Improving Performance of an HMM-Based ASR System by Using Monophone-Level Normalized Confidence Measure

Muhammad Ghulam, Takashi Fukuda, Takaharu Sato, and Tsuneo Nitta
Graduate School of Engineering, Toyohashi University of Technology
1-1 Hibari-gaoka, Tempaku, Toyohashi, Japan
ghulam@vox.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 (word candidates) and boundaries of all the monophones in the hypotheses. In the second pass, an SM (Sub-space Method)-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, how to combine an HMM-based likelihood of word level and an SM-based likelihood of monophone level, and also how to accept the correct words and reject OOV words. In the experiments performed on speaker-independent word recognition, the proposed confidence scoring method significantly reduced word error rate from 4.7% obtained by the standard HMM classifier to 2.0%, and it also reduced the equal error rate from 9.0% to 6.5% in an unknown word rejection task.

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 posteriori 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 detecting keywords and/or unknown words have been proposed. The proposed confidence measures include the likelihood ratio of $P(x|w)/P(x|p)$, where $P(x|p)$ is the accumulated likelihood of phonemes [1], sub-word [2] over a word x, and application of multiple features [3].

In this paper, we attempt to normalize the variations of likelihood affected by the acoustical difference 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 more 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, 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, and 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 [5]. 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 a normalized, accumulated likelihood of l_{SM}. If the score is above a threshold, then the word is accepted, otherwise the word is rejected. Feature extractions for the SM-based verifier are also discussed. In our previous work [6], the proposed method was evaluated in an isolated word recognition task and showed significant improvement comparing with a standard HMM classifier which implements the likelihood normalization in a whole word level by accumulating phoneme likelihoods. In this paper, we evaluate this phoneme likelihood normalization to reject the out-of-vocabulary (OOV) words.

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

2. SYSTEM OVERVIEW

Figure 1 shows a block diagram of the proposed two-pass spoken word recognition 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 that executes the first pass and outputs N-best hypotheses (word candidates) and all the intervals of monophones in the hypotheses; and the other is fed into the SM-based verifier, which performs the second pass and tests the hypotheses through the confidence score normalization. Then the score is compared with a threshold for the word to be accepted or rejected.

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 Gaussian mixtures with diagonal covariance matrices (mixture = 8).

