Bayes-Based Confidence Measure in Speech Recognition

Néstor Becerra Yoma, *Member, IEEE*, Jorge Carrasco, and Carlos Molina, *Student Member, IEEE*

***Abstract*—In this letter, Bayes-based confidence measure (BBCM) in speech recognition is proposed. BBCM is applicable to any standard word feature and makes use of information about the speech recognition engine performance. In contrast to ordinary confidence measures, BBCM is a probability, which is interesting itself from the practical and theoretical point of view. If applied with word density confidence measure (WDCM), BBCM dramatically improves the discrimination ability of the false acceptance curve when compared to WDCM itself.**

***Index Terms*—Bayes theorem, confidence measure, dialogue systems, speech recognition.**

I. Introduction

ONE OF THE most important motivations of speech recognition technology is to provide a very natural and familiar interface for human–machine interaction. However, speech input presents many challenges. Automatic speech recognition (ASR) engines currently operate without substantial parts of the human communication repertoire, such as gestures, intonation, and facial expressions. Moreover, real natural speech is usually unbounded, and the speaker could easily exceed the current vocabulary or grammar of the engine. In addition, ASR engines must often deal with large variations in the speaker's environment. In fact, background noise, microphone quality, and reverberation dramatically increase the word error rate (WER). Finally, pronunciation diversity within a city or country and intraspeaker variation also contribute to make the speech-to-text procedure one of the most challenging tasks in technology. As a consequence, robustness of ASR has continuously attracted the attention of the speech community, as interactive systems are getting more popular in fields like telephone services.

Speech recognition has created practical opportunities for call centers, internal company operations using the telephone, and telephone service providers. On the other hand, telephony is a natural market for speech recognition. Interactive voice response systems with ASR allow users to connect with the information they need, from anywhere at any time, by employing natural language. However, avoiding user's frustration is critical. To do so, interaction needs to be very effective and efficient, and confirmation loops should be avoided. In this context, reliably assessing the operation of ASR is necessary in all practical systems to decide whether a recognized word or sentence should be accepted or rejected [1]. Moreover, the information provided by confidence measures could be applied to unsupervised noise, environment and user adaptation algorithms [2], to out-of-vocabulary (OOV) detection approaches, and to reorder the hypotheses in a N-best decoder [3], [4]. There are several methods to estimate word-level confidence measure. They are based on word length, word acoustic score, or word density [5]. The word density approach—word density confidence measure (WDCM)—is probably the most popular in the specialized literature and is based on the frequency of the word's occurrences in the N-best list delivered by the Viterbi decoding [6]. The confidence measure based on the word density WDCM_i, where i is the word index, is computed as [5], [7]

$$\text{WDCM}_i = \frac{\sum_{r \in E(w_i,H)} Q(h_r)}{\sum_{l=1}^{N} Q(h_l)} \tag{1}$$

where $Q(h_r) = P(h_r)^{\gamma} P(O/h_r)$; h_l is the lth hypothesis in the N-best Viterbi list; $Q(h_r)$ is the likelihood score of h_r given by the Viterbi search; $P(h_r)$ is the language model probability of h_l; $P(O/h_r)$ is the observation probability of h_l; γ is the acoustic model scaling factor; $E(w_i, H)$ corresponds to the indices of the hypotheses, where word w_i is contained; and finally, H denotes all the N-best alignments or hypotheses obtained from Viterbi decoding.

Surprisingly, the confidence measures proposed so far, including WDCM, are not probabilities [1]. Moreover, they do not take into consideration *a priori* information about the performance of the ASR engine. The contribution of this letter concerns a Bayes-based confidence measure (BBCM) that is a probability itself and incorporates *a priori* information about the recognizer. When compared to WDCM, BBCM dramatically improves the discrimination of misrecognized words. BBCM can also provide more symmetrical false acceptation and false rejection curves if applied with WDCM in combination with word maximum hypothesis log-likelihood. It is worth emphasizing that, as a probability, BBCM is more applicable to stochastic adaptation algorithms and to the OOV problem. Finally, the approach presented here has not been found in the literature.

II. BBCM

Several word features could be extracted from the Viterbi decoding [1]: mean acoustic likelihood score, mean difference from maximum score, number of N-best, and number of acoustic observations. Notice that WDCM is a word feature

Manuscript received January 31, 2005; revised June 14, 2005. This work was supported by Conicyt/Chile: Fondecyt N°1030956 and Fondef N°D02I-1089. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Frederic Bimbot.

