Reliability Evaluation of Speech Recognition in Acoustic Modeling*

Bin Jia¹, Xiaoyan Zhu², Yupin Luo¹, and Dongcheng Hu¹
¹Dept. of Automation, Tsinghua University, Beijing 100084, P. R. China
²State Key Laboratory of Intelligent Technology and Systems
Dept. of Computer Science and Technology, Tsinghua University, Beijing 100084, P. R. China
Email: zxy-dcs@mail.tsinghua.edu.cn

ABSTRACT

In this paper, we present a new method for reliability evaluation of speech recognition in acoustic modeling. The new method incorporates the Integrated Model (IM), which is trained by all the speech data. For close-set verification, the IM method has theoretically the lowest equal error rate; and for open-set, the method often performs well. At the same time, it costs much less computation than other verification methods commonly used. Experiment results show the method is feasible.

1. INTRODUCTION

The reliable confidence measure is playing more and more important role in the broad applications of speech recognition. In many occasions, such as dictation, Utterance Verification [1] or UV, which is based on reliability measure, is used to decide whether to accept or to reject the recognition results. Dialog systems and speech-commanding ones often need UV to verify the keywords. As for unsupervised training, reliable confidence measure is crucial in deciding whether an utterance can be used to train a transcription or unit model. From these important applications (they are only part of its usage), we are sure reliable confidence measure of speech recognition deserves lots of research interest.

Like speaker verification, UV is usually based on model or unit likelihood score. From the point of likelihood score, UV methods are fundamentally divided into two kinds: verification on direct model score and verification on likelihood ratio. The former methods are simple and usually compare the model scores directly with thresholds determined in training stage. In practice, these methods have poor discriminability and it is difficult to calculate the effective thresholds. The latter kind compares with some threshold the ratio of keyword model likelihood to some other model likelihood and includes many methods, such as filler model method, antiword one, Cohort Set antiword one, and so on. The ratio methods usually outperform the direct likelihood ones, however with much larger computation. Antiword method and its simplification — Cohort Set method — are prevailing and in this paper, we use them as the baselines in experiments.

In this paper, we present the IM method. It is one of likelihood ratio methods in form, however comes from the well-known Bayes posterior probability. For close set, the IM method has theoretically the lowest equal error rate, which is proved in the paper; and for open set, the method often performs well. In experiments, it performs not worse than antiword method and much better than Cohort Set method. For another strength, the method is very timesaving in training and verification stages.

2. RELIABILITY EVALUATION UNDER BAYESIAN PROBABILITY EQUATION SCHEME

Most of likelihood ratio methods are based on Neyman-Pearson Lemma [2] in hypothesis testing theory. The lemma gives the optimal decision rule as following:

\[
T(Y, w) = \begin{cases} 
1 & \frac{p(Y \mid H_0)}{p(Y \mid H_1)} > \eta, \quad \text{accept} \\
0 & \frac{p(Y \mid H_0)}{p(Y \mid H_1)} \leq \eta, \quad \text{reject}
\end{cases}
\]  

(1)

¹ This project is supported by the Chinese National Natural Science Foundation (69982005) and “973” (G199803050703).
in which $H_o$ and $H_i$ stand for the hypothesis that utterance $Y$ is produced by the claimed transcription $w$ and the hypothesis that $Y$ is not produced by the transcription respectively, $p(Y \mid H_o)$ and $p(Y \mid H_i)$ are the probability density function, pdf, or probability function of $Y$ on the condition of $H_o$ and $H_i$ respectively, and $\eta$ is the threshold. However, because of the uncertainty of the actual model pdf, the pdfs in use are approximate estimations. The estimations of $p(Y \mid H_i)$, pdf of the alternative hypothesis, are usually different among methods. Typical among them is the antiword method. One of its log-likelihood versions is

$$
\log T(Y, w) = \log p(Y \mid w) - \frac{1}{M} \sum_{i=1}^{M} \log \{P(Y \mid w_i)\}, \\
\text{w} \neq w_i, i = 1, \ldots, M
$$

in which, there are $M$ models besides $w$ and $p(Y \mid \cdot)$ is pdf of $Y$ on condition of one model. The subtrahend is the approximation of the logarithm of $p(Y \mid H_i)$. The equation arrogantly requires all the models have the same prior probability $(1/M)$, which is not reasonable.

