Confidence Measures in Speech Recognition based on Probability Distribution of Likelihoods[♦]

Joel Pinto, R.N.V. Sitaram
HP Laboratories India

HPL-2005-144
August 10, 2005*

confidence measures, speech recognition, acoustic likelihood, duration, viterbi

In this paper, we propose two confidence measures (CMs) in speech recognition: one based on acoustic likelihood and the other based on phone duration. For a decoded speech frame aligned to an HMM state, the CM based on acoustic likelihood depends on the relative position of its output likelihood value in the probability distribution of likelihood value in that particular state. The CM of whole phone is the geometric mean of CMs of all frames in it. The CM based on duration depends on the deviation of the observed duration from the expected duration of the recognized phone. The two CMs are combined using weighted geometric mean to obtain a hybrid phone CM. The hybrid CM shows significant improvement over the CM based on time normalized log-likelihood score. On TI-digits database, at 20% false acceptance rate, the normalized acoustic log-likelihood based CM has a detection rate of 83.8% while the hybrid CM has a detection rate of 92.4%.

* Internal Accession Date Only

[♦] Interspeech 2005, Lisbon, Portugal, 4-8 September 2005

Confidence Measures in Speech Recognition based on Probability Distribution of Likelihoods

Joel Pinto, R. N. V. Sitaram

Hewlett Packard Labs India
Bangalore, India.
`{joel.pinto, sitaram}@hp.com`

Abstract

In this paper, we propose two confidence measures (CMs) in speech recognition: one based on acoustic likelihood and the other based on phone duration. For a decoded speech frame aligned to an HMM state, the CM based on acoustic likelihood depends on the relative position of its output likelihood value in the probability distribution of likelihood value in that particular state. The CM of whole phone is the geometric mean of CMs of all frames in it. The CM based on duration depends on the deviation of the observed duration from the expected duration of the recognized phone. The two CMs are combined using weighted geometric mean to obtain a hybrid phone CM. The hybrid CM shows significant improvement over the CM based on time normalized log-likelihood score. On TI-digits database, at 20% false acceptance rate, the normalized acoustic log-likelihood based CM has a detection rate of 83.8% while the hybrid CM has a detection rate of 92.4%.

1. Introduction

In any communication scenario using speech, either human to human or human to machine, the intelligibility of speech is an important factor. Poor articulation of speech, ambient noise etc, make speech less intelligible and thereby difficult to recognize. Humans overcome this problem by either asking the speaker to repeat or interpret speech using higher level knowledge like pragmatics and context. In the case of human to machine interaction, less intelligible speech can be dealt with confidence measures. Confidence measures assign a degree of confidence to the recognized words. Using confidence measures, the ASR could identify the words which are likely to be erroneous and the application using ASR could use corrective action.

The fundamental rule in statistical speech recognition is the Baye's rule given by:

$$\begin{aligned} W_{opt} &= \arg\max_{W} P(W \mid O) \\ &= \arg\max_{W} \frac{p(O \mid W).P(W)}{p(O)} \\ &= \arg\max_{W} p(O \mid W).P(W) \end{aligned}$$

The recognized word sequence W_{opt} is the one which maximizes the posterior probability $P(W \mid O)$, where $p(O \mid W)$ is the acoustic model, $P(W)$ is the language model and $p(O)$ is the unconditional acoustic likelihood of the observation sequence. While decoding, the unconditional acoustic likelihood $p(O)$ is normally omitted since it is invariant to the choice of a particular word sequence. As a result, the acoustic score $p(O \mid W)$ obtained during recognition will be unnormalized and cannot be used as a measure of confidence. Different approaches have been tried to approximate $p(O)$ to obtain the right confidence measure $P(W_{opt}|O)$. The unconditional acoustic likelihood $p(O)$ could be evaluated by summing up likelihoods given all speech models (generic catch-all models [1]). The likelihood $p(O)$ can also be approximated to the likelihood score obtained by doing a phone recognition of the same speech segment. In ASRs using lattice re-scoring, $p(O)$ could also be derived from the word lattice [2].

