SEGMENTAL CONFIDENCE WEIGHTING FOR RECOGNITION OF PARTLY CONTAMINATED SPEECH

Hoon-Young Cho, Lag-Young Kim and Yung-Hwan Oh

Department of Electrical Engineering and Computer Science, KAIST, Korea
{hycho, kly, yhoh}@bulsai.kaist.ac.kr

ABSTRACT

This paper proposed a segmental confidence weighting (SCW) method that makes the segments with higher confidence score contribute more to the recognition score at a pattern comparison step. As a two-stage approach, the proposed method obtains N-best candidates and calculates HMM (hidden Markov model) state-level segmental confidence scores using the most likely candidate. After that it normalizes segmental confidence scores and uses them as weighting factors for a modified Viterbi algorithm to select the best candidate among the N-best candidates. We added burst noises of various SINR (signal-to-impulsive noise ratio) to 3000 word utterances for the verification of the proposed method. The experimental results showed that the proposed method could reduce the word error rate by about 12.5% at SINR -5 dB.

1. INTRODUCTION

Speech recognition system operating in a real situation may have to deal with a variety of noise signals to prevent serious degradation of the performance. Great advances have been achieved in the area of robust speech recognition over the past two decades. The three major approaches are robust feature extraction, speech enhancement and model-based compensation. The first approach tries to extract speech features that are inherently immune to noise. The second one tries to restore clean speech by removing noise from noisy speech signal. Finally, the model-based techniques adapt or compensate model parameters according to a target noise environment [1].

Most of the previous work is concentrated on the stationary noise process like a white Gaussian noise or a slowly varying noise such as a car noise. However, a large portion of real world noise is highly non-stationary or impulsive, which cannot be dealt effectively with the previous approach. Therefore, new compensation techniques are required for the ASR systems running in these kinds of noise environments.

An impulsive or highly non-stationary noise contaminates certain segments of speech signal and leaves a large fraction of speech samples nearly unaffected while a slowly varying non-stationary noise degrades each segment on different levels. For this reason segments of a noisy speech have various degrees of confidence. In most of current ASR systems every segments contribute to a recognition score equally, but it would be better if a segment with higher confidence is taken more into account at the pattern comparison step.

As a related approach, the segmental signal to noise ratio (SNR) of a frame was used as a reliability measure and was applied as a weighting factor for a distance between two vectors in DTW (dynamic time warping)-based pattern matching procedure [3]. In another study, the segmental SNR and a noise suppression technique are combined to give a weighting factor [4]. A weighted Viterbi algorithm was proposed to apply weighting factors to HMM (hidden Markov model) based recognition system [5]. These approaches are effective for slowly varying noise environments because it is feasible to estimate the segmental SNR. However, they cannot be used in highly non-stationary or impulsive noise situation because of the difficulties in obtaining local SNR values.

This paper is based on the confidence measures that have been extensively studied for the non-keyword rejection or the utterance verification [7]. The previous confidence measure was mainly used to reject a whole word-level or a whole utterance-level speech, but in this work, we extended it to a subword segment level. The proposed segmental confidence weighting (SCW) method calculates confidence scores for HMM state level segments and normalize them to the values between 0 and 1. A modified Viterbi algorithm uses the normalized confidence scores as weighting factors to give a likelihood score. As the proposed method does not require an estimation of noise spectrum or a calculation of segmental SNR, it fits to the impulsive or the highly non-stationary noise environments.

This paper is organized as follows. Section 2 describes briefly on the segmental SNR weighting methods. Section 3 explains several different segmental confidence measures. In section 4, we propose two confidence normalization functions. Section 5 details the modified Viterbi algorithm and overall 2-stage procedure based on the proposed segmental confidence weighting (SCW) method. Section 6 gives experimental results followed by conclusions in section 7.

2. SEGMENTAL SNR WEIGHTING

The feature vectors extracted from a noisy speech are contaminated in different ways along the characteristics of the noise properties. As shown in the figure 1, the stationary noise contaminates the entire feature vector sequence while the impulsive noise contaminates only several portions of a speech without affecting the other portions of the speech. A non-stationary noise is an intermediate type of these two and it contaminates each segment to a different degree. Even if the noise is stationary, as the power of a speech signal changes fast, the segmental SNR of a noisy speech is time-varying [3].

