Verification of Normalized Confidence Measure Using n-Phone Based Statistics

Byoung Don Kim* · Jin Young Kim* · Seung You Na* · Seung Ho Choi**

ABSTRACT

Confidence measure (CM) is used for the rejection of mis-recognized words in an automatic speech recognition (ASR) system. Rahim, Lee, Juang and Cho's confidence measure (RLJC-CM) is one of the widely-used CMs [1]. The RLJC-CM is calculated by averaging phone-level CMs. An extension of the RLJC-CM was achieved by Kim et al [2]. They devised the normalized CM (NCM), which is a statistically normalized version of the RLJC-CM by using the tri-phone based CM normalization. In this paper we verify the NCM by generalizing tri-phone to n-phone unit. To apply various units for the normalization, mono-phone, tri-phone, quin-phone and ∞-phone are tested. By the experiments in the domain of the isolated word recognition we show that tri-phone based normalization is sufficient enough to enhance the rejection performance of the ASR system. Also we explain the NCM in regard to two class pattern classification problems.

Keywords: CM, RLJC-CM, NCM

1. Introduction

Speech is the most efficient and natural modality, with which human communicates easily. Also, speech is a convenient tool in man-machine communication: thus, ASR has been actively studied for more than three decades. Although much research has been done, it is not easy to apply ASR to real services due to noise and rejection problems.

In various service domains of ASR, a reliable confidence measure of a speech recognizer's output is useful and important. Basically, CM is of use for rejecting a hypothesis that is likely to be erroneous in a hypothesis test. For example, CM is a useful measure to detect out-of-vocabularies (OOV) in continuous speech recognition or mis-recognized words in isolated speech recognition. Many kinds of CM have been developed [1-6]. The rejection problem of mis-recognized words is a two-class

* Dept. of Electronics Eng., Chonnam National University.
** Depart. of Multimedia Communications Eng., Dongshin University.

classification of 'true' or 'false'. In the case of ASR we don not have the model of anti-hypotheses: therefore, the phone-models are commonly used as anti-hypotheses. One of the representative CMs is the RLJC-CM proposed by Rahim et al [1]. They obtain the word-level CM as the weighted sum of the phone-level confidence measures. Kim et al extended the RLJC-CM by introducing the concept of normalization [2]. They perceived the statistical inconsistency of word and phone-level confidence measures. Since CM statistics, mean and variance, of various phones and words are not same, some words seem to be always rejected under the specific threshold. To overcome this problem, the normalization of the phone-level CMs was tried. For normalizing the phone-level CMs Kim, et al used the statistics of the tri-phone CMs. They achieved a good result using the normalization process. According to the reference [2], however, there are no any experimental or theoretical explanations of the reasons that the tri-phone based normalization is applied.

In this paper we generalize the tri-phone based normalization to the n-phone based normalization. And we try to test the normalization processes using phone, tri-phone, quin-phone and ∞-phone units. From the recognition and confirmation experiments the normalized confidence measure is verified. We present some experimental results, which explain why the rejection performance is enhanced by the normalized confidence measure.

2. Experimental Setup and Baseline System

An isolated word recognition (IWR) system is used for verifying the NCM. The CNU-IWR, which was developed by the CNU DSP laboratory, is a middle-sized word recognition system. The CNU-IWR can recognize 1,000 words of Korean stock and road names. The speech DB used for training is composed of 1,000 distinctive words. Each word was spoken by 70 peaple. All the utterances were recorded in the office environment. The spoken words were digitized with 16 bit and 8 kHz sampling.

For the IWR system we use tri-phone based HMMs (hidden Markov model) and the token-passing algorithm for decoding the input speech [7,8]. The Mel-cepstrum feature and delta-parameter are used as the speech parameters. The Tri-phone HMM models has been trained using the HTK 3.0 (hidden Markov model tool kit) developed by Cambridge University. The HMMs are obtained based on the left-to-right three state and tied-state modeling. The recognition module is composed of pre-processing, decoding processing and post-processing. A speech signal is transformed to the Mel-cepstrum based feature vectors in the pre-processing. In the post-processing the utterance verification is adopted for the rejection of mis-recognized words. In our experiments of the NCM verification we

use 600 words as true words. The last 400 words are used as imposters or attack words against the IWR system. As a result the specifications of the IWR system are as follows.

