音声言語情報処理 44－2
(2002. 12. 19)

発話中における単音の音響的品質正規化の検討

Muhammad GHULAM† 福田 隆† 新田 恒雄‡

†豊橋技術科学大学 大学院工学研究科
〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

E-mail: †{ghulam,fukuda}@vox.tutkie.tut.ac.jp, ‡{nitta}@tutkie.tut.ac.jp

あらまし本報告では，これまでに提案したHMM-SM方式による音声認識を連続数字音声に適用すると共に，発話中における単音の音響的品質を正規化することの効果を，HMM に基づく様々な方式，すなわち発話全体の正規化，単語単位の正規化，単音単位の正規化，および正規化なしの場合と比較する。提案のHMM-SM方式では，最初にHMM分類器でN-best の単語候補と，候補単語内の全音素境界を求めた後，SM照合器で単音を構成する特徴ベクトル系列と，全単音の固有ベクトル群間で類似度を計算する。最後に，発話区間における単音の音響的品質の違いを反映させた類似度→尤度変換（正規化処理）を行うと共に，HMM 分類器による尤度を併用して発話内容を決定する。連続数字音声を対象に，標準的なＨＭＭ方式（発話全体で正規化）と比較した実験では，提案方式は単語正解率で 96.3%から 98.7%，また単語認識精度で 95.7%から 98.2% と大幅な向上を示した。また，他の様々な正規化改良を施したHMM方式と比較した際にも，提案方式はこれらの性能を大きく上回る結果を示した。

キーワード 音声認識，ＨＭＭ，音響品質の正規化，部分空間法

Normalizing the Acoustic Qualities of Monophones in an Utterance

Muhammad GHULAM†, Takashi FUKUDA† and Tsuneo NITTA‡

Graduate School of Engineering, Toyohashi University of Technology
1-1 Hibariga-oka, Tempaku, Toyohashi, 441-8580 Japan

E-mail: †{ghulam,fukuda}@vox.tutkie.tut.ac.jp, ‡{nitta}@tutkie.tut.ac.jp

Abstract In this paper, we expand our previously proposed HMM-SM-based speech recognition system [1,2, 3] to a connected digit recognition task by exploring the effect of normalizing the acoustic qualities of the monophones in an utterance and compare it with a number of HMM-based systems with utterance-level normalization, word-level normalization, monophone-level normalization and without normalization. In the proposed HMM-SM-based system, an HMM-based classifier classifies the N-best hypotheses (word candidates), and then an SM (Subspace Method)-based verifier tests the hypotheses after applying the monophone score normalization. Experimental results performed on a connected digit recognition task showed that the word correct rate and the word accuracy rate were significantly improved by the proposed method from 96.3% to 98.7% and from 95.7% to 98.2%, respectively, compared with the convenient HMM-based classifier with utterance-level normalization. The proposed method also showed high performance over the other HMM-based systems that we have compared.

Key words Speech Recognition, HMM, Normalization of Acoustic Quality, Subspace Method

1. INTRODUCTION

In standard 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. This means that these systems do not consider the difference of the acoustic quality throughout an utterance. But in practical cases, many factors affect the acoustic quality of the utterance. Hence various confidence scoring methods were proposed to verify an utterance for improving the word recognition accuracy. These confidence scoring methods include the likelihood ratio of $P(x|w)/P(x|p)$, where $P(x|p)$ is the accumulated likelihood of phonemes [4], sub-word [5] over a word x, etc.

In our previous works [1, 2, 3], the variations of likelihood affected by the acoustic quality in an utterance are normalized by applying the monophone-based Subspace Method (SM). In an HMM scheme, 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 contrary, SM can represent variations of fine structures in sub-words as a set of eigenvectors, and so has better performance at the phone level than HMM. In the proposed HMM-SM-based system [1, 2, 3], an HMM-based classifier not only classifies the N-best hypotheses but also determines the boundaries of all the monophones in each hypothesis. Then an SM-based verifier tests the hypotheses. In the verifier stage, we also normalize the acoustic quality of the monophones. Then this normalized score from SM-based verifier is combined with the word-level HMM-based score to give the competitive score of the target word. This HMM-SM-based system was successfully implemented on an isolated-word recognition task [1], and out-of-vocabulary word rejection task [2].

