Voice Quality Normalization in an Utterance for Robust ASR

Muhammad Ghulam, Takashi Fukuda 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 method of normalizing the voice quality in an utterance for both clean speech and speech contaminated by noise. The normalization method is applied to the N-best hypotheses from an HMM-based classifier, then an SM (Sub-space Method)-based verifier tests the hypotheses after normalizing the monophone scores together with the HMM-based likelihood score. The HMM-SM-based speech recognition system was proposed previously [1, 2] and successfully implemented on a speaker-independent word recognition task and an OOV word rejection task. We extend the proposed system to a connected digit string recognition task by exploring the effect of the voice quality normalization in an utterance for robust ASR and compare it with the HMM-based recognition systems with utterance-level normalization, word-level normalization, monophone-level normalization, and state-level normalization. Experimental results performed on connected 4-digit strings showed that the word accuracy was significantly improved from 95.7% obtained by the typical HMM-based system with utterance-level normalization to 98.2% obtained by the HMM-SM-based system for clean speech, from 88.1% to 91.5% for noise-added speech with SNR=10dB, and from 72.4% to 76.4% for noise-added speech with SNR=5dB, while the other HMM-based systems also showed lower performances.

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. It means that those typical systems do not consider the difference of voice quality throughout an utterance, but in practical cases many factors affect the voice quality in an utterance. Hence various acoustic confidence scoring methods to verify an utterance for improving word recognition accuracy have been proposed. The proposed scoring methods include the likelihood ratio of $P(x|w)/P(x|p)$, where $P(x|p)$ is the accumulated likelihood of phonemes [3], sub-word [4] over a word x, etc.

In our previous works [1, 2], the variations of likelihood affected by the voice difference in an utterance are normalized by applying the monophone-based Sub-space Method (SM). First, an HMM-based classifier calculates both the N-best hypotheses and monophone boundaries of all the monophones, then an SM-based verifier tests those hypotheses. In the verifier, after normalizing the voice quality of the monophones, the normalized score from SM is combined with the word-level HMM score to give the competitive score of the targeted word. This method was successfully implemented on an isolated-word recognition task [1] and an out-of-vocabulary word rejection task [2].

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 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 state-level normalization. Then we compare these performances with that of our proposed normalized monophone scoring method using an HMM-SM system. The systems are evaluated both with clean speech and with speech contaminated by additive white noise.

The paper is organized as follows. Section 2 outlines the system configuration and discusses the proposed normalization of voice quality in an utterance, section 3 describes the experimental setup and results of a connected digit recognition task, and section 4 draws some conclusions.

2. Overview of the recognition system

Fig. 1, 2, 3 and 4 show the block diagrams of the HMM-based systems with utterance-level, word-level, monophone-level, and state-level normalization, respectively. The HMM-based classifier in this paper adopts a standard monophone-based HMM with 5-states 3-loops left-to-right models.

2.1 HMM-based system with utterance-level normalization

The HMM-based system with utterance-level normalization, shown in Fig. 1, ranks the best candidates by using not only word models but also using filler models throughout an utterance (T frames). In this case the log likelihood score 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.

2.2 HMM-based system with word-level normalization

In this system, shown in Fig. 2, the difference of voice quality of each word in an utterance is normalized. 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 targeted word was observed and $l_{HMM}(w)$ is the HMM-based log likelihood of the w-th targeted 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 HMM-based system with monophone-level normalization

We investigate the effect of normalization at smaller unit like monophones, as many factors affect the voice quality of the monophones in an utterance. These factors may include breathing, accentuation, speaking rate, etc. To reduce the bias in scoring caused by the difference of voice quality, we perform normalization of voice quality at each monophone interval.

At first, we try to normalize the monophone scoring using HMM-based system only. In this system (Fig. 3), the HMM-based classifier estimates not only the N-best hypotheses, but also the boundaries of each monophone in the hypotheses. We normalize the voice quality of the monophones by using the following equation in the HMM-based verifier:

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

where, $l'_{HMM}(j, r)$ is the log likelihood of the r-th monophone at the interval where the j-th target monophone is observed and $l_{HMM}(j)$ is the HMM-based log likelihood of the j-th target monophone in the utterance. Finally the HMM-based monophone-level normalized score is given by 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.