In many cases, the computational complexity in evaluating $p(O)$ could even match the complexity of actual recognition. In scenarios where this is undesirable, a meaningful confidence measure could be obtained from unnormalized $p(O \mid W)$ by dividing it by the number of frames [3][4]. Though this approach achieves time normalization, the CM depends on the decoded phones. In this paper, we still use the unnormalized acoustic likelihood of each frame, but normalize it using a priori knowledge of the distribution of the likelihood scores. We also propose a similar normalization scheme for the duration of the recognized phones.

The paper is organized as follows: In section 2, we discuss the CM based on time normalized log-likelihood score and analyze its shortcomings as a confidence measure. In section 3, we propose the new confidence measures. The experiments and the results are presented in section 4.

2. Time Normalized Log-Likelihood Score

The ASR returns the word sequence $W_{opt} = \{W_1 \dots W_K\}$ as well as the underlying HMM state sequence. The time normalized log-likelihood score C_{NLS} for the word W_k is given by:

$$C_{NLS}^{W_k} = \frac{1}{T_k} \log p(O^k \mid W_k) \tag{1}$$

$$p(O^k \mid W_k) = b_{s_1}(O_1) \cdot a_{s_1 s_2} \dots b_{s_{T_k}}(O_{T_k}) \tag{2}$$

where $O^k = (O_1 \dots O_{T_k})$ is the feature vector sequence and $(s_1 \dots s_{T_k})$ is the corresponding HMM state sequence, $b_{s_i}(O_t)$ is the output likelihood of O_t in state s_i and $a_{s_i s_j}$ is the transition probability from state s_i to state s_j. Ideally, the confidence measure should be indicative of the correctness of the decoded string and should not depend on actual phones/words decoded so that CMs can be compared across different words. However, the normalized log-likelihood score c_{NLS} depends on the decoded phone sequence. To illustrate this point, Fig. 1 shows the output probability density functions (pdf) in HMM state 1 of context independent phones /z/ (e.g, /z/ in zero) and /ai/ (e.g. /ai/ in nine). Please note that plot is for one dimension only. The following two cases explain the inefficiency of C_{NLS} as a measure of confidence:

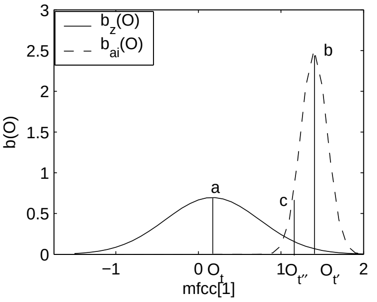

Figure 1: *Output pdfs in state 1 of phones /ai/ and /z/. The pdfs are plotted for the first component of the feature vector, $mfcc[1]$. The points a, b and c denote the output likelihood values $b_z(O_t)$, $b_{ai}(O_{t'})$ and $b_{ai}(O_{t''})$ respectively.*

Case-1: Let O_t and $O_{t'}$ be feature vectors exactly aligned to the mean of pdfs $b_z(O)$ and $b_{ai}(O)$ of state 1 of /z/ and /ai/ respectively. As $b_z(O_t) \ll b_{ai}(O_{t'})$, they contribute unequally to C_{NLS} in (1). Ideally, as O_t and $O_{t'}$ are the acoustic means in their respective states, their CMs should be equal.

Case-2: Let $O_{t''}$ be the feature vector aligned to state 1 in /ai/ and has a likelihood value $b_{ai}(O_{t''}) = b_z(O_t)$. As O_t and $O_{t''}$ have the same likelihood values, they contribute equally to C_{NLS} in (1). Ideally, O_t should have a higher CM as it is the acoustic mean in its state as opposed to $O_{t''}$ which is away from its mean.

In this paper, we explore a method to transform the output likelihood values using a priori knowledge of their distribution.

3. Proposed Confidence Measures

In this section, we propose two confidence measures, one based on acoustic likelihood value and the other based on phone duration. We also describe a method to combine the two CMs to obtain a hybrid confidence measure.