a. Speech database: speaker-independent (70 talkers, about 1,000 utterances per speaker)
b. Number of words in lexicon : 600 words
c. Number of attack words : 400 words
d. Recognition unit and model : tri-phone and HMM
e. Feature vectors: 8kHz sampling, 12 MEL-cepstrum, log energy and their delta-parameters
f. Number of utterances used in training : (52 talkers * 600 words) utterances
g. Number of utterances used in testing : (18 talkers * 600 words) utterances as true claims
(18 talkers * 400 words) utterances as attacks

On the other hand, we use the mis-detection ratio (MDR) and the false alarm ratio (FAR) to evaluate the performance of the confidence measures. The following equations are the definitions of MDR and FAR.

$$MDR = \frac{N_{MW}}{N_{CW}} \tag{1}$$

$$FAR = \frac{N_{FA}}{N_{CW}} \tag{2}$$

, where N_{MW} is the number of mis-recognized words, N_{CW} is the number of tried correct words, N_{FA} is the number of false alarmed words and N_{AW} is the number of attack words.

3. Generalization of the NCM

The normalized confidence measure is an extended version of the RLJC-CM proposed by Rahim et al. [1]. First the phone-level CMs are calculated by the eq. (4) in the RLJC-CM. And then the word-level CM is obtained by calculating the weighted sum of the phone-level CMs. The eq. (5) shows the averaging process.

$$logPr_a = \frac{1}{M}\sum_{i=1}^{M} logPr_a(i) \tag{3}$$

$$cm_p = \frac{logPr_p - logPr_a}{|logPr_p|} \tag{4}$$

$$CM = \frac{1}{f_{cm}}\log\left(\frac{\sum_{p=1}^{n_p} exp(f_{cm} cm_p)}{n_p}\right) \tag{5}$$

In the equations above, $logPr_a$ is an average log probability of the anti-phone models, and $logPr_p$ is the log probability of the p-th phone model. n_p is the number of the phones of a given word and M is a value of the anti-phone model. cm_p is the confidence of a phoneme unit and f_{cm} is a weighting value.

In the normalized confidence measure the phone-level CMs are regularized to have the same probability density functions (PDF) by a pdf transform. It is easily achieved by mean shifting and scaling. In the previous work [2] all the phone-level CMs are normalized using the statistics of tri-phone unit CMs. This transform is shown in the eq. (6). Thus the NCM can be redefined by the eq. (7).

$$ncm_{p,tri} = \frac{cm_p - \mu_{tri}}{\sigma_{tri}} + \alpha \tag{6}$$

$$NCM = \frac{1}{f_{ncm}}\log\left(\frac{\sum_{p=1}^{n_p} exp(f_{ncm} ncm_{p,try})}{n_p}\right) \tag{7}$$

, where f_{ncm} is a weighting value of the phone level CM. cm_p is the CM of the eq. (4) and $ncm_{p,tri}$ is a normalized phone-level CM. α is a shifting value used for controling the range of NCM values. In the equations. of (6) a (7) the weighting values of f_{cm} and f_{ncm} are empirically. The shifting value α does not change the rejection performance. In our experiments f_{cm} is -5, f_{ncm} is -0,5 and α is 2. Now we can generalize the eq. (6) by substituting the statistics of n-phone CMs in the place of those of tri-phone CM. The eq. (8) represents the generalization of the eq. (6).

$$ncm_{p,n-phone} = \frac{cm_p - \mu_{n-phone}}{\sigma_{n-phone}} + \alpha \tag{8}$$

, where $\sigma_{n-phone}$ is the standard deviation and $\mu_{n-phone}$ is the mean in the n-phone based CM.

When considering the phonetic environments we always use the left and right phones of the given phone as the phonetic environments. So, n is commonly odd number in *n-phone*. For example, phone, tri-phone and quin-phones are used as recognition units. Table 1 shows n-phones applied in our NCM verification.