The authors are with the Department of Electrical Engineering, Universidad de Chile Santiago, Santiago, Chile (e-mail: nbecerra@ing.uchile.cl).

Digital Object Identifier 10.1109/LSP.2005.856888

itself in this context. If WF denotes a given word feature, BBCM is defined as

$$\begin{aligned}\text{BBCM}(\text{WF}_i) &= P(w_i \text{ is correct}/\text{WF}_i) \\ &= \frac{P(\text{WF}_i/w_i \text{ is correct}) \cdot P(w_i \text{ is correct})}{P(\text{WF}_i)}\end{aligned} \tag{2}$$

where the event "*w_i is correct*" corresponds to the fact that word w_i, which is contained at least in one of the N-best hypotheses, was properly recognized (i.e., it is in the transcription of the testing utterance). Notice that $\text{BBCM}(\text{WF}_i)$ is a probability itself. Moreover, the distributions $P(\text{WF}_i/w_i \text{ is correct})$ and $P(\text{WF}_i)$ and the probability $P(w_i \text{ is correct})$ provide information about the recognition engine performance.

In this letter, BBCM was tested with the following word features: WDCM_i and $\text{ML}_i = \max[Q(h_r)]|r \in E(w_i, H)$, where i is the word index. ML_i is the maximum hypothesis log-likelihood within the N-best list, where w_i is found. The use of ML_i is motivated by the fact that WDCM is a ratio of summation of likelihoods, so the real value of the hypothesis likelihood is lost. Observe that BBCM could be applied to any other feature obtained from the Viterbi decoding.

A. Probability of a Word Is Correct Given $WDCM_i$

The probability $\text{BBCM}(\text{WDCM}_i)$=$P(w_i \text{ is correct } \text{WDCM}_i)$ is estimated according to

$$\begin{aligned}&P(w_i \text{ is correct}/\text{WDCM}_i) \\ &\quad = \frac{P(\text{WDCM}_i/w_i \text{ is correct})P(w_i \text{ is correct})}{P(\text{WDCM}_i)}.\end{aligned} \tag{3}$$

The functions $P(\text{WDCM}_i/w_i \text{ is correct})$ and $P(\text{WDCM}_i)$, and the probability $P(w_i \text{ is correct})$ are computed with the evaluation data that are different from the training and testing databases. $P(\text{WDCM}_i/w_i \text{ is correct})$ and $P(\text{WDCM}_i)$ were approximated with discrete probability distributions, as shown in Fig. 1.

B. Probability of a Word Is Correct Given ML_i

The probability $\text{BBCM}(\text{ML}_i) = P(w_i \text{ is correct}/\text{ML}_i)$ is estimated according to

$$\begin{aligned}&P(w_i \text{ is correct}/\text{ML}_i) \\ &\quad = \frac{P(\text{ML}_i/w_i \text{ is correct})P(w_i \text{ is correct})}{P(\text{ML}_i)}.\end{aligned} \tag{4}$$

$P(\text{ML}_i/w_i \text{ is correct})$ and $P(\text{ML}_i)$ are modeled with Gaussian probability density function (pdf). As mentioned in Section II-A, these pdfs and $P(w_i \text{ is correct})$ are computed with the evaluation database. Fig. 2 shows $P(\text{ML}_i/w_i \text{ is correct})$ and $P(\text{ML}_i)$.

C. Combining the Information From Several Word Features

Employing several word features should lead to more reliable confidence evaluations. If $\text{WF}_1, \text{WF}_2, \ldots, \text{WF}_J$ are the word features used in the confidence analysis, BBCM will correspond to

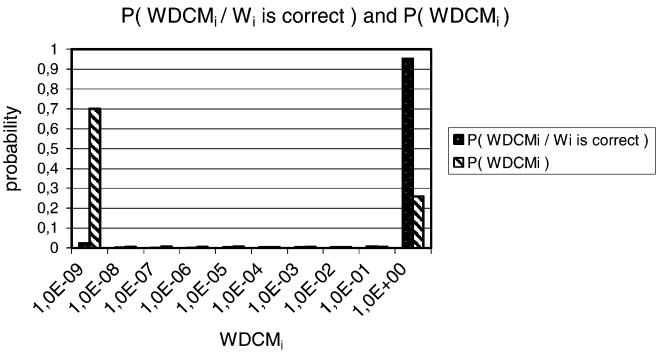

Fig. 1. *A priori* distribution probabilities with WDCM.