The segments with low SNR lose more linguistic information than the segments with high SNR do. Therefore, the confidence of segment is proportional to its SNR value. Because a noisy speech is composed of segments with various confidence values, the pattern comparison algorithm should deal each segment with different degrees of importance. However, most of current systems take the segments equally.

Several approaches that reflect local SNR at the pattern comparison step were proposed [3][4][5]. They first estimate segmental SNR and normalize it to the values between 0 and 1.

Then use the values as weighting factors at the pattern comparison step of DTW or HMM based recognition systems.

These methods showed performance improvements with the reliable estimation of segmental SNR. However, when the noise is fast time-varying or impulsive, an exact estimation of the segmental SNR becomes difficult. Therefore, in this work we take another approach that does not need an estimation of local SNR. The proposed method is based on the confidence measure, which is described in the following sections.

3. SEGMENTAL CONFIDENCE MEASURE

For the reliability test of a recognition result, the confidence measure has been extensively studied in the area of utterance verification to reject incorrect utterances such as out-of-vocabulary words, speaker's hesitations or noise tokens. This approach has rejected or accepted the whole utterance at a sentence or a word level only.

In this study, we extend the previous confidence measure to a subword level, for example, a phoneme or a HMM state level, to reduce the contribution of severely contaminated segments to a likelihood score. Section 3.1 summarizes the framework of the utterance level confidence measure [6] followed by the proposed subword level segmental confidence measure.

3.1 Word level confidence measure

Given a feature vector sequence $X = \{x_1, x_2, \cdots, x_T\}$, the Viterbi decoding is employed in the recognition process to determine the most likely word W, where

$$W = \arg\max_j \; L(X \mid W_j) \tag{1}$$

In the context of subword recognition, W is a concatenation of subword units that can be written as

$$W = p_1 p_2 \cdots p_m \cdots p_M \tag{2}$$

where M is the number of subword units comprising W. Assuming independence among subword units, the Viterbi decoding implies that we can write the likelihood in (1) as

$$L(X \mid W) = \max_{t_1, t_2, \cdots, t_{M-1}} L(X_{t_0}^{t_1} \mid p_1) L(X_{t_1}^{t_2} \mid p_2) \cdots L(X_{t_{M-1}}^{t_M} \mid p_M) \tag{3}$$

where $X_{t_{i-1}}^{t_i}$ is the feature vectors between t_{i-1} and t_i corresponding to the speech segment for subword unit p_i. Given the subword model $p_1 p_2 \cdots p_M$, subword level verification is performed independently on each subword in the string, implying M independent likelihood ratio tests. For a given subword p_m in W, the likelihood ratio can be written as

$$T(X_{t_{m-1}}^{t_m}; p_m) = \frac{L(X_{t_{m-1}}^{t_m} \mid H_0)}{L(X_{t_{m-1}}^{t_m} \mid H_1)}, \quad 1 \le m \le M \tag{4}$$

where H_0 is the hypothesis that the segment $X_{t_{m-1}}^{t_m}$ correspond to the true instance for subword p_m, and H_1 is the hypothesis that the segment does not correspond to the true instance for

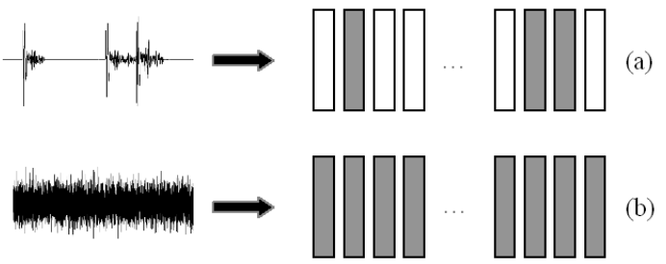

Figure 1. Contamination of a feature vector sequence by (a) an impulsive noise (b) a stationary noise.