Table 1. n-phones tested in the experiments of NCM verification.

n-number	name	example of '메뉴명'
1	phone	mz, ee, nq, yu, mz, yv, ng
3	tri-phone	mz+ee, mz-ee+nq, ee-nq+yu, nq-yu+mz yu-mz+yv, mz-yv+ng, yv-ng
5	quin-phone	mz+ee+nq, mz-ee+nq+yu, mz-ee-nq+yu+mz ee-nq-yu+mz+yv, nq-yu-mz+yv+ng yu-mz-yv+ng, mz-yv-ng
∞	∞-phone	

In Table 1 ∞-phone considers the case that the phones from different words and positions are treated separatively. In the isolated word recognition the phone number of each word is not actually infinite. In that case n is variable according to the number of phones. The infinite, ∞, explain all the cases.

4. Experimental Results and Verification of NCM

We performed the IWR experiments with the 600 word recognition system. 10,800 (600 words * 18 persons) utterances were tested as true claims. and 7,200 (400 words * 18 persons) spoken words were used as false claims. Table 2 shows the numbers of the units of phone, tri-phone, quin-phone and ∞-phone.

Table 2. Number of each normalized factor

normalized factor	mono-phone	tri-phone	quin-phone	∞-phone
number	51	2,347	5,470	6,217

As we expected, the number of units is increasing as n get larger. The example of the statistics of n-phone based CMs is shown in Table 3. The experimental results are

explained in the following two sections.

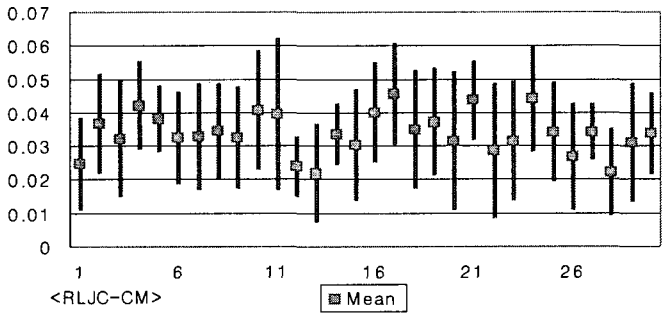
Table 3. Example of CM statistics of the word '메뉴명'

	construct models	appeared number	Mean	Standard deviation
Mono-phone	mz	8,040	0.063	0.102
	ee	7,426	0.036	0.033
	nq	13,873	0.038	0.037
	yu	3,367	0.024	0.049
	mz	8,040	0.063	0.102
	yv	8,651	0.037	0.050
	ng	41,173	0.012	0.039
Tri-phone	mz+ee	155	0.075	0.015
	mz-ee+nq	104	0.031	0.026
	ee-nq+yu	104	0.034	0.052
	nq-yu+mz	52	0.050	0.026
	yu-mz+yv	52	0.050	0.031
	mz-yv+ng	520	0.034	0.026
	yv-ng	312	-0.076	0.022
Quin-phone	mz+ee+nq	52	0.011	0.017
	mz-ee+nq+yu	52	0.036	0.026
	mz-ee-nq+yu+mz	52	0.042	0.054
	ee-nq-yu+mz+yv	52	0.050	0.026
	nq-yu-mz+yv+ng	52	0.050	0.030
	yu-mz-yv+ng	52	0.035	0.024
	mz-yv-ng	208	-0.078	0.022
Word dependent	mz	52	0.113	0.175
	ee	52	0.036	0.026
	nq	52	0.042	0.054
	yu	52	0.050	0.026
	mz	52	0.050	0.031
	yv	52	0.035	0.024
	ng	52	-0.072	0.022

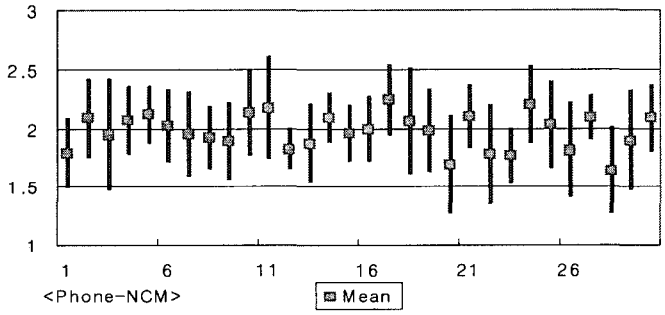
4.1 The Normalization Effects of CM

The main purpose of the normalization is to accomplish the statistical consistency of word-level CMs. That is, the means and standard deviations of word-level CMs get similar after the normalization of phone-level CMs. To verify the normalization effect we show the some statistics of word-level CMs with and without the normalization in Figure 1. As shown in Figure 1 the rectangle boxes represent the mean values and the lines represent the standard deviation values. In the case without the normalization the ratio of the maximum to the minimum is approximately 2.5 in the mean values. After the normalization the ration is reduced to about 1.3. Thus it is confirmed that the

