Confidence Scoring for Accurate HMM-based Word Recognition
By Using SM-based Monophone Score Normalization

Takaharu Sato, Muhammad Ghulam, Takashi Fukuda, Tsuneo Nitta
Graduate School of Engineering, Toyohashi University of Technology
1-1 Hibariga-oka, Tempaku, Toyohashi, JAPAN
E-mail: nitta@tutkie.tut.ac.jp

ABSTRACT

In this paper, we propose a novel confidence scoring method that is applied to N-best hypotheses output from an HMM-based classifier. In the first pass of the proposed method, the HMM-based classifier with monophone models outputs N-best hypotheses and boundaries of all the monophones in the hypotheses. In the second pass, an SM-based verifier tests the hypotheses by comparing confidence scores. We discuss how to convert a monophone similarity score of SM into a likelihood score, how to normalize the variations of acoustic quality in an utterance, and how to combine an HMM-based likelihood of word level and an SM-based likelihood of monophone level. In the experiments performed on speaker-independent word recognition, the proposed confidence scoring method significantly improves correct word recognition rate from 95.3% obtained by the standard HMM classifier to 98.0%.

1. INTRODUCTION

In typical HMM-based speech recognition systems, an input utterance $x$ is converted into a word $w$ (or a sequence of words) by evaluating the probability score $P(w|x) = P(x|w)P(w)/P(x)$ and, in the usual case, $P(x)$ is omitted because it is assumed to be invariant over an utterance. However, because many factors affect the acoustic quality in an utterance, various confidence scoring methods to verify an utterance for improving word recognition accuracy or for detecting keywords and/or unknown words have been proposed. The proposed confidence measures include the likelihood ratio of $P(x|w)/P(x|p)$ [1], where $P(x|p)$ is the likelihood of phoneme level, sub-word level verification scoring [2], and application of multiple features [3].

In this paper, we attempt to normalize the variations of acoustic quality in an utterance by applying the monophone-based Sub-space Method (SM) [4]. In an HMM scheme, speech events are represented by stochastic transition networks, and time variation of acoustic features in a state is simplified as a set of piecewise uniform regions, even if the variation is complicated. Moreover, likelihood scores of sub-words are accumulated over an utterance, and the classification result is output according to the accumulated score without checking the phones that the utterance consists of. On the other hand, SM can represent variation of fine structures in sub-words into a set of eigen vectors [5], however, the method needs accurate sub-word boundaries and a procedure to convert a similarity score into a likelihood score.

In the proposed method, an HMM-based classifier with monophone models calculates both N-best hypotheses and boundaries of all the monophones in the hypotheses, then an SM-based verifier tests the hypotheses. In hypotheses testing, firstly similarity between the input pattern of monophone with a fixed point and an eigen vector set is calculated by using SM after re-sampling each monophone interval. Secondly, the similarity score $S$ is converted into likelihood score $l_{SM}$ by using the maximum similarity normalization method [6]. Finally, after likelihood normalization of acoustic quality that is described in section 2.4 in detail, the normalized confidence scores of all the hypotheses are compared by combining an HMM word-level likelihood and an accumulated likelihood of $l_{SM}$. Feature extractions for the SM-based verifier are also discussed.

This paper is organized as follows. Section 2 outlines the system configuration and discusses the proposed confidence scoring, then section 3 describes the experimental setup and the results.

2. SYSTEM OVERVIEW

Figure 1 shows a block diagram of the proposed two-pass word-utterance verification system. The system is divided into three parts: the feature extractor, which converts the input speech into two types of acoustic features, one of which is fed into the HMM-based classifier and the other into the SM-based verifier; the HMM-based classifier, which executes the first pass and outputs N-best hypotheses (word candidates) and all the intervals of monophones in the hypotheses; and the SM-based verifier, which performs the second pass and tests the hypotheses through the confidence score normalization.
2.1 HMM-based Classifier (Baseline)

The HMM-based classifier in this paper adopts a standard monophone-based HMM with 5-states 3-loops left-to-right models, 38 standard MFCC parameters, and diagonalized Gaussian mixtures (mixture = 8).

The classifier estimates not only the N-best hypotheses but also the boundaries of each monophone in the hypotheses by a back-tracking procedure.

2.2 Sub-space Method (SM)

The Sub-space Method (SM) [4], or Multiple Similarity Method (MSM), incorporates variations in fixed-dimensional patterns of class c into an eigen vector set \( \phi_{cm} \), or an orthogonalized codebook, by using KLT from a learning database. The multiple similarity score \( S_c \) of class c between the codebook \( \phi_{cm} \) and a normalized input pattern \( x \) is defined as follows:

\[
S_c = \sum_{m=1}^{M} (x \cdot \phi_{cm})^2
\]  

where, \( \cdot \) denotes inner product and \( M \) is the number of eigen vectors.