subword p_m. Taking the log of (4) with simplifying the notation $X_{t_{m-1}}^{t_m}$ as X_m, results in a log likelihood difference, which is

$$G(X_m; p_m) = \log[\, L(X_m \mid H_0)] - \log[\, L(X_m \mid H_1)], \quad 1 \le m \le M \tag{5}$$

Since the density functions of $L(X_m \mid H_0)$ and $L(X_m \mid H_1)$ are not known exactly, (5) can be approximated as a measure of the classification of two classes, which is

$$V_m(X_m; p_m) = \log[\, L(X_m \mid p_m)] - \log\left[\frac{1}{K-1} \sum_{\substack{k=1 \\ k \ne m}}^{K} \exp(\, \gamma \log[\, L(X_m \mid p_k)]) \right]^{1/\gamma} \tag{6}$$

where the first class expressed in the first term represents the case of correct subword recognition. The second class expressed in the second term is the complement of the first class and represents the case of incorrect subword recognition.

Assuming independence, the utterance level likelihood ratio for word W can be written as a product of subword level likelihood ratios as in (7).

$$T(X; W) = \prod_{m=1}^{M} T(X_m; p_m) \tag{7}$$

Taking the log of (7) followed by applying (6) results in (8).

$$V(X; W) = \sum_{m=1}^{M} V(X_m; p_m) \tag{8}$$

The utterance is rejected or accepted by comparing $V(X;W)$ with a predefined threshold.

3.2 Frame level confidence measure

In the context of subword recognition, W can be expressed as a concatenation of subword units as in (2), further, it can be expressed as a concatenation of HMM states of each subword that can be written as

$$W = p_1 \cdots p_m \cdots p_M = s_1^{(1)} \cdots s_N^{(1)} \cdots s_j^{(m)} \cdots s_1^{(M)} \cdots s_N^{(M)} \quad (9)$$

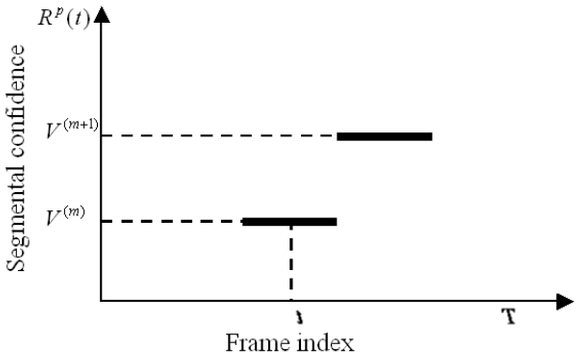

Figure 2. An example of phoneme level segmental confidence on input speech; $R^p(t)$ is a confidence score at t-th frame and $V^{(m)}$ is a confidence score of the segment for m-th phoneme.

where N is the number of states comprising a subword. Suppose that a vector x_t is generated from a state $s_j^{(m)}$, a frame-level confidence measure $V_t(x_t \mid s_j^{(m)})$ can be written as

$$V_t(x_t; s_j^{(m)}) = \log[\, L(x_t \mid s_j^{(m)})] - \log\left[\frac{1}{K-1} \sum_{\substack{k=1 \\ k \neq m}}^{K} \exp(\gamma \log[\, L(x_t \mid s_j^{(k)})]) \right]^{1/\gamma} \quad (10)$$

3.3 Phoneme and HMM state level segmental confidence measure

Based on the frame-level confidence measure that was defined previously, an HMM state-level segmental confidence measure can be defined as in (11) by averaging frame-level confidence scores on a given state segment.

$$V_j^{(m)} = \frac{1}{T_j^{(m)}} \sum_{x_t \in s_j^{(m)}} V_t(x_t; s_j^{(m)}) \quad (11)$$

where $T_j^{(m)}$ is the number of frames composing the given state segment. As a function of t, the state-level confidence can be expressed as in (12).

$$R^s(t) = V_j^{(m)}, \quad x_t \in s_j^{(m)} \quad (12)$$

Similarly, a phoneme-level segmental confidence measure can be defined as in (13) by averaging frame-level confidence scores on a given phoneme segment.

$$V^{(m)} = \frac{1}{T^{(m)}} \sum_j \sum_{x_t \in s_j^{(m)}} V_t(x_t; s_j^{(m)}) \quad (13)$$

where $T^{(m)}$ is the number of frames composing the given phoneme segment. Also, as a function of t, the phoneme-level confidence can be expressed as in (14).

$$R^p(t) = V^{(m)}, \quad x_t \in p_m \quad (14)$$

An example for the phoneme-level segmental confidence as a function of t is shown in figure 2.