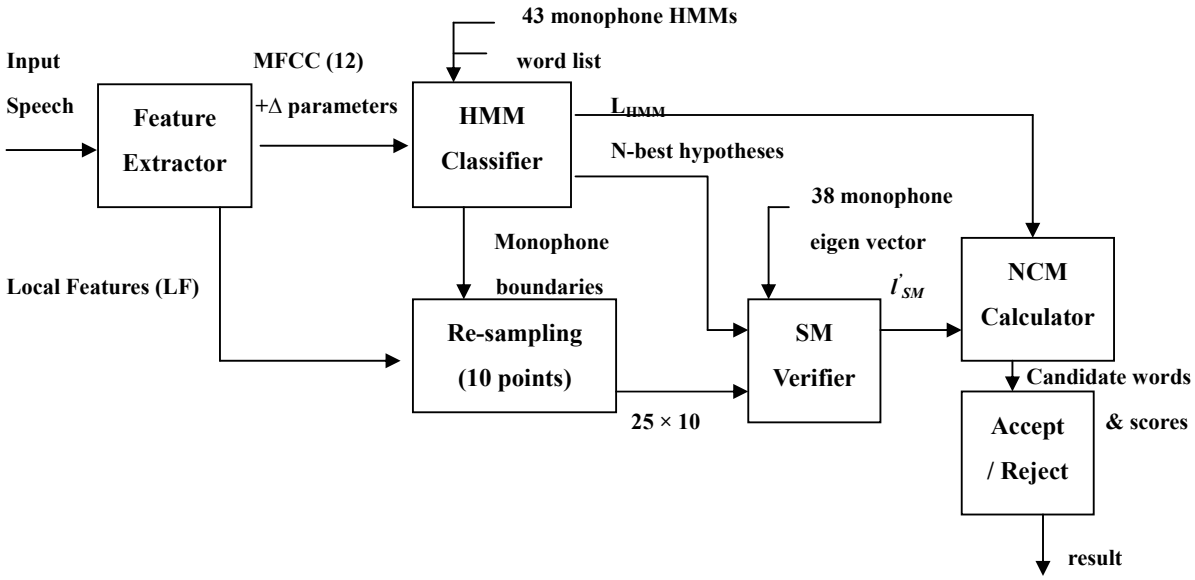

Figure 1 Block diagram of a spoken word recognition system

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 ϕ_{cm}, m=1,2,…,M, or an orthogonalized codebook, by using KLT from a learning database. The multiple similarity score Sc of class c between the codebook ϕ_{cm} and a normalized input pattern **x** is defined as follows:

$$S_c = \sum_{m=1}^{M} (\mathbf{x} \bullet \phi_{cm})^2 \qquad (1)$$

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

Here, we apply the local features (LFs) [7], extracted by using 3×3 derivative operators applying linear regression(LR) calculation along the time axis and frequency axis, to the SM-based verifier. After linearly re-sampling 10 frames of LFs between the monophone boundaries, the verifier calculates multiple similarity Sc, then converts the similarity into a posteriori probability described in the next section.

2.3 Conversion of Similarity into a posteriori Probability

Figure 2-A shows probability P(S|p), P(S), and P(S|p) / P(S) observed in real monophone speech data (section 3.1, data set D1). Here, S and p are the multiple similarity and monophone, respectively. Firstly, P(S|p) / P(S) is modeled by the following equation:

$$P(S \mid p) / P(S) = AB^{S} \qquad A > 0, B \geq 1 \qquad (2)$$

Next, the model is simplified as shown in **Figure 2-B**. By considering $\log_B B = 1$, likelihood score l_{SM} is given by the following maximum similarity normalization procedure [5]:

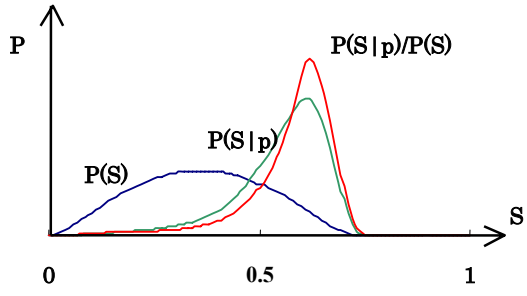

Figure 2-A Observed probability distributions

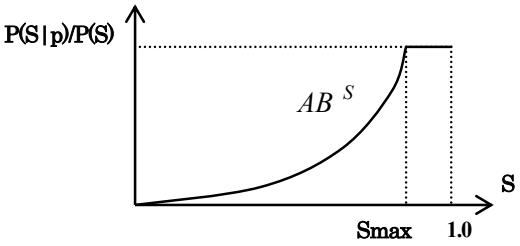

Figure 2-B A model of log[P(S|p) / P(S)]

$$\log[P(S \mid p_c) / P(S)] = \begin{cases} S - S_{\max} & , S \leq S_{\max} \\ 0 & , S > S_{\max} \end{cases} \qquad (3)$$

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

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

$$CM = \alpha L_{HMM} + (1-\alpha) L_{SM}$$

$$= \alpha L_{HMM} + (1-\alpha) \sum_{j=1}^{J} l_{SM}(j) \quad (4)$$

where, α and L_{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 and the number of monophones in a hypothesis, respectively.

Figure 3-A shows an example of confidence scoring in SM-based verifier. **Figure 3-B** compares two scores, L_{HMM} and CM = αL_{HMM} + (1- α)L_{SM}, for an input utterance [ro:do:] (labor), where the HMM classifier outputs [kodomo] (child) for the best hypothesis and [ro:do:] is the second best.