One of the authors previously showed that local features extracted by using 3x3 derivative operators along the time axis and frequency axis significantly reduced phonetic segment classification errors [7]. Here, we apply the same types of local features to the SM-based verifier, but extract them by using linear regression calculation along the time axis and frequency axis [8].

After linearly re-sampling 10 frames of local features between the monophone boundaries, the verifier calculates multiple similarity \( S_c \) between the codebook \( \phi_{cm} \) and a normalized input pattern \( x \), and then converts the similarity into posteriori probability described in the next section.

2.3 Conversion of Similarity into a posteriori Probability

Figure 2.1 shows probability \( P(S|p) / P(S) \) observed in real monophone speech data. Here, \( S \) and \( p \) are the multiple similarity and monophone, respectively.

Firstly, \( P(S|p) / P(S) \) is modeled by the following equation:

\[
P(S|p) / P(S) = A B^S, \quad A>0, \quad B \geq 1
\]  

Next, the model is simplified as shown in Figure 2.2. Considering \( \log_B{A}=1 \), likelihood score \( l_{SM} \) is given by the following maximum similarity normalization procedure [6]:

\[
\log[P(S|p) / P(S)] = S - S_{\text{max}} \quad S \leq S_{\text{max}}
\]

\[
\log[P(S|p) / P(S)], \quad 0 \quad S > S_{\text{max}}
\]

2.4 Word-level Confidence Scoring and Estimation of \( P(S|p)/P(S) \) of Monophone

The confidence score (CS) in the word-utterance verifier is calculated by the following equation:

\[
CS = \alpha L_{\text{HMM}} + (1- \alpha) L_{SM}
\]

\[
\quad = \alpha L_{\text{HMM}} + (1- \alpha) \sum_{j} L_{SM}(j)
\]

where, \( \alpha \) and \( L_{\text{HMM}} \) are weighting coefficient and the word-level likelihood output by the HMM classifier, and \( L_{SM} \) and \( J \) is an accumulated likelihood over all the monophones in
each hypothesis ($\Sigma (S(j) - S_{\text{max}})$) and the number of monophones in a hypothesis, respectively.

Figure 3.1 shows an example of confidence scoring in which an utterance is [ro:do:] (labor). The HMM classifier outputs [kodomo] (child) for the best hypothesis and [ro:do:] is the second best. In the figure, two sequences of likelihood corresponding to two monophone strings given by the SM verifier are plotted, and two accumulated scores $L_{\text{SM}}$ show the change of rank. Figure 3.2 compares two scores, $L_{\text{HMM}}$ and $CS = \alpha L_{\text{HMM}} + (1 - \alpha)L_{\text{SM}}$.

The acoustic quality of monophones in an utterance is influenced by many factors such as breathing, accentuation, speaking rate, and so forth. Here, we propose a simple but effective normalization method to solve the degradation phenomena at the monophone level. Figure 4.1 shows two types of $P(S|p)/P(S)$ models: one corresponds to a monophone at the clearly uttered position $t_1$ and the other at the unclearly uttered position $t_2$. Next, we simplify the model by assuming that $A_{t_1} \cong A_{t_2}$ and $B_{t_1} \cong B_{t_2}$, namely all the $P(S|p)/P(S)$ can be approximated with a shift model in which only one parameter is $S_{\text{max}}$ that reflects the acoustic quality around the observing position. Figure 4.2 shows the approximated model. Thus, the likelihood normalization of acoustic quality is realized by the following procedure.

$$l_{\text{SM}}(j) = l_{\text{SM}}(j) - \max \{l_{\text{SM}}(j, r)\}$$

$$= (S(j) - S_{\text{max}}) - \max \{S(j, r) - S_{\text{max}}\}$$

$$= S(j) - \max \{S(j, r)\}$$

(5)

where, $S(j, r)$ is the similarity score of the $r$-th monophone at the $j$-th observing monophone position ($r = 1, 2, \ldots, 38$). We call the confidence score, which is calculated with equation (4) after substitution of equation (5) for $l_{\text{SM}}(j)$, the normalized CS (Figure 5: utterance is [jo:se:] (situation)).

The other simpler scoring procedures to normalize the variations of the acoustic quality is rank scoring [9]. In the procedure, such as for the 10-best case, the first ranked monophone is given the score of 10/55, the second 9/55,..., and the 10th 1/55.