In this paper, we evaluate the connected digit recognition using HMM-based systems with four different approaches: utterance-level normalization, word-level normalization, monophone-level normalization, and without normalization. Then we compare these performances with that of our proposed HMM-SM-based system with monophone-level normalization.

This paper is organized as follows. Section 2 outlines the system configuration and discusses the proposed normalization of acoustic quality of the monophones, section 3 describes the experimental setup and the results, and section 4 draws some conclusion.

2. OVERVIEW OF THE RECOGNITION SYSTEM

Figure 1, 2 and 3 show the block diagrams of the HMM-based systems with utterance-level normalization, word-level normalization, and monophone-level normalization, respectively. The HMM-based classifier in this paper adopts a standard monophone-based HMM with 5-state 3-loop left-to-right models, 38 standard MFCC parameters, and Gaussian mixtures with diagonal covariance matrices (mixture = 8).

2.1 HMM-based system with utterance-level normalization

In this system, we normalize the difference of the acoustic quality in the utterance level. The HMM-based system with utterance-level normalization ranks the best candidates by using not only the word models but also using filler models throughout an utterance. In this case, the log likelihood using HMM is as follows:

$$L^{utterance-level}{}_{HMM} = (L^{u}{}_{HMM} - L^{f}{}_{HMM}) \qquad (1)$$

where, $L^{u}{}_{HMM}$ and $L^{f}{}_{HMM}$ are the log likelihood of HMM using word models and filler models, respectively. This HMM-based system with utterance-level normalization is shown in Figure 1, where T is the total number of frames in an utterance.

2.2 HMM-based system with word-level normalization

A block diagram of the HMM-based system with word-level normalization that normalizes the difference of the acoustic quality between the words throughout an utterance is shown is Figure 2. The HMM-based classifier outputs the N-best hypotheses with their log likelihood $L^{u}{}_{HMM}$, and also determines the word boundaries by a backtracking procedure. Then the HMM-based verifier normalizes the word scoring with the following equation:

$$l_{HMM}(w) \leftarrow l_{HMM}(w) - \max_{r}(l'_{HMM}(w,r)) \qquad (2)$$

where, $l'_{HMM}(w, r)$ is the log likelihood of the r-th word at the interval at which the w-th target word was observed and $l_{HMM}(w)$ is the HMM-based log likelihood of the w-th target word in the utterance. The HMM-based word-level normalized score can be found by the following equation:

$$L^{word-level}{}_{HMM} = \frac{1}{W}\sum_{w=1}^{W} l_{HMM}(w) \qquad (3)$$

where, W is the number of words in a word string.

2.3 Normalization of acoustic quality at monophone-level

We investigate the effect of normalization even at smaller unit like monophones, because many factors affect the acoustic quality

of the monophones in an utterance. These factors may include breathing, accentuation, speaking rate, etc. The difference of acoustic quality at each monophone interval should be reflected on phone scoring.

To reduce the degree of bias of scoring caused by the difference of acoustic quality, we perform normalization of acoustic quality at each monophone interval.

(1) HMM-based system with monophone-level normalization

At first, we try to normalize the monophone scoring using HMM-based system only. The block diagram of this system is shown in Figure 3. An HMM-based classifier estimates not only the N-best hypotheses but also the boundaries of all the monophones in each hypothesis. We normalize the acoustic quality of the monophones in the HMM-based verifier stage by the following equation:

$$l_{HMM}(j) \leftarrow l_{HMM}(j) - \max_{r}(l'_{HMM}(j,r)) \qquad (4)$$

where $l'_{HMM}(j, r)$ is the log likelihood of r-th monophone at the interval at which the j-th target monophone was observed and $l_{HMM}(j)$ is the HMM-based log likelihood of the j-th target monophone in the utterance. We get the score of HMM-based monophone-level normalization with the following equation:

$$L^{monophone-level}{}_{HMM} = \frac{1}{J}\sum_{j=1}^{J} l_{HMM}(j) \qquad (5)$$

where, J is the total number of monophones in the utterance.

(2) The proposed HMM-SM-based system with monophone-level normalization

Figure 4 shows the block diagram of the proposed HMM-SM-based system with monophone-level normalization. The details are described in our previous works [1, 2, 3]. In this method, an HMM-based classifier estimates the N-best hypotheses and the boundaries of all the monophones in each hypothesis. Then, in an SM-based verifier stage, the similarity score for each monophone is computed and later this score is converted into log likelihood. The SM-based verifier also normalizes these monophone log likelihood by the following equation:

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

where, l_{SM} is the SM-based log likelihood. This normalization process is illustrated in Figure 5.