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. We can normalize l_{SM} by the following equation.

$$l'_{SM}(j) = l_{SM}(j) - avg\{l_{SM}(j)\}$$

$$= l_{SM}(j) - \log[\frac{1}{R} \sum_{r=1}^{R} \exp\{l'_{SM}(j,r)\}] \quad (5)$$

where, $l'_{SM}(j, r)$ is the log-likelihood of the r-th monophone at the j-th observing monophone position (R = 38). We can simplify the equation (5) as follows:

$$l'_{SM}(j) = l_{SM}(j) - \max_{r}\{l'_{SM}(j,r)\} \quad (6)$$

We call the confidence measure, which is calculated with equation (4) after substitution of equation (5) or (6) for $l_{SM}(j)$, the **normalized CM (NCM).** Here we apply the equation (6). In practice, the evaluation test on comparing (5) and (6) showed almost the same result. **Figure 4** illustrates an example of normalized L_{SM} for input utterance [awa].

2.5 Unknown Word Rejection

After calculating CM and NCM, the scores for each candidate word are compared with a threshold. If the score of an OOV word is greater than the threshold, then it is considered as falsely accepted (FA), while the score of a word within vocabulary is smaller than the threshold, false reject (FR) is occurred.

3. EXPERIMENTS

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 200 words uttered by 10 unknown male speakers. Of the 200 words, second 100 words are Out Of Vocabulary (OOV) words. The sampling rate was converted from 24 kHz to 16 kHz.

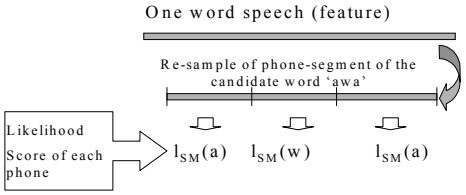

Figure 3-A Example of L_{SM} scoring in SM verifier

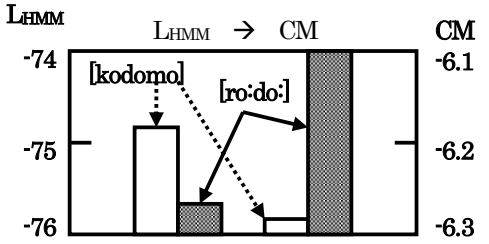

Figure 3-B L_{HMM} vs. CM (= αL_{HMM}+(1-α)L_{SM})

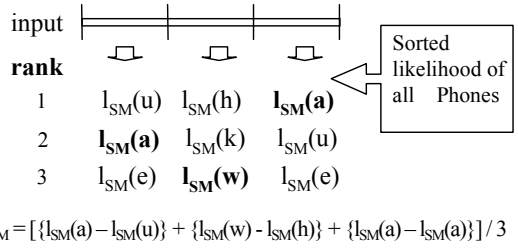

Figure 4 Example of Normalized L_{SM} scoring

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 Δ -parameters (12-Δt and 12-$\Delta\Delta$t), ΔP, and $\Delta\Delta$P.

The other is for the SM-based verifier, and two types of 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 ΔP. 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 .

The D1 data set was used to design 43 Japanese monophone-HMMs with five states and three loops. The D1 data set was also used to design 38 eigen vector 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 correct rate within the best N, N=1,2,…10. This score gives the upper limit for the succeeding SM-based verifier.

[A] Comparison of feature parameters and scoring methods

Figure 5 shows the result of two evaluation tests. Ten hypotheses were tested in the verification process. CM, based on SM, significantly improves WER. CM with LF (25) reduced WER by 2% comparing with baseline, while CM with MFCC (13) reduced it only 0.4%. The proposed NCM far more improves WER. NCM with LF (25) reduced WER by 2.7%, while NCM with MFCC (13) reduced it by 1.2%

[B] Rejection of unknown words

Three feature extractors for the SM-based verifier were evaluated. In the experiments, the NCM was applied to a D2 data set. **Figure 6, 7** show FRR for within vocabulary words vs. FAR for OOV words, and FRR vs. WER, respectively. From the figure, we can see that the proposed NCM with LF (25) reduced the equal error rate from 9.0% , obtained by the standard HMM classifier with a whole-word level normalization, to 6.5%.