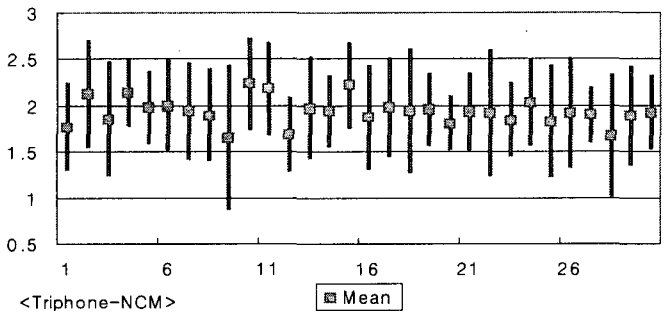
normalization process attains the statistical consistency.



(a) Statistics of Word-level CMs without normalization



(b) Statistics of Word-level CMs with mono-phone based normalization

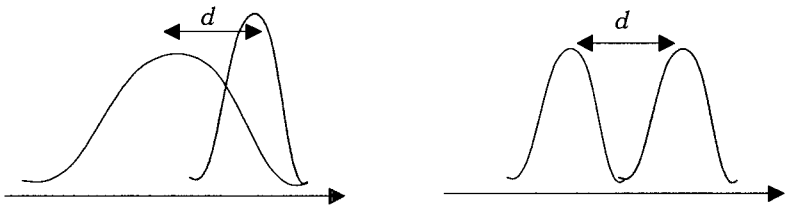


(c) Statistics of Word-level CMs with tri-phone based normalization

Figure 1. Examples of the statistics of word-level CMs

On the other hand, we can explain the normalization effect in regard to two-class classification problem (TCCP). The different point of our TCCP from other common TCCPs is that just the true claim model is known. We have no information about the imposters. We can image two cases shown in Figure 2. In Figure 2 (a) the models of two

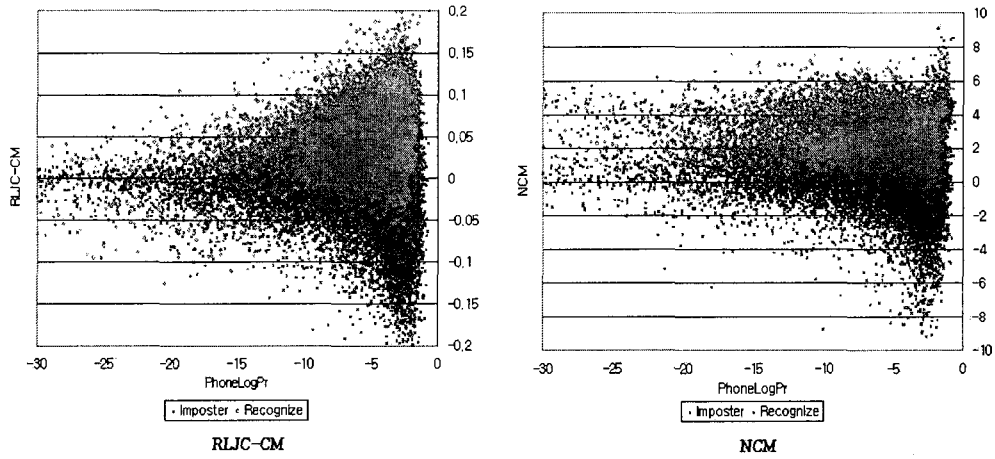
classes are very similar. Contrary to the case of Figure 2 (a) the models of two classes in Figure 3 (b) are distinctive. According to the pattern classification theory we can select the optimum threshold when we know the pdfs of the two classes. As mentioned above we do not know the pdf of the imposter claim. Thus it is impossible to determine the threshold adaptively depending on the imposter pdf. But we can enhance the classification performance utilizing the true claim's statistics. In Figure 2 both confusable and separable cases have the same value d of the mean differences. However, if we normalize the pdfs with true claim's statistics, we can enhance the classification performance. The reason is that after the pdf transformation the mean difference of the well-separable case is comparably larger than that of the confusing case. Obviously, the pdf transformation has no effect on the confusing case.