3.1. Acoustic Confidence Measure

In ASR, a triphone is normally modeled by an N-state left-right HMM. The output density in an HMM state is modeled as a Gaussian mixture. For state j in triphone i (hereafter denoted by $state(i,j)$), the output density $b_{ij}(O)$ is given by:

$$b_{ij}(O) = \sum_{k=1}^{M} w_{ijk}\mathcal{N}(\mu_{ijk}, U_{ijk}; O) \tag{3}$$

where, M is the number of mixtures, w_{ijk} is the k^{th} mixture weight and $\mathcal{N}(\mu, U; O)$ is the unimodal Normal density with mean μ, covariance matrix U and given by:

$$\mathcal{N}(\mu, U; O) = (2\pi)^{-\frac{N}{2}} |U|^{-\frac{1}{2}} \exp(-\frac{1}{2}(O-\mu)^T U^{-1}(O-\mu))$$

The ASR returns the recognized word sequence as well as the HMM state sequence. Suppose that the t^{th} speech frame O_t is aligned to $state(i,j)$ and has an output likelihood value of $b_{ij}(O_t)$, we define the new acoustic CM c_t^A for that frame as:

$$c_t^A = P[B_{ij} \leq b_{ij}(O_t)] \tag{4}$$

Where B_{ij} is a single dimensional random variable denoting the output likelihood value of feature vectors that are correctly aligned to $state(i,j)$. The CM c_t^A is the probability that the output likelihood value in $state(i,j)$ is lesser than the observed test vector likelihood $b_{ij}(O_t)$. The normalized CM c_t^A is a better measure of confidence than the CMs obtained directly from the output likelihood scores. B_{ij} is the transformed random variable $B_{ij} = b_{ij}(O)$. To simplify the notations, we drop the subscript ij in further discussions. Solving for $B = b(O)$ in (4) and moving to logarithmic domain, we get:

$$c_t^A = P[\log(b(O)) \leq \log(b(O_t))] \tag{5}$$

Assuming unimodal output density in (3), the acoustic CM in (5) reduces to (6) where R is the region of integration (ROI).

$$c_t^A = 1 - \int_R b(O')dO' \tag{6}$$

$$R = \{O : (O-\mu)^T U^{-1}(O-\mu) < (O_t-\mu)^T U^{-1}(O_t-\mu)\}$$

If the feature vector O_t is the acoustic mean of the state to which it aligns to i.e, $O_t = \mu$, then the ROI is the null set $\{\Phi\}$ and vector O_t has a maximum CM of $c_t = 1$.

$$R = \{O : (O-\mu)^T U^{-1}(O-\mu) < 0\} = \{\Phi\} \tag{7}$$

In Fig. 1, the feature vectors O_t and $O_{t'}$ are the acoustic means of state 1 in phones /z/ and /ai/ respectively. As explained above, both O_t and $O_{t'}$ will have the same confidence measure of $c_t^A = c_{t'}^A = 1$. Thus, the proposed CM has overcome the shortcomings of C_{NLS} as discussed in case-1 in section 2.

In practice, the output density in an HMM state will be multimodal and it is not possible to obtain a closed form solution for c_t^A in terms of $b(O)$. Hence we evaluate c_t^A from (5) as:

$$c_t^A = \int_{b'=-\infty}^{\log(b(O_t))} f_{B'}(b')db' \tag{8}$$

where $f_{B'}$ is the pdf of the transformed random variable $B' = log(B) = log(b(O))$ denoting the log-likelihood values of speech feature vectors when aligned to the correct HMM state. The pdf $f_{B'}(b')$ is estimated non-parametrically from the training data as explained in section 3.2.