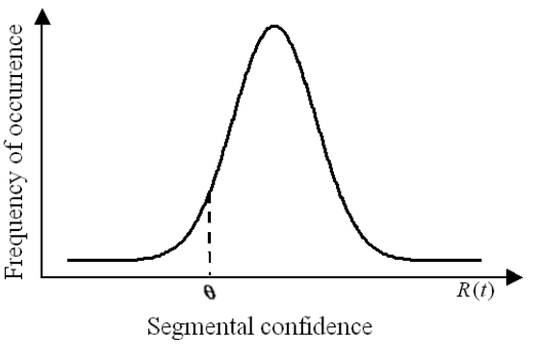

Figure 3. Distribution of segmental confidence scores.

4. CONFIDENCE NORMALIZATION

There are no upper and lower bound for a confidence score making it improper to be used as a weight factor directly. In this work we normalized the score to have the value between 0 and 1. Let us simplify the segmental confidence function $R^p(t)$ or $R^s(t)$ as $R(t)$. Figure 3 shows the distribution of $R(t)$ values. In the figure, the confidence gets higher from left to right. θ is a threshold value, which will be used for the normalization.

We propose two kinds of normalization functions. The first one is a hard limit function (HLF) in which all confidence scores that are lower than the threshold take a value $\varepsilon << 1$, and those that are higher than the threshold take the value $1-\varepsilon$. According to this function, a segment that has a very low confidence value because of the severe contamination does not contribute to the likelihood score while the other segments contribute equally to the likelihood score. The HLF is written as

$$R_N(t) = \begin{cases} 1-\varepsilon & \text{if } R(t) \geq \theta \\ \varepsilon & \text{if } R(t) < \theta \end{cases} \quad (15)$$

where θ is a threshold and is set as $\mu - \sigma$, where μ and σ are the mean and the standard deviation of the confidence distribution that is shown in figure 3.

The second normalization function is a threshold logic function (TLF) where the confidence values that lie between the upper and lower thresholds are linearly transformed to a value between $1-\varepsilon$ and ε, and those that are out of the interval take either $1-\varepsilon$ or ε as their normalized confidence value. The TLF is written as

$$R_N(t) = \begin{cases} 1-\varepsilon & \text{if} \quad R(t) \geq \theta_H \\ \dfrac{R(t)-\theta_L}{\theta_H - \theta_L} & \text{if} \quad \theta_L \leq R(t) < \theta_H \\ \varepsilon & \text{if} \quad R(t) < \theta_L \end{cases} \quad (16)$$

where $\theta_H = \mu + \sigma$ and $\theta_L = \mu - \sigma$ are the upper and the lower threshold respectively. According to the TLF, a segment that has a confidence value lower than or higher than the threshold values does not contribute at all or fully contribute to

the likelihood score while a segment whose confidence value lies between the two thresholds contribute to the likelihood score in proportion to their confidence values.

5. CONFIDENCE WEIGHTING VITERBI ALGORITHM

In this section we introduce a modified Viterbi algorithm that uses the normalized confidence as a weighting factor. The algorithm is as follows. The notations have the same meaning as in [9].

Step 1: Initialization. For each state i of HMM,

$$\delta_1(i) = \pi_i \times [b_i(x_1)]^{R_N(1)}$$
$$\psi_1(i) = 0$$

Step 2: Iteration. For $2 \le t \le T$ and $\forall j$,

$$\delta_t(j) = \max_i \; [\delta_{t-1}(i) \times a_{ij}] \times [b_j(x_t)]^{R_N(t)}$$
$$\psi_t(j) = \arg\max_i \; [\delta_{t-1}(i) \times a_{ij}]$$

Step 3: Termination.

$$P^* = \max_{s \in s_f} \; [\delta_T(s)]$$

In the algorithm, the normalized confidence value $R_N(t)$ controls the degree of contribution for a given frame. For $R_N(t) \approx 1$, the output probability of a given frame contributes fully to the likelihood score. Similarly, for $R_N(t) \approx 0$, the output probability does not contribute to the likelihood score.