(a) Confusable classes (b) Well-separable classes

Figure 2. Two class classification problem.
(solid line : true claim, dotted line : false claim)

Figure 3 shows the distribution of the log probability and the confidence measure value in the case of the phone-level CMs with and without the normalization. Figure 3 (a) shows the results of the RLJC-CM and Figure 3 (b) shows the results of the tri-phone based NCM. From Figure 3 (a) we can observe that it is not easy to distinguish the true claims from the imposter claims on the phone level. The distribution of the true claim (DTC) is very confusable with that of the false claim (DFC). According to Figure 3 (b) the distributions of DTC and DFC are still similar, but DTC is a little separable from DFC comparatively. We think that this normalization effect can enhance the rejection performance.



(a) RLJC-CM case (b) Normalized CM case

Figure 3. logPr-CM plots of phone level CMs

4.2 The Performance Results of n-phone Based NCM

In this section we represent the performances of *n*-phone based NCM. Table 3 shows some statistics of *n*-phone based CMs. From Table 3 it is observed that the statistics of tri-phones are similar to those of quin-phones and ∞-phones. But the differences of the statistics between mono-phone and tri-phone are somewhat large. From these facts we can expect that tri-phone (3-phone) based normalization is sufficient enough to achieve the enhancement of the rejection performance.

Figure 4 shows the experimental results of RLJC-CM and NCM. The performances of phone, tri-phone, quin-phone and ∞-phone based normalizations are compared. This figure shows that the rejection performances of NCMs are better than those of RLJC-CM. Also, according to Figure 4 the performances of tri-phone based NCM is similar to those of quin-phone and ∞-phone based NCMs. Also the trip-hone based NCM reduces FAR(false alarm rate) from 0.35 to 0.29 FA/words with 9% MDR. It achieves 17% reduction of phone based NCM's FAR. Although the tri-phone based NCM is similar to quin-phone and ∞-phone based NCMs, the model numbers of the normalizations are greatly different. Approximately, the number of quin-phone is double the number of tri-phone. That is, it causes the increase of recognition time. Consequently we may conclude that tri-phone based NCM achieves the best performance.

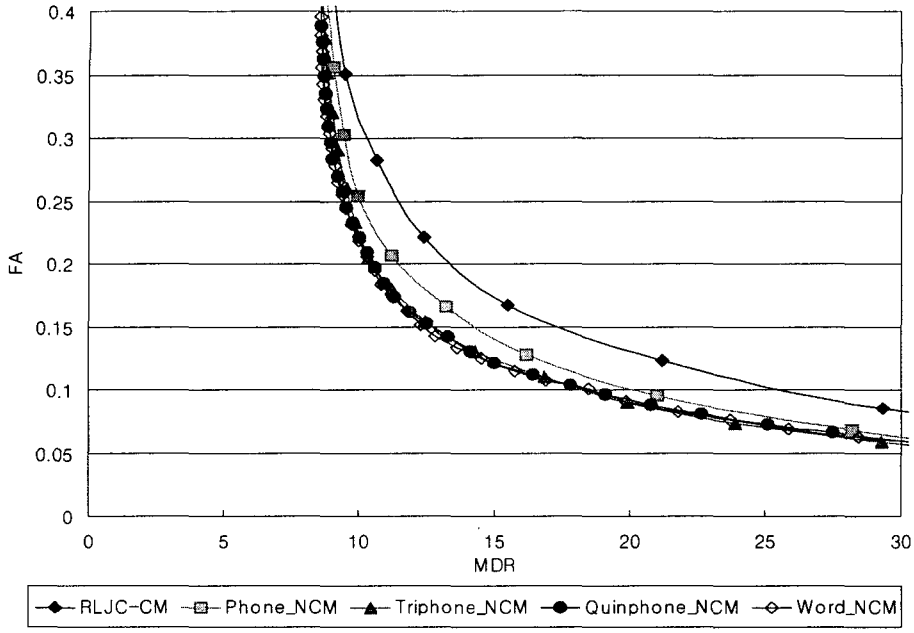


Figure 4. Experimental results of n-phone based normalizations

5. Conclusions

In this paper we studied the verification of the normalized confidence measure. Using the concept of n-phone, we extended the normalized confidence measure into a general formula. We tested n-phone based normalizations on the isolated word recognition domain with different n values. From the experimental results we observed that