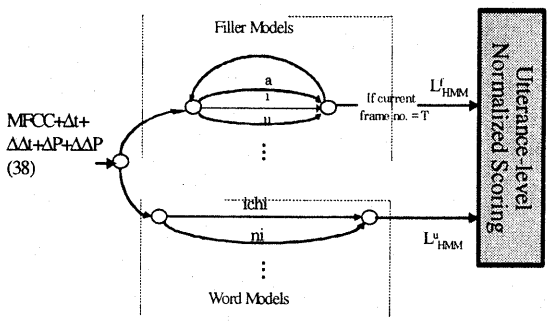

Figure 1 An HMM-based system with utterance-level normalization

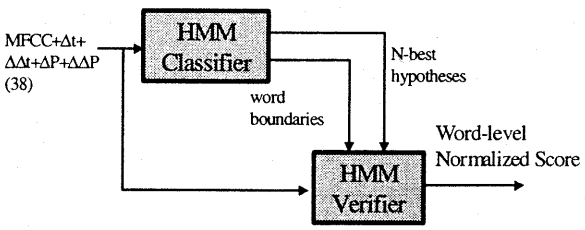

Figure 2 An HMM-based system with word-level normalization

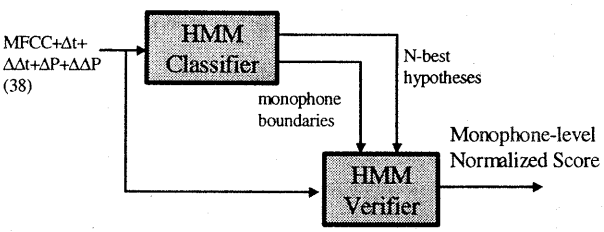

Figure 3 An HMM-based system with monophone-level normalization

We derive the proposed monophone-level normalization by a linear combination of the HMM-based score with the normalized monophone-level SM-based score using the following equation:

$$L^{monophone-level}{}_{HMM-SM} = \alpha\frac{L^u{}_{HMM}}{T} + \frac{(1-\alpha)}{J}\sum_{j=1}^{J} l_{SM}(j) \qquad (7)$$

