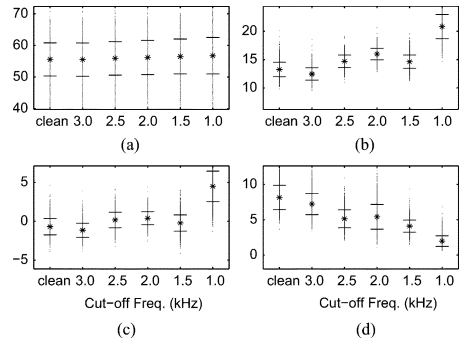

Fig. 4. Distributions of CMs in band-restricted speech conditions (a) SNR (b) U-CM (c) C-CM (d) P-CM.

To observe the performance of each confidence measure as a means of reflecting the acoustic condition in speech, two types of conditions were generated that include additive background noise corruption and band-restricted channel distortion. To obtain background noise corrupted speech, white noise samples from NOISEX92 data were added to produce five different SNR speech conditions (i.e., 20, 15, 10, 5, and 0 dB). The band-limited speech samples were generated by low-pass filtering the original clean speech to 3.0, 2.5, 2.0, 1.5, and 1.0 kHz cutoff frequencies.

Figs. 1 and 2 show probability density estimates (i.e., normalized histograms) of a) union anti-model based (U-CM), b) competitive anti-model based (C-CM), and c) the proposed phonetic distance based (P-CM) confidence measures for clean (dashed line), white noise corrupted (5 dB SNR, solid line), and band-limited speech (1.5 kHz cutoff, "+" line) conditions for vowels and consonants respectively. In our experiment, a consistent framework is used for P-CM, U-CM, and C-CM, which employs a phone recognizer that determines target and anti models every five frames without Viterbi decoding, in order to avoid a failure in obtaining the subword sequence in adverse conditions, and also to perform a fair comparison with the proposed P-CM. From these figures, we see that the modes of P-CM pdfs of both corrupted conditions (white noise and band-limited) are shifting to the left from those of clean conditions, with a higher peak and narrow width, resulting in a more distinguished pdf shape from clean conditions compared to the other two confidence measures (i.e., U-CM and C-CM). These pdf transitions of P-CM in corrupted speech conditions are expected to present more robustness for detection of the input speech corruption.

Figs. 3 and 4 show distributions of SNR, U-CM, C-CM, and P-CM for cases of various SNRs and cutoff frequencies for white noise corrupted and band-limited speech conditions. Here, the SNR was obtained using the NIST Speech Quality Assurance tool [11], which was found to show on average 5–10 dB more than the actual SNR due to blind knowledge on silence duration content. In each figure, mean values of each statistic are presented together with the corresponding standard deviations using small bars. For an equivalent comparison, the vertical axis of each figure is normalized to the same scale, therefore the slope generated by the mean values should reflect the correlation of the given measure to the acoustic conditions (i.e., the steeper the slope, the higher the correlation). While the SNR measure unsurprisingly shows a high correlation with actual SNR of white noise corrupted speech, no correlation is seen for the band-limited speech. Unlike U-CM and C-CM which show less correlation to SNR, the proposed phonetic distance based measure (P-CM) shows higher correlation trends for both acoustic corrupted conditions such as white noise corruption SNR and the degree of frequency restriction respectively.

We also evaluated the correlation of each confidence measure to the WER of the test utterance in the adverse acoustic conditions. To obtain the WER, an HMM based speech recognizer with tri-phone models was employed with a 6233-word vocabulary and a trigram language model adapted on the TIMIT task using a Broadcast News model. Table III shows the correlation coefficient of each confidence measure to WER in each acoustic

TABLE III
CORRELATION COEFFICIENTS BETWEEN CONFIDENCE MEASURE AND WER ON TIMIT CORPUS

	SNR	U-CM	C-CM	P-Base[6]	**P-CM**
White Noise	-0.678	-0.275	0.348	-0.720	**-0.683**
Band-Limited	0.069	0.554	0.445	-0.502	**-0.660**

TABLE IV
CORRELATION COEFFICIENTS BETWEEN CONFIDENCE MEASURE AND WER ON CDP CORPUS

SNR	All-Phone	Lang	WD	U-CM	C-CM	**P-CM**
-0.235	0.257	0.350	-0.265	-0.212	0.176	**-0.435**

condition, which represents a similar trend to that seen in Figs. 3 and 4. The results in Table III show that the proposed confidence measure (P-CM) is highly correlated to WER for both white noise corruption and band-restricted conditions, compared to other statistics. By employing separate phonetic distance knowledge for vowels and consonants, we obtain a more consistent correlation to WER for both adverse conditions compared to our earlier work [6] (P-Base in Table III) which shows less correlation in band-limited conditions.