Fig. 2 shows the non-parametrically estimated pdf of the output log-likelihood value in state 1 of phones /z/ and /ai/. The points a, b and c on the x-axis correspond to the output likelihood values a, b and c of vectors O_t, $O_{t'}$ and $O_{t''}$ as shown in Fig. 1. Integration in (8) can be interpreted as the area under a pdf curve. It is clear from the figure that O_t and $O_{t'}$ will have equal CMs of $c_t^A = c_{t'}^A = 1$ as the total area under a pdf is always unity. The CM for the feature vector $O_{t''}$ is the area under $f_{B'_{ai}}(b')$ from $b' = -\infty$ to the point c which is less than unity. It should be noted that $c_{t''}^A < c_t^A$ even though the log-likelihood values are the same for $O_{t''}$ as well as O_t. This explains how the proposed CM handles case-2 explained in section 2.

The confidence measure for a phone is computed as the geometric mean of the CMs of the speech frames in the phone. If a phone p has T_p frames, its acoustic CM, C_p^A is given by:

$$C_p^A = \exp(\frac{1}{T_p}\sum_{t=1}^{T_p} \log c_t^A) \tag{9}$$

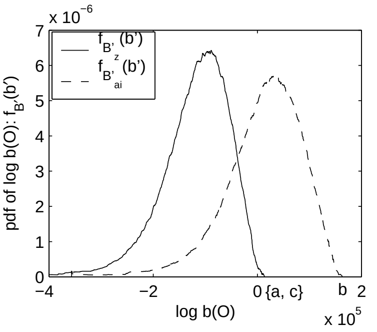

Figure 2: *The pdf of log-likelihood (log base = 1.0001) scores in state* 1 *of phones /ai/ and /z/. The points* a, b *and* c *on the x-axis denote the log-likelihoods* $\log b_z(O_t)$, $\log b_{ai}(O_{t'})$ *and* $\log b_{ai}(O_{t''})$ *respectively*

3.2. Non-Parametric pdf estimation

If the output pdf in the HMM state is unimodal Gaussian i.e, M=1 in (3), it can be shown that B' will have a Gamma distribution. However, in the multimodal case, we cannot assume a parametric form for $f_{B'}$. Non parametric methods of pdf estimation are useful when there is no a priori knowledge of the underlying distribution. Parzen window [5] method is a kernel based non parametric pdf estimation method. In this method a kernel function $\varphi_{x_k}(y)$ (Rectangular, Gaussian etc) is generated around each data point x_k in the training set and the pdf is evaluated by adding these kernel functions and scaling the sum.

$$f(y) = \frac{1}{V}\sum_{k=1}^{K}\varphi_{x_k}(y) \tag{10}$$

The training data is force aligned to its correct transcript using the Viterbi algorithm [6]. Suppose the feature vectors $O^1 \ldots O^K$ are aligned to a particular state in a triphone and each frame has a log-likelihood score of $\log(b(O^k))$, $k = 1, \ldots K$, the pdf $f_{B'}(b')$ is given by:

$$f_{B'}(b') = \frac{1}{Kh}\sum_{k=1}^{K}\varphi(\frac{b' - \log(b(O^k))}{h}) \tag{11}$$

where, $\varphi(u)$ is the rectangular kernel function given by (12) and h is the scaling factor.

$$\varphi(u) = \begin{cases} 1 & \mid u \mid \leq 1/2 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

3.3. Duration Confidence Measure

Different approaches have been tried in the past to use phone duration as a feature in utterance verification [7]. Let D be the discrete random variable denoting the phone duration (in terms of number of frames), $p_D(n)$ the probability mass function (pmf) of D and μ_D the expected duration of phone. Suppose d is the observed duration of the recognized phone, the new confidence measure is based on the deviation ($d' = |d - \mu_D|$) of the observed duration from the expected duration as opposed to directly using $p_D(n = d)$. If the random variable $D' = |D - \mu_D|$ denotes the deviation, we define the duration based CM as:

$$\begin{aligned} C_p^D &= P[D' \geq d'] = 1 - P[D' < d'] && (13)\\ &= 1 - P[\mu_D - d' < D < \mu_D + d'] && (14)\\ &= 1 - \sum_{n=\lceil \mu_D - |d-\mu_D| \rceil}^{\lfloor \mu_D + |d-\mu_D| \rfloor} p_D(n) && (15) \end{aligned}$$

Closer the value of observed duration to its expected duration, the higher is the duration confidence measure of that phone. To evaluate the duration pmf of each triphone, we force align the training data to its correct transcript and obtain the histogram of the phone duration. The histogram is then smoothened and normalized to get the pmf $p_D(n)$.