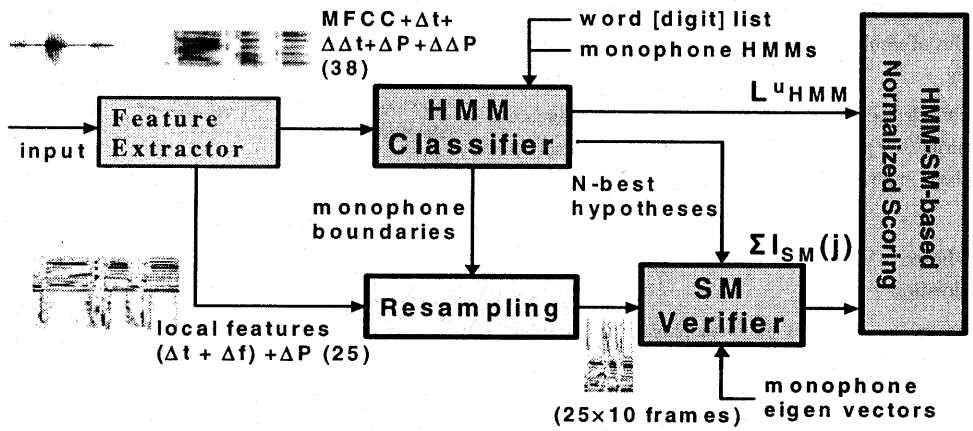

Figure 4 An HMM-SM-based system with monophone-level normalization

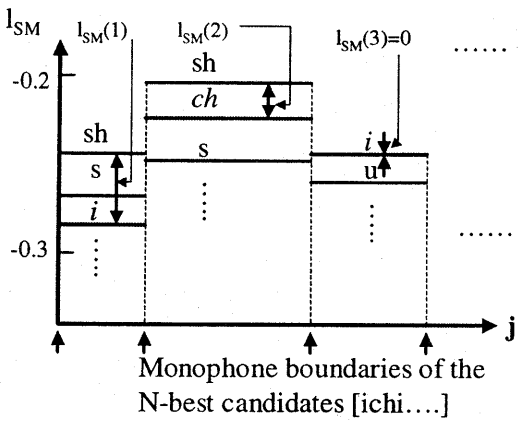

Figure 5 Example of normalizing the monophone scores

where, α, T, J are the weighting coefficient, number of frames in the utterance and number of monophones in the utterance, respectively, whereas L^u_{HMM} and $l_{SM}(j)$ are the log likelihood of HMM using word models and the j-th monophone log likelihood of SM, respectively.

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 set of 35 connected 4-digit strings as shown in Table 1 [6] uttered by 12 male speakers, once by each speaker (16 kHz).

Table 1 List of 35 connected 4-digit strings used as a test data set

0287	5732	9601	4156	1199
1398	6843	0712	5267	6633
2409	7954	1823	6378	8877
3510	8065	2934	7489	2244
4621	9176	3045	8590	5500
6972	5861	3649	0316	7083
8194	9205	1427	2538	4750

3.2 Experimental setup

An input speech is sampled at 16kHz and a 512-point FFT of the 25ms Hamming-windowed speech segments is applied every 10 ms. The resultant FFT power spectrum is then integrated into a 24-ch BPF output with mel-scaled center frequencies.

Two types of features are extracted, one for the HMM-based classifier and the other for the SM-based verifier. For the HMM-based classifier, 24 outputs of a BPF bank are converted into cepstrum (MFCC) by using DCT, then combined with Δ-parameters (12-Δt and 12-ΔΔt), Δp, and ΔΔp, where p stands for

log power.

For the SM-based verifier, two types of LFs (Local Features) [7] with dimensions of 24 each, extracted from BPF outputs by using LR, are converted into cepstrum with the dimension of 12 each, then combined with Δp. It was verified before that LF 25 has better performance than MFCC parameters in SM-based verifier stage [1, 2, 3].

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 eigenvector sets of SM [1, 2, 3]. A speaker independent connected digit recognition test was then carried out with the D2 data set.

3.3 Experimental results

We conducted the experiments on connected digit recognition task using five different methods: HMM-based system without normalization, HMM-based systems with utterance-level normalization, word-level normalization and monophone-level normalization, and the proposed HMM-SM-based system with monophone-level normalization. We also varied the value of insertion penalty (IP) for each of the methods.

Figures 6 and 7 compare the word correct rate (WCR) and the word accuracy rate (WAR) between the five methods for different values of IP. Here, WCR is calculated as $[(N - S - D) / N] \times 100\%$ and WAR is calculated as $[(N - S - D - I) / N] \times 100\%$, where N is the total number of digits, and S, D, I are the number of substituted, deleted and inserted digits, respectively. From these figures, it can be seen that, as IP increases, WCR is decreased whereas WAR is increased. The proposed HMM-SM-based system with monophone-level normalization ($L^{\text{monophone-level}}_{\text{HMM-SM}}$) has significantly high WCR and WAR compared with other HMM-based systems.

From Figure 7, we can see that, there is no improvement of WAR beyond IP = 3. Considering this fact with the steep degradation of WCR in Figure 6, we can say that the systems give the best performance with IP = 2. The results of comparing WCR and WAR for the five different methods for IP = 2 are shown in Figure 8.

The Figure 8 illustrates the superiority of the proposed HMM-SM-based system with monophone-level normalization in both WCR and WAR. The Figure shows that the HMM-based system without normalization has the least WCR and WAR. Applying normalization in utterance-level and in word-level in the HMM-based systems gradually increase WCR and WAR, where word-level normalization has better performance. For example, HMM-based system without normalization has 95.8% WCR and 94.2% WAR, whereas HMM-based system with word-level normalization has 96.6% WCR and 95.9% WAR compared with