3. EXPERIMENTS

3.1 Speech Database

The following two data sets were used:

**D1.** Acoustic model design set: A subset of “ASJ (Acoustic Society of Japan) Continuous Speech Database”, consisting of 4,503 sentences uttered by 30 male speakers (16 kHz, 16-bit).

**D2.** Test data set: A subset of “Tohoku University and Matsushita Spoken Word Database”, consisting of 100 words uttered by 10 unknown male speakers. The sampling rate was converted from 24 kHz to 16 kHz.

3.2 Experimental Setup

An input speech is sampled at 16 kHz and a 512-point FFT of the 25 ms Hamming-windowed speech segments is applied every 10 ms. The resultant FFT power spectrum is then integrated into 24-ch BPFs output with mel-scaled center frequencies. At the acoustic-feature extraction stage, two types of features are extracted. One is for the HMM-based classifier, and 24 outputs of a BPF bank are converted into
cepstrum (MFCC) by using DCT, then combined with \( \Delta \) parameters (12- \( \Delta \) t and 12- \( \Delta \) d), \( \Delta \) P, and \( \Delta \) \( \Delta \) P.

The other is for the SM-based verifier, and two types of local features (LFs) with the dimension of 24 each, extracted from BPF outputs by using LR, are converted into cepstrum with the dimension of 12 each, then combined with \( \Delta \) P. A compressed LFs set is also extracted for the SM-based verifier. The compressed feature set with the dimension of 12 is extracted from two LFs with the dimension of 24 each by using not only DCT but also DST [10].

The D1 data set was used to design 43 Japanese monophone-HMMs with five states and three loops. In the HMM, output probabilities are represented in the form of Gaussian mixtures, and covariance matrices are diagonalized (mixture = 1 to 8). The D1 data set was also used to design 38 eigen isolated-word sets (M=8) of SM. A speaker-independent isolated-word recognition test was then carried out with the D2 data set.

### 3.3 Experimental Results

Firstly the baseline performance with the HMM-based classifier is evaluated. Table 1 shows the word recognition rate within the best n, n=1,2,…10. This score gives the upper limit for the succeeding SM-based verifier.

#### [A] Comparison for scoring method

Table 2 shows the result of three evaluation tests. Ten hypotheses were tested in the verification process. Confidence scoring with “LHMM + rank score” and “LHMM + LSM”(n=10) improved the correct recognition rate by 1.5 % and 1.9 %, respectively. The normalized CS, “LHMM + normalized LSM”, achieved the highest gain of 2.7%.

#### [B] Comparison for acoustic feature extraction in an SM-based verifier

Three feature extractors for the SM-based verifier were evaluated. In the experiments, the normalized CS was used. The result in Table 3 shows that local features (LFs) with the same dimension as standard MFCC improve the correct recognition rate by 0.4%, and when LFs are not compressed, or are directly converted by DCT, LFs improve the correct recognition rate by 1.8%.

### 4. CONCLUSION

The normalized confidence scoring method based on SM was proposed and showed significant improvement in speaker-independent word recognition tasks. Application of the proposed scoring method to a key-words spotting task and an unknown-words detection task will be investigated in a future study.

---

### Table 1  Performance of the baseline system

<table>
<thead>
<tr>
<th>N (NUMBER OF HYPOTHESES)</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (LHMM)</td>
<td>95.3</td>
<td>96.6</td>
<td>98.2</td>
<td>98.7</td>
<td>98.8</td>
<td>98.8 [%]</td>
</tr>
</tbody>
</table>

### Table 2  Comparison for scoring method

<table>
<thead>
<tr>
<th>CONFIDENCE SCORING METHOD</th>
<th>CORRECT WORD RECOGNITION RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (LHMM)</td>
<td>95.3 [%]</td>
</tr>
<tr>
<td>LHMM + Rank scoring</td>
<td>96.8</td>
</tr>
<tr>
<td>LHMM + LSM</td>
<td>97.2</td>
</tr>
<tr>
<td>LHMM + normalized LSM</td>
<td>98.0</td>
</tr>
</tbody>
</table>

### Table 3  Comparison for acoustic feature extraction

<table>
<thead>
<tr>
<th>ACOUSTIC FEATURES</th>
<th>DIMENSION</th>
<th>CORRECT WORD RECOGNITION RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC + ( \Delta ) P</td>
<td>13×10</td>
<td>96.2 [%]</td>
</tr>
<tr>
<td>Compressed LFs + ( \Delta ) P</td>
<td>13×10</td>
<td>96.6</td>
</tr>
<tr>
<td>LFs + ( \Delta ) P</td>
<td>25×10</td>
<td>98.0</td>
</tr>
</tbody>
</table>

### REFERENCES