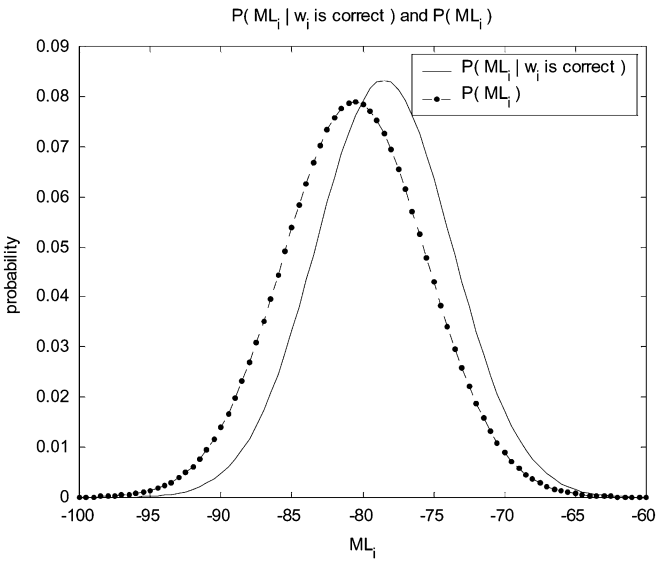

Fig. 2. *A priori* pdfs with ML_i.

$$\begin{aligned}&\text{BBCM}(\text{WF}_{1,i}, \text{WF}_{2,i}, \ldots, \text{WF}_{J,i}) \\ &\quad = P(w_i \text{ is correct}/\text{WF}_{1,i}, \text{WF}_{2,i}, \ldots, \text{WF}_{J,i}).\end{aligned} \tag{5}$$

Estimating the distributions $P(\text{WF}_{1,i}, \text{WF}_{2,i}, \ldots, \text{WF}_{J,i}/w_i$ is correct) and $P(\text{WF}_{1,i}, \text{WF}_{2,i}, \ldots, \text{WF}_{J,i})$ to compute (5) usually requires a high amount of data. To counteract this limitation, $\text{BBCM}(\text{WF}_{1,i}, \text{WF}_{2,i}, \ldots, \text{WF}_{J,i})$ could be approximated as

$$\begin{aligned}&\text{BBCM}(\text{WF}_{1,i}, \text{WF}_{2,i}, \ldots, \text{WF}_{J,i}) \approx \text{BBCM}(\text{WF}_{1,i}) \\ &\qquad \cdot \text{BBCM}(\text{WF}_{2,i}), \ldots, \text{BBCM}(\text{WF}_{J,i}).\end{aligned} \tag{6}$$

Approximation in (6) loses accuracy if $\text{WF}_1, \text{WF}_2, \ldots, \text{WF}_J$ are statistically dependent. However, despite the fact that the word features are not independent, (6) could still be considered a confidence metric. The combination of word features according to (6) was tested with WDCM_i and ML_i in this letter

$$\text{BBCM}(\text{WDCM}_i, \text{ML}_i) \approx \text{BBCM}(\text{WDCM}_i) \cdot \text{BBCM}(\text{ML}_i). \tag{7}$$

III. EXPERIMENTS

The approach proposed in this letter was tested with a Spanish database recorded on the telephone line. Users phoned to a ASR-based cinema enquiry system implemented with Galaxy II [8] at the Speech Processing and Transmission Lab., Universidad de Chile. The dialogue sequence was as follows: First, the system asked the user to choose one film from of a list composed of 80 films; second, the system prompted for the name and neighborhood of the cinema; and finally, the user had to say if he/she wanted to go to the cinema in the morning, afternoon, or evening. The ASR employed a language model based on trigrams and allowed the user to employ natural language to input the information required by the system. The vocabulary was composed of 221 words. The training database corresponded to 13 897 utterances. All of the training signals were employed to train the CDHMMs. The distributions $P(\text{WDCM}_i/w_i \text{ is correct}), P(\text{WDCM}_i)$, $P(\text{ML}_i/w_i \text{ is correct})$ and $P(\text{ML}_i)$ in (3) and (4) and the *a priori* probability $P(w_i \text{ is correct})$ were evaluated with 2826 evaluating utterances. N-best analysis was based on the ten best hypotheses $(N = 10)$ obtained from Viterbi algorithm. The testing database corresponded to 1036 utterances.