1) the normalization procedure of the phone-level CMs could achieve the statistical consistency of the word-level CMs and that
2) the tri-phone based normalization is sufficient enough to get the optimal performance.

Furthermore, since our experiment is limited to the medium-sized IWR, it is necessary to test our n-phone based NCM on a large vocabulary recognition system in the future. Importantly, we need to devise accurate and adoptable anti-models for enhancing rejection performance.

Acknowledgement

This work was supported by RRC-HECS, CNU.

Reference

[1] Rahim, M. G., Lee, C. H., Juang, B. H. and Chou, W. 1996. "Discriminative utterance verification using minimum string verification error (MSVE) training." *Proc. of ICASSP96*, pp.3585-3589.

[2] Kim, J. Y., Choi, S. H., Lee, Joohun. 2002. "Hybrid confidence measure for domain-specific keyword spotting." LNAI 2358, *Developments in Applied Artificial Intelligence*, Tim Hentlass and Moonis Ali (Eds.), pp.736-745.

[3] Jiang, L. and Huang, X. 1998. "Vocabulary-independent word confidence measure using subword featuresm." *Proceedings of ICSLP'98*, pp.3245-3258.

[4] Rahim, M., Lee, C. H., Juang, B. H. and Chou, W. 1995. "Discriminative utterance verification for connected digits recognition." *Proceedings of Eurospeech95*, 1995.

[5] Kim, H., Yi, S. and Lee, H. 1999. "Out-of-vocabulary rejection using phone filler model in variable vocabulary word recognition," *Proceedings of ICSP* Seoul, pp.337-339.

[6] Tsiporkova, E., Vanpoucke, F. and Hamme, H. V. 2000. "Evaluation of various confidence-based strategies for isolated word rejection." *Proceedings of ICSLP 2000*, pp.803-806.

[7] Young, S. J., Russell, N. H., Thornton, J. H. S. 1989. "Token Passing: a simple conceptual model for connected speech recognition systems." *Technical report of Cambridge University Engineering Department*, TR38.

[8] Huang, X. Acero, A. and Hon, H. 2001. *Spoken Langauage Processing : A Guide to Theory, Algorithm and System Development.* Prentice Hall.

[9] Conrad Sanderson, Kuldip K. Paliwal, 2002. "Likelihood normalization for face authentication in variable recording conditions." *Proceedings of IEEE ICIP 2002*, vol. I, pp. 301-304.

[10] Duda, R., Hart, P. and Stork, D. 2001. *Pattern Classification.* John-Wiley and Sons.

received: January 30, 2005
accepted: March 14, 2005

▲ Byoung Don Kim
Dept. of Electronics Engineering, Chonnam National University
300 Yongbong-Dong, Buk-Gu, Gwangju, 500-757 Korea
Tel: +82-62-530-0472 FAX: +82-62-530-1750
E-mail: doni96@empal.com

▲ Jin Young Kim
Dept. of Electronics Engineering Chonnam National University
300 Yongbong-Dong, Buk-Gu, Gwangju, 500-757 Korea
Tel: +82-62-530-1757 FAX: +82-62-530-1750
E-mail: beyondi@chonnam.ac.kr

▲ Seung Yu Na
Dept. of Electronics Engineering Chonnam National University
300 Yongbong-Dong, Buk-Gu, Gwangju, 500-757 Korea
Tel: +82-62-530-1757 FAX: +82-62-530-1750
E-mail: syna@chonnam.ac.kr

▲ Seung Ho Choi
Dept. of Multimedia Communications Engineering Dongshin University
252 Daeho-Dong, Naju, Chonnam, 520-714 Korea
Tel: +82-61-330-3194 FAX: +82-61-330-3194
E-mail: shchoi@dsu.ac.kr