Moreover, the computation cost is conspicuous—there are $M+1$ likelihood scores to be calculated. The approximate computations are still very large, although some techniques are taken, such as the adaptation of Cohort Set, which is equivalent to decreasing $M$ and therefore decreasing the performance dramatically. Can we seek an effective method that keeps good performance and costs less computation at the same time?

To cope with the problem, we can think of the essential of verification. In fact, the confidence measure in verification is the Bayes posterior probability $p(w \mid Y)$, which can determine how sure we are of the transcription $w$ on the condition of utterance $Y$. We here have the notation $\overline{w}$ for the transcription of the set of models not including $w$. Then, the well-known Bayes equation is expressed as:

$$
p(w \mid Y) = \frac{p(Y \mid w)p(w)}{p(Y)} = \frac{p(Y \mid w)p_w}{p(Y \mid w)p_u + p(Y \mid \overline{w})p_\overline{w}} \quad (2)
$$

in which, $p(Y \mid w)$ and $p(Y \mid \overline{w})$ are the probabilities of utterance $Y$ on the condition of the transcription $w$ and $\overline{w}$ respectively. $p_u$ and $p_\overline{w}$ are the priors of $w$ and $\overline{w}$ respectively and include the language modeling information. There are some relationship between Formulation (1) and Equation (2). Actually, we can find that $p(Y \mid w)$ is $p(Y \mid H_o)$ and $p(Y \mid \overline{w})$ is $p(Y \mid H_i)$. In the conventional likelihood ratio methods, a keyword has its own anti-part likelihood $p(Y \mid H_i)$, which is approximately calculated with large computation load in training and verification stages. Whereas, our idea is to try to find a likelihood easy to calculate, and as for the simplicity of system, it should better be common in or shared by all the keywords. Actually, $Y$ is uttered either by $w$ or by $\overline{w}$. If we integrate $w$ and $\overline{w}$ into a union transcription which are sure to utter $Y$, then all keywords can have such counterpart. Corresponding to the general antiword model, the union transcription model here is named as Integrated Model or IM. Any utterance is uttered by IM. So,

$$
p(Y \mid \lambda_w) = p(Y) \quad (3)
$$

In Equation (2), $p_u$ is constant in fixed application domain. Then, we can simplify the confidence measure from $p(w \mid Y) = \frac{p(Y \mid w)p_w}{p(Y)}$ to

$$
C(Y, w) = \frac{p(Y \mid w)}{p(Y)} = \frac{p(Y \mid w)}{p(Y \mid \lambda_w)} \quad (4)
$$

The IM verification technique is as following:

$$
\log C(Y, w) = \log p(Y \mid w) - \log p(Y \mid \lambda_w) \begin{cases} > \varepsilon, & \text{accept} \\ \leq \varepsilon, & \text{reject} \end{cases}
$$

(5)

3. EFFECTIVENESS OF IM METHOD

We can prove that in the situation of close-set verification, verification with formulation (5) performs best—with the lowest equal error rate.