2.4 HMM-based system with state-level normalization

We further investigate the effect of normalization even at smaller unit like states. Fig. 4 shows a block diagram of this system, where the HMM-based classifier outputs the N-best hypotheses with their state boundaries, and the HMM-based verifier normalizes the state scores as with equation (4), where the term monophone is replaced by state. Then the HMM-based state-level normalized score is found by summing all the normalized state scores in an utterance.
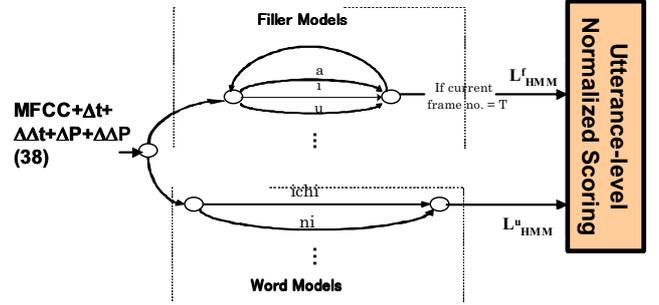

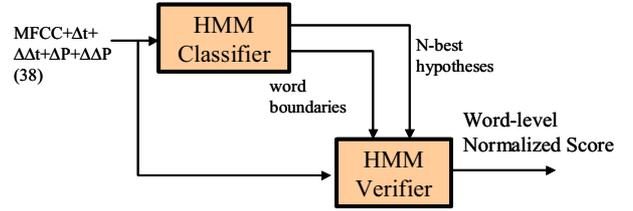*Figure 1*: An HMM-based system with utterance-level normalization

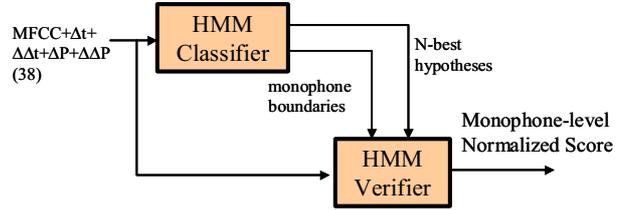*Figure 2*: An HMM-based system with word-level normalization

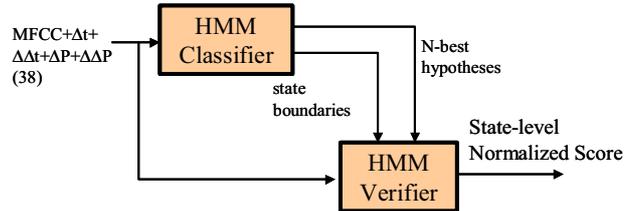*Figure 3*: An HMM-based system with monophone-level normalization

Figure 4: An HMM-based system with state-level normalization

2.5 The proposed HMM-SM-based system with monophone-level normalization

Fig. 5 shows a block diagram of the proposed HMM-SM-based system with monophone-level normalization. The details   are
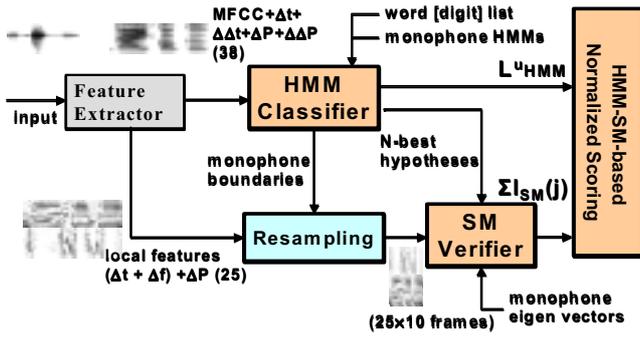
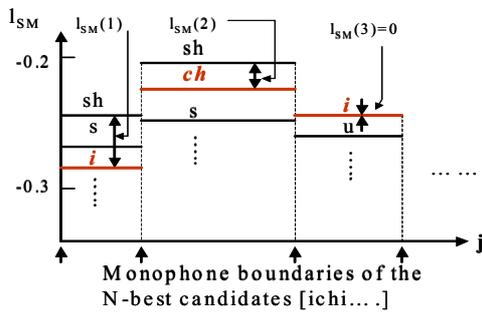*Figure 5*: An HMM-SM-based system with monophone-level normalization