Table 1 word correct rate of the baseline system [%]

N (NUMBER OF HYPOTHESES)					
1	2	4	6	8	10
95.3	96.6	98.2	98.7	98.8	98.8

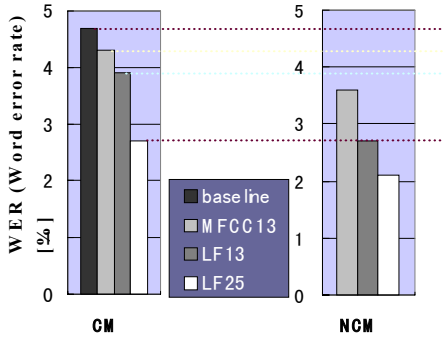

Figure 5 comparison of feature parameters and scoring methods

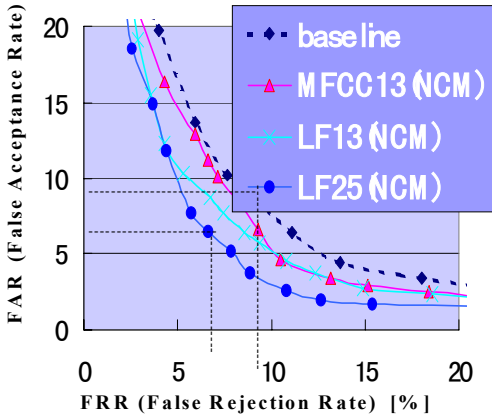

Figure 6 FRR vs. FAR

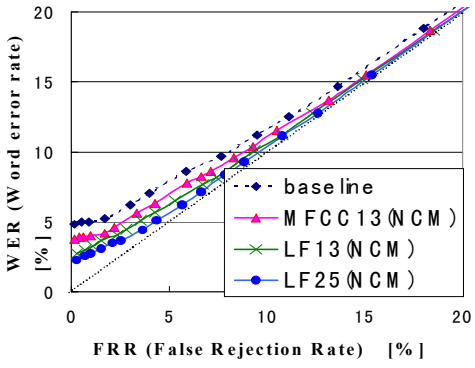

Figure 7 FRR vs. WER

4. CONCLUSION

The normalized confidence scoring method based on SM was proposed and showed significant improvement in speaker-independent word recognition tasks both for within vocabulary and out of vocabulary words. Application of the proposed scoring method to continuous speech recognition task will be investigated in a future study.

REFERENCES

[1] Asadi, R. Schwartz, and J. Makhoul, "Automatic detection of new words in a large vocabulary continuous speech recognition system", Proc. ICASSP'90, pp.125-128 (1990).

[2] R.A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition", IEEE Trans. Speech and Audio Process., Vol.4, No.6, pp.420-429 (1996).

[3] T. Schaaf and T. Kemp, "Confidence measures for spontaneous speech recognition", Proc. ICASSP, pp.875-878 (1997).

[4] E. Oja, "Subspace Method of Pattern Recognition", Research Studies Press (1983).

[5] T. Ukita, E. Saito, T. Nitta, and S. Watanabe, "A speaker-independent connected digit recognition system concatenating statistically discriminated words", IEEE Trans. Signal Process., Vol.40, No.10, pp.2414-2424 (1992).

[6] T. Sato, M. Ghulam, T. Fukuda, T.Nitta, "Confidence Scoring for Accurate HMM-based Word Recognition By Using SM-base Monophone Score Normalization ", Proc ICASSP'02 (2002)

[7] T. Nitta, "Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA," Proc. ICASSP'99, pp.421-424 (1999).