Thirty-three MFCC parameters per frame were computed: the frame energy plus ten static coefficients and their first and second time derivatives. Spectral subtraction and stochastic weighted Viterbi was applied as in [9]. Cepstral mean normalization (CMN) was also employed to reduce the channel distortion. Each triphone was modeled with a three-state left-to-right topology without skip-state transition, with eight multivariate Gaussian densities per state with diagonal covariance matrices. Recognized words (RW) are those contained in the N-best hypotheses with confidence measures above a given threshold. Testing words (TW) are those contained in the transcription of the testing utterances. As a consequence, false acceptation (FA) and false rejection (FR) errors were estimated as $(\text{N}^\circ \text{ of RW that were incorrect})/(\text{N}^\circ \text{ of RW})$ and $(\text{N}^\circ \text{ of TW that were not recognized})/(\text{N}^\circ \text{ of TW})$, respectively. Equal error rate (EER) is defined as the intersection of FA and FR curves. Results are presented in Tables I and II and Fig. 3. The system gave a WER equal to 13.1%.

IV. DISCUSSION AND CONCLUSION

As can be seen in Fig. 1, the range of WDCM_i was logarithmically divided in ten intervals from 10^{-9} to 1, where the last interval corresponds to $\text{WDCM}_i \leq 10^{-9}$. $P(\text{WDCM}_i/w_i \text{ is correct})$ and $P(\text{WDCM}_i)$ show a high concentration in the highest interval but are also fuzzily distributed when WDCM_i tends to zero. According to Fig. 2, $P(\text{ML}_i/w_i \text{ is correct})$ and $P(\text{ML}_i)$ are highly overlapped but their means are different, which in turn suggests that ML_i may still provide some useful information to asses the result of Viterbi decoding.

As shown in Table I, $\text{BBCM}(\text{WDCM}_i)$ (3) and the combination with $\text{ML}_i, \text{BBCM}(\text{WDCM}_i) \cdot \text{BBCM}(\text{ML}_i)$ (7), gave an EER that is 2% lower than the one provided by the ordinary WDCM (1). Moreover, as can be seen in Fig. 3, $\text{BBCM}(\text{WDCM}_i)$ dramatically improves the discrimination

TABLE I
EQUAL ERROR RATE WITH THE CONFIDENCE METRICS PROPOSED HERE

Confidence metrics	EER (%)
$WDCM_i$	10.46
$BBCM(WDCM_i)$	10.29
$BBCM(ML_i)$	58.50
$BBCM(WDCM_i) \cdot BBCM(ML_i)$	10.28

TABLE II
AVERAGE $|g(\text{WF})|$ WHERE $g(.)$ IS DEFINED AS IN (8), AND WF IS EQUAL TO WDCM_i AND $\text{BBCM}(\text{WDCM}_i)$

$\overline{\lvert g(WDCM_i)\rvert}$	$\overline{\lvert g[BBCM(WDCM_i)]\rvert}$
2.05	3.94

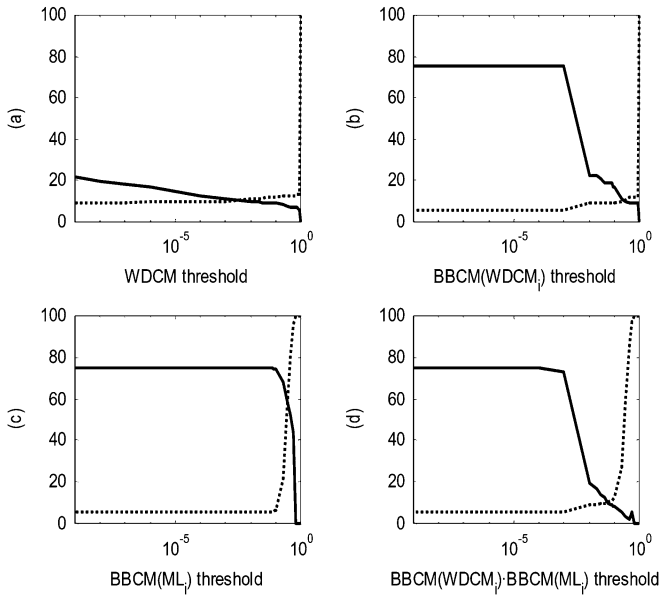

Fig. 3. FA (—) and FR (- - - -) curves: (a) WDCM. (b) BBCM with WDCM. (c) BBCM with ML. (d) $\text{BBCM}(\text{WDCM}_i) \cdot \text{BBCM}(\text{ML}_i)$.