Close-set verification means that the testing data come from some keywords involved in training stage (It should not be considered that verification data is from training data set). In
this situation, we can first prove the IM method has the lowest FA rate among all methods given the upper boundary of FR rate, then prove that IM method has the lowest equal error rate. FA rate can be defined as:

$$P_{fa} = E(q(Y)|H_0) = \int q(Y)p(Y|H_0)dY$$  \hspace{1cm} (6)

where $q(Y) \ (0 \leq q(Y) \leq 1)$ is the acceptance probability based on utterance $Y$. Likely, FR rate can be defined as:

$$P_{fr} = 1 - E(q(Y)|H_0) = 1 - \int q(Y)p(Y|H_0)dY$$  \hspace{1cm} (7)

The IM method with formulation (5) is actually:

$$q(Y) = \begin{cases} 1, & \text{when } p(Y|w) > \theta p(Y|\lambda_{\text{nl}}) \\ 0, & \text{when } p(Y|w) \leq \theta p(Y|\lambda_{\text{nl}}) \end{cases}, \hspace{0.5cm} \theta \geq 0$$  \hspace{1cm} (8)

Now, our first step is to prove: for closed set, if the IM method’s FR rate $P_{fr}$ is equal to $\nu$, $\nu \geq 0$, then its FA rate is the lowest among all the verification methods whose FR rate is not larger than $\nu$.

**Proof:** suppose there exists another verification method whose acceptance probability is $q^\prime(Y)$ and FA rate is $P^\prime_{fa}$.

(1) When $\theta < \frac{1}{P_w}$.

Since $P_{fr} \geq P^\prime_{fr}$,

$$P_{fr} - P^\prime_{fr} = \int \left( (1 - q(Y))p(Y|w)dy - (1 - q^\prime(Y))p(Y|w)dy \right)
\geq \int \left( (q(Y) - q^\prime(Y))p(Y|w)dy \right) \geq 0$$  \hspace{1cm} (9)

If $p(Y|w) > \theta p(Y|\text{AIM})$ or $p(Y|\bar{w}) \leq \frac{1-\theta}{\theta} p_Y p(w)$, then $q(Y) = 1$, and $q(Y) - q^\prime(Y) \geq 0$; else $q(Y) - q^\prime(Y) < 0$. Then,

$$\int (q(Y) - q^\prime(Y)) \left( p(Y|\bar{w}) - \frac{1-\theta}{\theta} p_{\bar{w}} \right) dY \leq 0$$  \hspace{1cm} (10)

From Expression (9) and (10),

$$P_{fr} - P^\prime_{fa} = \int \left( (q(Y) - q^\prime(Y))p(Y|\bar{w})dy \right) \geq \frac{1-\theta}{\theta} \int (q(Y) - q^\prime(Y))p(Y|w)dy \geq 0$$

Thus, $P^\prime_{fa} \geq P_{fa}$.

(2) When $\theta \geq \frac{1}{p_w}$.

Since $p(Y|\lambda_{\text{nl}}) = p(Y|w)p_w + p(Y|\bar{w})p_{\bar{w}}$, $q(Y) = 0$.

Hence, $P_{fr} = 1$ and $P^\prime_{fa} \geq P_{fa} = 0$.

From (1) and (2), it is obvious that IM technique has the lowest equal error rate for close set verification. [end of the first proof]

Based on the above, second step is to prove IM method has the lowest equal error rate.

**Proof:** Suppose verification method $V$ has lower equal error rate than IM method and $V$’s FR rate is $r$. Adjust the verification threshold of IM method so that the FR rate of IM method is also $r$. According to the above step: when FR rate is $r$, the FA rate of IM method is not larger than that of $V$, thus smaller than the IM’s FA rate at its equal error rate point. Now, we find that when the IM’s FR rate is $r$, both FR rate and FA rate are smaller than FR rate and FA rate at its equal error point. This opposes to the fact that with the adjustment of the threshold in a verification method, the FA rate and FR rate change oppositely. So, IM method has the lowest equal error rate. [end of the second proof]