B. Analysis on Real-Life Conditions: CDP Corpus

Next, we conducted an analysis of the proposed confidence measure using the CDP corpus, which contains a range of real-life acoustic conditions with a total 32-h evaluation set. The CDP collaboration with CRSS-UTD represents a collection of 1300 h of audio material from libraries and archives [5]. In this experiment, the all-phone model score (All-Phone), language model score (Lang), and word density (WD) [10] were also investigated, including the confidence measures discussed in the previous section. Table IV [1] shows the correlation relationships of the confidence measures to WER. The proposed phonetic distance based confidence measure (P-CM) shows the highest correlation to WER, while other anti-model based measures (U-CM and C-CM) do not show comparable results to Table III.

C. Transcript Reliability Estimator

As an application, the transcript reliability estimator was evaluated using the CDP corpus. A total of 32 h of training data and 7 h of test data were used in this experiment. The acoustic model for each class (*good*, *fair* and *poor*) [2] consists of a 4-component Gaussian mixture. To evaluate the reliability estimator, we employ an evaluation criterion named "critical error rate" together with recognition accuracy [6]. The critical error rate reflects a combination of false alarms and "significant" misses which would critically mislead users depending on the retrieved transcripts.

Table V shows performance of the transcript reliability classifier employing several combinations of confidence measures. The baseline system employs four confidence measures as a feature vector, including SNR, all-phone model score, language model score, and word density. From these results, by adding the union anti-model confidence measures (U-CM) and the proposed phonetic distance based confidence measure (P-CM) to the baseline, improved overall performance was obtained that is consistent with the correlation trend shown in Table IV. We obtained the best performance by adding both U-CM and P-CM to the baseline system, showing a 16.21% relative improvement in critical error. These results demonstrate that the proposed confidence measure is superior at improving the performance of reliability estimation, thus ensuring intelligibility of the retrieved utterances for the user.

[1]The difference from results presented in [6] is due to a different implementation of U/C-CM and a significant increase in the speech corpus size.

[2]The overall WER was 71.07% and the classes were determined considering class distribution and data amount for train/test; 13%, 67%, and 20% distributions for *good* ($\mathrm{WER} \leq 45\%$), *poor* ($\mathrm{WER} \geq 65\%$), & *fair* (others).

TABLE V
PERFORMANCE OF RELIABILITY ESTIMATOR FOR THREE CLASSES (%)

System	Accuracy	Critical Error	(Relative)
Baseline	60.54	29.68	-
+U-CM	62.87	27.46	(7.48)
+C-CM	59.03	31.08	(-4.72)
+P-CM	**63.84**	**26.11**	**(12.03)**
+U-CM+P-CM	**64.70**	**24.87**	**(16.21)**

V. CONCLUSIONS

In this study, a novel confidence measure was presented based on the phonetic distance among subword models. The proposed method employed an anti-model which was determined to be acoustically discriminative to a target model obtained from offline training. By utilizing separate phonetic similarity knowledge for vowels and consonants, more reliable performance of the proposed confidence measure was obtained in various corrupted speech conditions. Analysis using various corpora demonstrated that the proposed confidence measure is more effective in detecting corrupted speech due to acoustic conditions or unarticulated speaking style, showing high correlation to WER in speech recognition. Experimental results also showed that the proposed confidence measure is effective at increasing the reliability estimator performance with a 16.21% relative improvement. This study showed that a reliability estimator employing the proposed confidence measure will be effective at improving user acceptance of a SDR system.

REFERENCES

[1] J. H. L. Hansen, R. Huang, B. Zhou, M. Seadle, J. R. Deller, Jr, A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "SpeechFind: Advances in spoken document retrieval for a National Gallery of the Spoken Word," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 712–730, 2005.
[2] [Online]. Available: http://SpeechFind.utdallas.edu
[3] [Online]. Available: http://www.ngsw.org
[4] [Online]. Available: http://cdpheritage.org
[5] W. Kim and J. H. L. Hansen, "Advances in spoken document retrieval for the U.S. collaborative digitization program," in *ASRU2007*, 2007, pp. 687–692.
[6] W. Kim and J. H. L. Hansen, "Advances in speechfind: Transcript reliability estimation employing confidence measure based on discriminative sub-word model for SDR," in *Interspeech2007*, 2007, pp. 2409–2412.
[7] R. C. Rose, B. H. Juang, and C. H. Lee, "A training procedure for verifying string hypotheses in continuous speech recognition," in *ICASSP-95*, 1995, pp. 281–284.
[8] R. A. Sukkar and C. H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 6, pp. 420–429, 1996.
[9] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Commun.*, vol. 45, pp. 455–470, 2005.
[10] F. 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, 2001.
[11] [Online]. Available: http://www.nist.gov/speech