of misrecognized words when compared to WDCM_i: The $\text{BBCM}(\text{WDCM}_i)$ threshold that increases the FA error within RW is much more distinguishable. This problem could be analyzed from the pattern recognition theory point of view, and the following discriminant function can be defined for this "two-category case"[10]:

$$g(\text{WF}) = P(w_i \text{ is correct}/\text{WF}) - P(w_i \text{ is not correct}/\text{WF}) \tag{8}$$

where WF is a given word feature. The higher the discriminant function, the lower the classification error rate. Table II shows the average $|g(\text{WF})|$, where WF is equal to WDCM_i and $\text{BBCM}(\text{WDCM}_i)$. As can be seen in Table II, the average $|g[\text{BBCM}(\text{WDCM}_i)]|$ is 92% higher than the average $|g(\text{WDCM}_i)|$. Consequently, $\text{BBCM}(\text{WDCM}_i)$ provides a discriminative ability 92% higher than the one given by WDCM_i to decide if w_i is correct or not.

When compared with WDCM_i and $\mathrm{BBCM}(\mathrm{WDCM}_i)$, $\mathrm{BBCM}(\mathrm{WDCM}_i)\cdot\mathrm{BBCM}(\mathrm{ML}_i)$ provides more symmetrical FA and FR curves, i.e., the resolution of the FR curve increases when the BBCM threshold is higher than 0.1. In other words, the estimation of a threshold as a function of a given FR rate should be more reliable. Nevertheless, the increase in the resolution of the FR curve could also be interpreted as a reduction in the discrimination ability, as discussed above. This is due to the fact that ML_i still gives some useful information about recognized words (see Fig. 2), although it is not a very discriminating confidence measure itself.

It is worth emphasizing that BBCM is a probability, which is interesting from the practical and theoretical point of view. Consequently, BBCM is more applicable to stochastic adaptation and OOV algorithms. Finally, the applicability of BBCM to other word features is suggested as future work.

ACKNOWLEDGMENT

The authors would like to thank Prof. R. Cole and Dr. B. Pellom from CSLR, Colorado University, for supporting the use of Galaxy II to develop dialogue systems.

REFERENCES

[1] T. J. Hazen, Th. Burianek, J. Polifroni, and S. Seneff, "Recognition confidence scoring for use in speech understanding systems," in *Proc. ISCA Tutorial Research Workshop*, Paris, France, 2000, pp. 213–220.

[2] C. H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proc. IEEE*, vol. 88, no. 8, pp. 1241–1267, Aug. 2000.

[3] M. Andorno, P. Laface, and R. Gemello, "Experiments in confidence scoring for word and sentence verification," in *Proc. ICSLP*, 2002, pp. 1377–1380.

[4] A. Stolcke, Y. König, and M. Weintraub, "Explicit word error minimization in N-best list rescoring," in *Proc. 5th Eur. Conf. Speech Communication Technology*, vol. 1, 1997, pp. 163–166.

[5] K. Y. Kwan, T. Lee, and C. Yang, "Unsupervised N-best based model adaptation using model-level confidence measures," in *Proc. ICSLP*, 2002, pp. 69–72.

[6] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA: The MIT Press, 1998.

[7] M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting," in *Proc. ICASSP*, Detroit, MI, 1995, pp. 297–300.

[8] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "Galaxy-II: A reference architecture for conversational system development," in *Proc. ICSLP*, Sydney, Australia, Nov. 1998, pp. 931–934.

[9] N. B. Yoma, I. Brito, and J. Silva, "Language model accuracy and uncertainty in noise canceling in the stochastic weighted Viterbi algorithm," *Proc. EUROSPEECH*, pp. 2193–2196, 2003.

[10] R. O. Duda, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.