The overall procedure of the proposed segmental confidence weighting (SCW) is as follows. Firstly N-best candidates are selected using the Viterbi algorithm [9]. Using the most likely word, the state-level segmentation is performed followed by the segmental confidence calculation. The confidence values are then normalized and the modified Viterbi algorithm chooses the best word among the N-best candidates. This 2-stage procedure is shown in figure 4.

6. EXPERIMENTAL RESULTS

For the verification of the proposed segmental confidence weighting method, a speaker independent isolated word recognition test was done. 100 words were selected from a 452 phoneme balanced word (PBW) database. The training database consisted of 25 male and 25 female speakers. For the test database 3000 utterances of 30 male and female speakers were used. The test data were contaminated artificially by the white burst noises of SINR 10, 5, 0, -5, -10 dB. Besides, at a given SINR, 5, 10 and 20 percent of a signal was partly contaminated by a burst noise. The burst noise was used because it is easy to control the duration and the power of each burst. We used 12 MFCC (mel frequency cepstral coefficient), delta and acceleration of it together with energy, delta energy and acceleration energy. Monophone HMM models as well as triphone HMM models were generated using HTK 3.0 toolkit. Monophone models were used for the calculation of segmental confidence score and triphone models were used for the N-best calculation and for the modified Viterbi decoding.

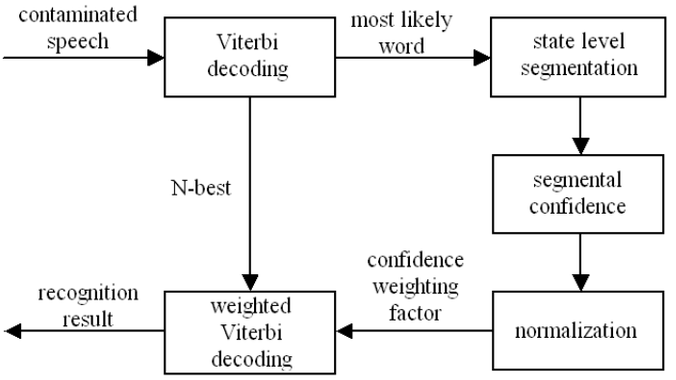

Figure 4. Overall 2-stage procedure based on the proposed SCW (segmental confidence weighting) method.

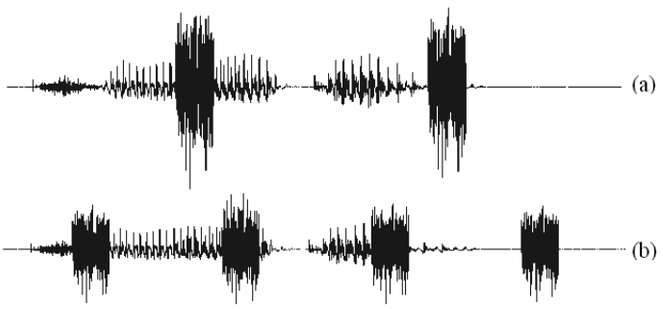

Figure 5. Examples of partly contaminated speech; (a) SINR –5 dB, 10% contamination (b) SINR –5 dB, 20% contamination.

Firstly, we investigated the effects of varying the fraction of contaminated segments and degree of contamination on the performance. The results are listed in Table 1. In this experiment, the duration of a burst is set to be 5% of the input length. Therefore 2 and 4 bursts correspond to a 10% and a 20% contamination respectively. An example of speech signals added by 10% and 20% burst at SINR –5 dB are shown in figure 5.

