ROBUST SPEECH RECOGNITION AGAINST SHORT-TIME NOISE

Manhung Siu, Yu-Chung Chan

Hong Kong University of Science and Technology
Electrical Electronic Engineering
Clear Water Bay Road

ABSTRACT

This paper develops a robust speech recognition algorithm against short-time noise, of which no prior knowledge of their spectral characteristics is known. However, we assume that these noises only affects certain part of the speech and are also known as partially temporal corruption in [1]. Examples of these short-time noises include door slam, click sound of keyboard or packet loss in network transmission of voice. These noises are found to degrade the performance of an automatic speech recognizer (ASR) significantly. In our previous work [2], we proposed a robust algorithm, called frame-skipping Viterbi algorithm (FSVA), which ignores the likelihood contributions of the K worst performing frames during Viterbi decoding. We showed that FSVA algorithm is robust against random replacement of speech frames by Gaussian noise. One important issue that we have not addressed is the determination of the number of skips. This paper extends our work by applying the FSVA to *additive short-time* noise with *unknown spectral characteristic* and *unexpected occurrence* under different SNR, rate of corruption and the length of corruption. We also propose a solution for determining the appropriate number of skips on log likelihood ratio.

1. INTRODUCTION

Automatic Speech Recognition (ASR) performance is known to be degraded under noise conditions. These noise conditions include the corruption by stationary additive noise across the whole speech waveform such as air-conditioning noise, or short-time noises that affects only a portion of the speech waveform such as door slam, click sound of keyboard or packet loss in network transmission. Many researches focused on making speech recognition robust against additive noise such as the parallel model combination (PMC) proposed by Gales in [3]. For PMC, one first captured the statistical properties of noise and used the noise statistical properties to create a noisy-speech model that is adapted for a particular noise environment. However, for short-time noises that affect only a portion of the speech waveform, which is also called partially temporal corruption in [1], an alternative approach is to gather information from the unaffected portion of speech signal [1]. This approach is more attractive compared to noise compensation approach such as PMC because even if we can obtain the enough noise samples to capture the noise statistics, the noisy-speech model may not match with the unaffected (or clean) portion of the speech waveform. In [1], Ming and Smith suggested one proposal to make use of the unaffected part of speech for speech recognition. They partitioned the speech waveform as a sequence of consecutive segments and considered noise corruption at the segment-level (i.e. a segment is corrupted if any of its frames are corrupted). They proposed a criterion function, called segment-based union model, in which corrupted segments would only contribute small effect to the criterion function. It was so-called because the criterion effectively maximizes the disjunction of segments (and "or" operation) instead of conjunction of segments (and "and" operation). However, recognition of continuous speech using the segment-based union model could only be performed in the N-best paradigm.

In our previous work [2], we proposed a different solution based on getting a robust estimate on the observation likelihood of a given state sequence to deal with similar short-time corruptions. The idea is similar to the use of trimmed-mean in statistics. For each state sequence, we ignored contributions of the K least likely frames in the computation of likelihood. Because we are skipping corrupted frames, we can easily handle short-time corruptions of a couple of frames. We also developed a dynamic algorithm called frame-skipping Viterbi algorithm (FSVA), which is a modified version of the standard Viterbi algorithm, for performing searching for the best robust state sequence efficiently.

While we had shown that FSVA performed well in one special kind of short-time noise in which speech frames were randomly *replaced* by white noise, that however, assumed we could determine the appropriate number of skips for each sentence. The problem of determining number of skips is difficult because it depends on factors such as the rate of corruption and length of utterance which are not known a'priori in real-life situation. We have not addressed this problem in our previous paper.

In this paper, we extend our work on FSVA to handle *additive short-time noise* with *unknown spectral characteristic* and *unexpected occurrence* with different SNR, corruption rate and the length of the corruption. In addition, we propose an approach called log likelihood ratio thresholding (LLRT) to determine the appropriate number of skips automatically in different noise conditions.