3.4. Combination of Confidence Measures

We use weighted geometric mean to combine the acoustic and duration CMs to obtain a hybrid phone confidence measure.

$$C_p = \exp(w_a \log(C_p^A) + (1 - w_a)\log(C_p^D)) \tag{16}$$

Where $w_a (0 \leq w_a \leq 1)$ is the acoustic CM weight factor. The word confidence measure C_W is the geometric mean of the hybrid phone confidence measures of the constituent phones.

4. Experiments and Results

To test the proposed confidence measures, we conducted experiments using open source speech recognition toolkit Sphinx-Train and Sphinx-3 flat decoder [8]. In this section, we discuss the experiments and the relevant results.

4.1. Database

Experiments were conducted on speaker independent continuous digit recognition task. The vocabulary consisted of 11 words (ten digits and oh). TI-digits [9] database was used for training the ASR as well as evaluating the performance of the confidence measures. The training set consisted of 12549 digit utterances from 163 speakers and the test set consisted of 12547 utterances from 163 speakers

4.2. ASR System

Mel Frequency Cepstral Coefficients (MFCCs) were used as features for speech recognition. The speech signal sampled at 16 kHz is frame blocked with a window length of 20 msec and frame shift of 10 msec. The 13-dimension MFCC vector, delta coefficients and delta-delta coefficients form a 39-dimensional feature vector. Triphone is used as the basic speech modeling unit, modeled by a 5-state left-right HMM. The output density in each state is modeled as mixture of 4 Gaussians. The word error rate of the ASR was 3.1% while the sentence error rate was 8.0%

4.3. Confidence Measure Results

To evaluate the performance of the confidence measures, recognized words are compared against the correct transcription of the utterance and each word is classified as *correct* or *wrong*. The confidence measures of hypothesized digits are compared against a threshold to either *accept* or *reject* the digit. Receiver Operator Characteristics (ROC) [10] - plot of the detection rate versus the false acceptance rate - is plotted by varying the confidence threshold between 0 and 1. A confidence measure is good if it has higher detection rates at lower false acceptance rate.

The proposed confidence measures are tested for its efficiency in detecting putative errors [10] (erroneous but in-vocabulary words) as well as Out-Of-Vocabulary (OOV) words. Table 1 compares the performance of different confidence measures in rejecting the putative errors. The proposed confidence measures have out performed the baseline confidence measures. Also, the hybrid CM ($w_a = 0.8$) based on acoustic likelihood as well as phone duration has performed better than the CM based only on acoustic likelihoods ($w_a = 1.0$).

Table 1: *Performance of the CMs in rejecting putative errors. Detection rates for false acceptance rates of 30, 20 and 10%*

Confidence Measure	*Word Detection Rate*		
	30%	*20%*	*10%*
Baseline Acoustic CM (C_{NLS})	90.9	82.3	67.2
Proposed Acoustic CM ($w_a = 1$)	93.5	90.0	81.4
Proposed Hybrid CM ($w_a = 0.8$)	94.5	91.8	82.5

To evaluate the performance of the proposed CMs in detecting the OOVs, we simulate recognition errors in the following manner. For every digit (*oh, zero, ... nine*), we decode utterances containing the digit using language models not containing the particular digit. Table 2 shows the overall performance of the CMs in detecting OOVs. The results clearly indicate that the hybrid method has the best performance.