Figure 6: Example of monophone score normalization

Figure 7: Comparison of normalization methods: WCR
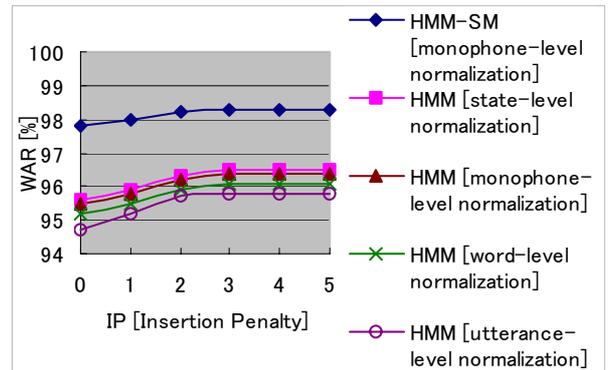
Figure 8: Comparison of normalization methods: WAR

described in our previous works [1,2].

In the proposed system, we perform the normalization of monophone scores in an SM-based verifier with the following equation:

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

where, l_{SM} is SM-based log likelihood. This normalization process is illustrated in Fig. 6.

We derive the score of the proposed HMM-SM-based system by a linear combination of the HMM-based utterance score, L^u_{HMM}, and the normalized monophone-level SM-based score, $l_{SM}(j)$, using the following equation:

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

where, α, T, J, and $l_{SM}(j)$ are a weighting coefficient, the total number of frames in the utterance, the total number of monophones in the utterance, and the j-th monophone 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 [5] uttered by 12 male speakers, once by each speaker (16 kHz). For noisy data, white noise was artificially added to the speech with two types of SNRs (10dB and 5dB).
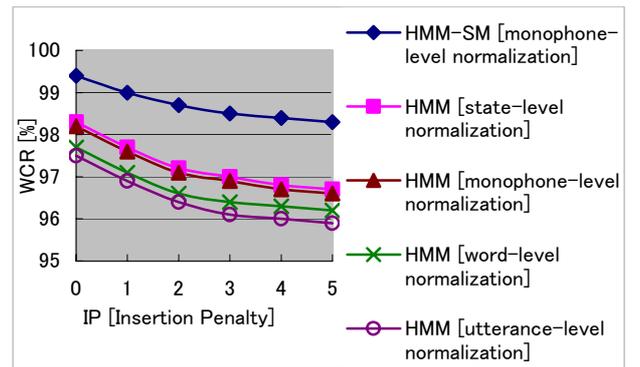
3.2 Experimental setup

The 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.

Two types of features were extracted, one for the HMM-based classifier and the other for the SM-based verifier. For the HMM-based classifier. MFCC with 38 dimensions including 12-Δ_t, 12-$\Delta\Delta_t$, Δ_P, and $\Delta\Delta_P$, where P stands for log power, was used. For the SM-based verifier, LFs (Local Features) with 25 dimensions (combined with Δ_P) are used [1, 2]. The HMM-based classifier uses Gaussian mixtures with diagonal covariance matrices (mixture=8).

The D1 data set was used to design 43 Japanese monophone HMMs with five states and three loops. This data set was also used to design 38 eigenvector sets [1, 2]. A speaker-independent connected digit recognition test was then carried out with the D2 data set.

3.3 Experimental results and discussion

We conducted experiments on a connected digit recognition task using the five different approaches, varying insertion penalty (IP).

Fig. 7 and Fig. 8 compare the word correct rate (WCR) and word accurate rate (WAR) between the five approaches with clean speech. Here, WCR is calculated as [(N-D-S)/N] × 100%, and WAR is calculated as [(N-D-S-I)/N] × 100%, where N, D, S, I are the total number of digits, and the number of deleted, substituted 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 has significantly high WCR and WAR compared with other HMM-based systems. Comparing Fig. 7 and Fig. 8, we can say that the systems give the best performance with IP = 2.