Table 1. Baseline word error rate on burst noise (with respect to several SINR and contamination percentage). The length of a burst is 5% of input speech signal

Burst rate	Singal to Impulsive Noise Ratio (SINR)				
	10 dB	5 dB	0 dB	-5 dB	-10 dB
5 %	5.7	9.3	14.4	22.7	46.4
10 %	7.5	16.1	27.4	40.5	56.9
20 %	9.0	20.5	39.2	59.1	76.4

According to the definition of SINR that was mentioned previously, at a given burst rate the SINR decreases as the power of a burst increases. In addition, for a fixed SINR the power of a burst decreases as the burst rate increases. In table 1,

the error rate increases as SINR decreases. At a given SINR, the error rate increases with increasing number of weaker bursts.

In the second experiment, the proposed 2-stage compensation based on the segmental confidence weighting (SCW) was evaluated and the two normalization functions were compared. The results are listed in the table 2. The results in table 2 in comparison to the results in table 1 reveals that the error rate decreased by the SCW method in most of the cases. The error rate was reduced by 12.5% on average for 5~20% burst rate at SINR –5 dB. As a normalization method, the TLF (threshold logic function) showed better performance than the HLF (hard limit function), which means that it is more effective to use a continuous value between 0 and 1 rather than to use the binary values as a confidential weighting factor.

Table 2. Word error rate of segmental confidence weighting method and comparison of the two normalization functions (NF); TLF (threshold logic function) and HLF (hard limit function)

Burst rate	NF	SINR (dB)				
		10	5	0	-5	-10
5 %	TLF	4.4	7.5	12.4	22.3	40.4
	HLF	4.7	8.4	13.5	24.1	47.5
10 %	TLF	7.7	13.4	20.2	30.7	46.3
	HLF	6.9	13.8	23.0	34.6	52.7
20 %	TLF	10.4	20.1	35.5	52.2	68.5
	HLF	9.5	20.1	37.6	56.3	73.9

7. CONCLUSION

In this paper, we proposed segmental confidence weighting (SCW) method to recognize a speech that is partly contaminated by burst noise. To make the scores from the segments of high confidence contribute more to the likelihood, the SCW normalizes HMM state-level segmental confidences and uses them as a weighting factor for a modified Viterbi algorithm. The proposed method does not require the estimation or the *a priori* knowledge of noise. The experimental results using the burst noises at various SINR showed that it could reduce the word error rate by 12.5% on average about various burst rates. A study on an improved method to obtain segmental confidence score is remained as a further work in addition to a study on a maximum utilization of normalization and weighting of confidence information.

REFERENCES

[1] Y. Gong, "Speech Recognition in Noise environments: A Survey," *Speech Communication*, vol. **16**, pp. 261-291, 1995.

[2] Zhong-Hua Wang, Patrick Kenny, "Speech Recognition in Non-Stationary Adverse Environments," Proc. IEEE Int. Conf., Acoustic Speech Signal Processing, pp. 265-268, 1998.

[3] H. Kobatake and Y. Matsunoo-, "Degraded Word Recognition based on Segmental Signal-to-Noise Ratio Weighting," Proc. IEEE Int. Conf., Acoustic Speech Signal Processing, pp. 425-428, 1994.

[4] N. B. Yoma, F. McInnes and M. Jack, "Weighted Matching Algorithms and Reliability in Noise Cancelling by Spectral Subtraction," Proc. IEEE Int. Conf., Acoustic Speech Signal Processing, pp. 1171-1174, 1997.

[5] N. B. Yoma, F. McInnes and M. Jack, "Improved Algorithms for Speech Recognition in Noise Using Lateral Inhibition and SNR Weighting," Proc. Eurospeech, pp. 461-464, 1995.

[6] 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 Processing*, Vol **4**. No. 6, pp. 420-429, 1996.

[7] N. Moreau and D. Jouvet, "Use of Confidence Measure Based on Frame Level Likelihood Ratios for the Rejection of Incorrect Data," Proc. Eurospeech, pp. 291-294, 1999.

[8] S. V. Vaseghi and B. P. Milner, "Speech Recognition in Impulsive Noise," Proc. IEEE Int. Conf., Acoustic Speech Signal Processing, pp. 437-440, 1995.

[9] X. D. Huang, Y. Ariki and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.