Table 2: *Performance of the CMs in rejecting OOVs. Detection rates for false acceptance rates of 30%, 20% and 10%*

Confidence Measure	*Word Detection Rate*		
	30%	*20%*	*10%*
Baseline Acoustic CM (C_{NLS})	91.6	83.8	68.5
Proposed Acoustic CM ($w_a = 1$)	94.2	90.4	81.3
Proposed Hybrid CM ($w_a = 0.8$)	95.4	92.4	84.2

Fig. 3 is the plot of ROC curves of the baseline method C_{NLS} and the proposed confidence measures in detecting the OOVs. A higher ROC curve indicates a better confidence measure. The goodness of the confidence measure is given by the figure-of-merit (FOM) which is the area under the ROC curve. The optimum acoustic weight w_a in (16) is obtained by empirically plotting (Fig. 4) the FOM of the hybrid CM for different values of w_a. The hybrid CM with $w_a = 0.8$ has the best performance. This is consistent with existing knowledge that acoustic likelihood scores are more reliable than duration as a feature for confidence measure.

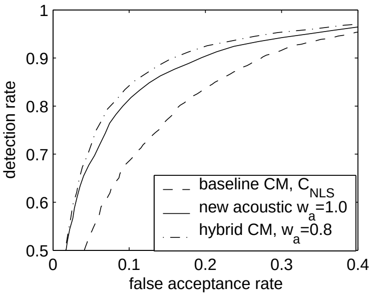

Figure 3: *ROC curve for the normalized likelihood score C_{NLS} and proposed confidence measures for $w_a = 1.0$ and $w_a = 0.8$*

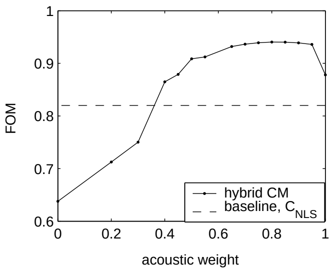

Figure 4: *The figure-of-merit (FOM) of the hybrid CM as a function of acoustic weight w_a.*

5. Conclusion

In this paper we have investigated two confidence measures, one based on transforming the output acoustic likelihood and another based on the duration of recognized phone. We have also explored a method to combine these confidence measures to obtain a hybrid confidence measure. In our experiments, the duration based confidence measure depends on the deviation of the observed phone duration from the mean (expected) duration. However, if the pmf of the duration is not symmetric, the mode may be a better statistic for computing the deviation than the mean. Work is being carried out to explore this further.

6. References

[1] Kamppari, S.O. and Hazen, T.J., “Word and Phone Level Acoustic Confidence Scoring”, ICASSP, 3:1799–1802, 2000.

[2] Wessel, F., Schluter, R. Macherey, K. and Ney, H., “Confidence Measures for Large Vocabulary Continuous Speech Recognition”, IEEE Trans. on Speech and Audio Proc., 9(3):288–298, 2001.

[3] Jia, B., Zhu, X., Luo, Y. and Hu, D., “Utterance Verification using Modified Segmental Probability Model”, Proceedings of EuroSpeech, 45–48, 1999.

[4] Charlet, D., “Optimizing Confidence Measure based on HMM Acoustical Rescoring”, Automatic Speech Recognition: Challenges for the new Millenium, ISCA Tutorial and Research Workshop, 203–206, 2000.

[5] Duda, R.O. and Hart. P.E., Pattern Classification and Scene Analysis, Wiely Publishers, 1973.

[6] Rabiner, L. A., “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, Proc. of IEEE, 77(2):256–286, 1989.

[7] Koo, M.-W., Lee, C.-H., and B.-H., Juang, “Speech Recognition and Utterance Verification Based on Generalized Confidence Score”, IEEE Trans. Speech and Audio Proc., 9(8): 821–832, 2001.

[8] “The CMU Sphinx Group Open Source Speech Recognition Engines” http://www.speech.cs.cmu.edu/sphinx/

[9] Leonard, R.G., “A Database for Speaker Independent Speech Recognition”, ICASSP, 9:328–331, 1984.

[10] Rahim, M.G., Lee, C.-H. Lee and Juang, B.-H., “Discriminative Utterance Verification for Connected Digits Recognition”, IEEE Trans. on Speech and Audio Proc., 5(3):266–277, 1997.