The experimental results performed on the same task with noise-added speech (SNR=10dB, 5dB) are shown if Fig. 9 for WCR, and in Fig. 10 for WAR. The insertion penalty in the experiments is set to 2. The Fig. 9 and Fig.10 illustrate the superiority of the proposed HMM-SM-based system with monophone-level normalization both in clean and in noise-added speech. For example, the proposed HMM-SM-based system has WCR=98.7%, 92.5%, and 77.5% compared to 96.4%, 90.2%, and 73.3% obtained by the typical HMM-based system with utterance-level normalization for clean and noise-added speech (SNR=10dB, 5dB), respectively. Again, the proposed system gives WAR=98.7% for clean speech, 91.5% for noise-added speech with SNR=10dB, and 76.4% for noise-added speech with SNR=5dB, whereas the HMM-based system with utterance-level normalization gives WAR=95.7%, 88.1%, and 72.4% for clean speech, and noise-added speech with SNR=10dB and 5dB, respectively. The proposed system also outperforms HMM-based system with monophone-level normalization with 2% absolute gain in WAR for clean speech, and 2.4% and 1.1% absolute gains in WAR for noise-added speech with SNR = 10dB and 5dB, respectively, justifying the use of monophone normalization in the SM-based verifier, rather than in the HMM-based verifier for robust ASR.

For the HMM-based systems, normalization of voice quality in smaller units gradually increases WCR and WAR. For example, the HMM-based systems with word-level, monophone-level and state-level normalization had absolute WAR gains of 0.2%, 0.5%, and 0.6% for clean speech, 0.3%, 1.0%, and 1.1% for noise-added speech (SNR=10dB), and 1.4%, 1.9%, and 2.0% for noise-added speech (SNR=5dB), respectively, over the HMM-based system with utterance-level normalization.

4. Conclusion

A method for normalizing the voice quality in an utterance was developed and applied to a speaker-independent connected digit recognition task both with clean speech and with speech contaminated by noise. The HMM-based systems with utterance-level normalization, word-level normalization, monophone-level normalization, and state-level 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 had 2.5%, 2.5%, and 3.4% absolute gains in WAR for clean speech and speech contaminated by noise (SNR=10dB, 5dB), respectively, over the HMM-based system with utterance-level normalization. The monophone normalization in an SM-based verifier was found to have far better performance than that in an HMM-based verifier.

The effects of our proposed system on an LVCSR task will be investigated in a future study.
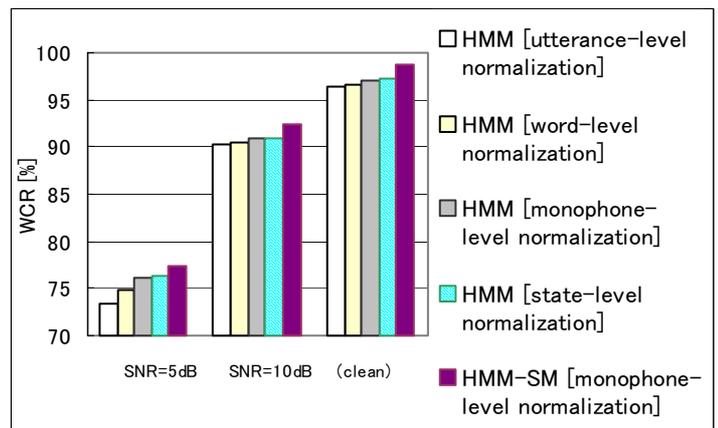
Acknowledgement

Figure 9: Comparison between normalization methods: WCR [%], for noise-added speech (SNR=10dB, 5dB)
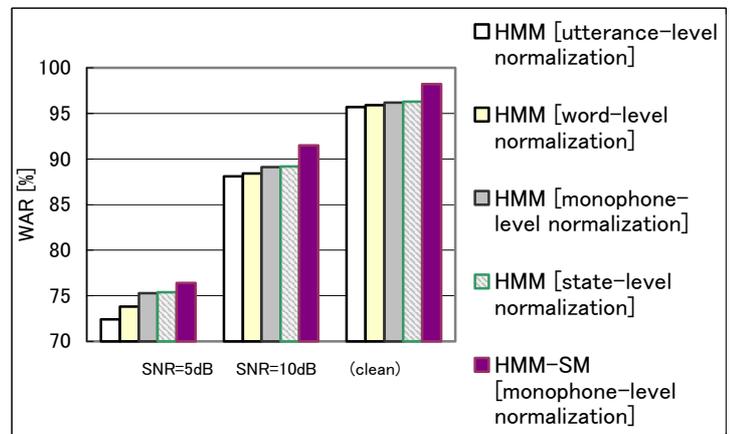

Figure 10: Comparison between normalization methods:WAR[%], for noise-added speech (SNR=10dB, 5dB)

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

[4] 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.

[5] 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).