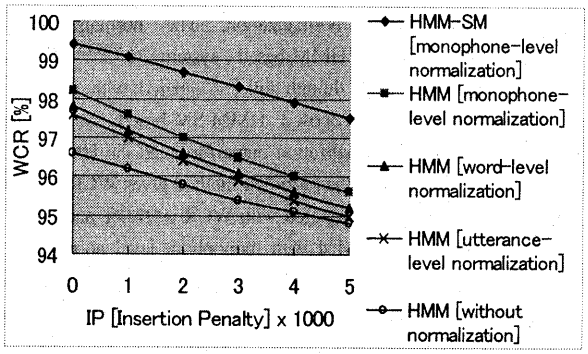

Figure 6 Comparison of the normalization method: WCR

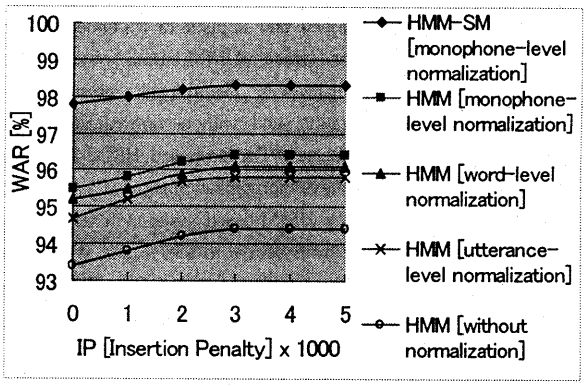

Figure 7 Comparison of the normalization method: WAR

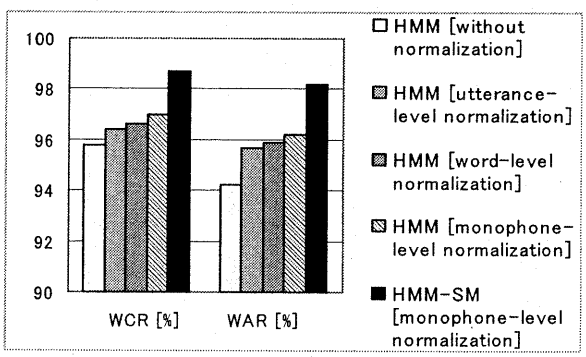

Figure 8 Comparison between normalization methods (IP=2)

96.3% WCR and 95.7% WAR obtained by HMM-based system with utterance-level normalization. The normalization of monophones in an HMM-based system ($L^{\text{monophone-level}}_{\text{HMM}}$) achieves further gain, though little, compared with word-level normalization. The proposed HMM-SM-based system with monophone-level normalization outperforms all the HMM-based systems. For example, the proposed method gives WCR = 98.7% and WAR = 98.2%, compared with WCR = 97.0% and WAR = 96.2% obtained by HMM with monophone-level normalization that has comparatively better performance over other HMM-based systems. It signifies that the monophone normalization in SM-verifier stage gives better performance than that in HMM-verifier stage.

4. CONCLUSION

A method for normalizing the acoustic quality of monophones in an utterance was developed and applied to a connected digit recognition task. Some typical HMM-based systems with monophone-level normalization, word-level normalization, utterance-level normalization, and without normalization were also investigated for this task. The proposed HMM-SM-based system with monophone-level normalization showed a significant improvement over all the HMM-based systems. The proposed system improved WCR from 96.3% to 98.7% and WAR from 95.7% to 98.2% compared with the convenient HMM-based system with utterance-level normalization. It also achieved 1.7% gain in WCR and 2% gain in WAR compared with HMM-based system with monophone-level normalization, justifying the use of normalization of monophone score in an SM-based verifier rather than in an HMM-based verifier.

The effects of the proposed HMM-SM-based system on a continuous word recognition task will be investigated as a future study.

5. REFERENCES

[1] M. Ghulam, T. Fukuda, T. Sato, and T. Nitta, "Improving performance of an HMM-based ASR system by using monophone-level normalized confidence measure," Proc. ICSLP'02, vol.4, pp.2453-2456 (2002).

[2] T. Sato, M. Ghulam, T. Fukuda, and T. Nitta, "Confidence scoring for accurate HMM-based word recognition by using SM-based monophone score normalization," Proc. ICASSP'02, pp.I-217-I-220 (2002).

[3] M. Ghulam, T. Sato, T. Fukuda and T. Nitta, "Confidence scoring for accurate HMM-based speech recognition by using monophone-level normalization based on subspace method," Technical report of IEICE, SP2002-41, pp.31-36, (2002-06).

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

[5] 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.

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

[7] T. Nitta, "Feature extraction for speech recognition based on orthogonal acoustic feature planes and LDA," Proc. ICASSP, pp.421-424, 1999.