The above conclusion is entirely valid when IM stands for all the utterance. Although limitation of the model structure and training results in modeling inaccuracy to some extent, the close-set verification can be considered like this valid situation. Unfortunately, as for the open-set verification, in which IM is not trained by the utterances of some non-keywords, and utterances of the non-keywords might appear in verification stage, the performance is not always so good. However, in fact, since the IM is trained with more uterance than arbitrary keyword model $w$, thus often have stronger ability to model arbitrary utterance. Then, for a non-keyword utterance $Y$, the likelihood of IM is often larger than that of $w$:

$$p(Y|\lambda_{\text{IM}}) \geq p(Y|w)$$ or $$\frac{p(Y|w)}{p(Y|\lambda_{\text{IM}})} \leq 1.$$

This means that a fixed threshold can be determined. So, IM method is usually effective in open-set verification.
4. EXPERIMENTS

Primitive experiments are done on Mandarin digit verification, which has wide applications, such as digit dialing, speech commanding, and so on.

The acoustic feature includes 16 LPC-CEPT, 16 delta LPC-CEPT, 1 log energy, and 1 delta log energy. The acoustic model is the modified segmental probability [3]. As for the digit keywords, each acoustic model has 3 states and 8 Gaussian mixtures per state. The models are trained with utterance from 102 people (62 males and 40 female). The IM acoustic model has 4 states and 16 mixtures per state. IM model is trained with all the training data of keywords. As for open set verification, some utterances of non-digit syllable are randomly chosen and tested. Cohort antiword technique is involved as the baseline. The close-set and open-set results are as following.

<table>
<thead>
<tr>
<th>Digit</th>
<th>Close set</th>
<th>Open set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cohort %</td>
<td>IM %</td>
</tr>
<tr>
<td>0</td>
<td>6.86%</td>
<td>4.90%</td>
</tr>
<tr>
<td>1</td>
<td>1.96%</td>
<td>0.98%</td>
</tr>
<tr>
<td>2</td>
<td>5.88%</td>
<td>2.94%</td>
</tr>
<tr>
<td>3</td>
<td>5.88%</td>
<td>3.92%</td>
</tr>
<tr>
<td>4</td>
<td>3.92%</td>
<td>1.96%</td>
</tr>
<tr>
<td>5</td>
<td>0.98%</td>
<td>0.98%</td>
</tr>
<tr>
<td>6</td>
<td>4.90%</td>
<td>4.90%</td>
</tr>
<tr>
<td>7</td>
<td>7.84%</td>
<td>3.92%</td>
</tr>
<tr>
<td>8</td>
<td>5.88%</td>
<td>3.92%</td>
</tr>
<tr>
<td>9</td>
<td>0.98%</td>
<td>2.94%</td>
</tr>
<tr>
<td>Average</td>
<td>4.70%</td>
<td>2.94%</td>
</tr>
<tr>
<td></td>
<td>5.68%</td>
<td>3.72%</td>
</tr>
</tbody>
</table>

Table: Equal error rates of Mandarin digit verification

The results show that IM technique performs better than Cohort technique. In close-set verification, the average improvement is 37.4%; and in open-set verification, the average improvement is 34.5%.

The IM technique in open set (average equal error rate: 3.72%) performs a little worse than itself in close set (average equal error rate: 2.94%). One of the major reasons is that IM model is trained only with the keyword utterance and so it doesn’t properly depict some non-keyword utterances in open set. One of the future efforts is to gather some typical non-keyword utterance to train the IM. With better modeling of non-keyword utterance, IM technology is expected to perform better.

As for the average verification time, the Cohort method requires 55 ms, while IM method requires 15 ms, only 27% of Cohort. It is obvious that the IM is much faster in computation or more timesaving than Cohort method, as is important in online application.

5. CONCLUSION

IM is a model for all utterances. Under the scheme of the famous Bayesian probability equation, IM is used in utterance verification. IM method outperforms convention verification techniques in two aspects: 1) Theoretically, it has the lowest equal error rate, which is proved in the paper. 2) It cost less computation than the conventional technologies. So, IM method is feasible in practice.

How to model IM is important to its highly effective applications. The key is to gather proper non-keyword utterances for training, which needs more study.

REFERENCES