The rest of this paper is organized as follows. In Sec. 2, we review the theory and implementation of FSVA. In Sec. 3, we describe the motivation and implementation of LLRT. for determining the number of skips. In Sec. 4, we describe the evaluation result in short-time additive noise. Finally, we conclude this paper and describe the future work in Sec. 5

2. FRAME-SKIPPING VITERBI ALGORITHM

As described in our previous paper [2], the underlying motivation is to get an estimate of the observation likelihood given a state sequence that is robust against outlier. In Sec. 2.1, we review the

This works is partially supported by Hong Kong Reseach Grant Council CERG Project number : HKUST 6049/00E

key steps in arriving at the resulting algorithm and describes some of its basic properties. In Sec. 2.2, we discuss the key limitation of our previous work which is the necessity for knowing the number of skip beforehand.

2.1. Theory of Frame-Skipping Viterbi Algorithm

HMM-based speech recognition is typically viewed as finding the most likely state sequence $\hat{Q}_1^T = (q_1, \ldots q_T)$ given the observation $O_1^T = (o_1, \ldots o_T)$ as:

$$\hat{Q}_1^T \quad = \quad \arg\max_{Q_1^T} \log p(\, Q_1^T \,|\, O_1^T\,). \tag{1}$$

By using the Viterbi algorithm, the above maximization can be done efficiently by computing the partial path score $\delta_t(i)$ that explains the observation O_1^t and that $q_t = j$. A recursive relation can be derived given as:

$$\delta_t(j) = \max_i [\delta_{t-1}(i)a_{ij}]\, b_j(o_t) \quad 1 \le t < T,$$

where $b_j(o_t)$ is the observation probability for observing o_t at state j.

From Equation 1, the maximization can be written as the sum of observation likelihoods and transition probabilities

$$\hat{Q}_1^T = \arg\max_{Q_1^T} \underbrace{\sum_t \log p(o_t|q_t)}_{\text{obs. likelihood for } Q_1^T} + \underbrace{\sum_t \log p(q_t|q_{t-1})}_{\text{transition prob.}}.$$

2.1.1. Robust Likelihood

To find a state sequence that is more robust against corruption, we denote the robust log likelihood of a path Q_1^T by its best $T-K$ frames, denoted as $\phi_K(\,O_1^T|Q_1^T)$. Suppose we compute the rank-order, $t_1, \ldots, t_T$, based on the observation likelihood such that

$$p(o_{t_i}|q_{t_i}) \le p(o_{t_{i+1}}|q_{t_{i+1}}).$$

Then, instead of summing the observation likelihood over time, we can sum over t_i and rewrite $\phi_K(\,O_1^T|Q_1^T)$ as

$$\phi_K(\,O_1^T|Q_1^T) \quad = \quad \sum_{i=K+1}^{T} \log p(o_{t_i}|q_{t_i}). \tag{2}$$

In speech recognition, because of the high dimensionality and sparseness of the space, it is usually true to assume $p(o_t|q_t) < 1$ for all t. Then, for a give path Q, since

$$\begin{aligned} \phi_K(\,O_1^T|Q) &= \phi_{K+1}(\,O_1^T|Q) + \log p(o_{t_{K+1}}|q_{t_{K+1}}), \\ &\le \phi_{K+1}(\,O_1^T|Q). \end{aligned} \tag{3}$$

2.1.2. Frame-Skipping Viterbi Algorithm

In recognition, we search over all possible state sequences to find the best robust path when K frames are skipped, denoted as $\hat{Q}_1^T(K)$. $\hat{Q}_1^T(K)$ is given by,

$$\hat{Q}_1^T(K) = \arg\max_{Q_1^T} \left(\sum_{t=1}^{T} \log p(q_t|q_{t-1}) + \phi_K(O_1^T|Q_1^T). \right) \tag{4}$$

Note that in this process, we only ignore contributions of the least likely frames to the total observation likelihood computation but retain contributions of state transitions and the length of the state sequence remains T. This preserves the duration information.

A recursive relationship, similar to the one in Viterbi algorithm can be obtained. We define the partial path score that explains the first t observations with k skips at state j as $\delta_t(j,k)$. At each state, K+1 different partial path scores are needed to store the likelihood scores for 0 to K skips. When we compute $\delta_t(j,k)$, in addition to finding the best previous state, we also decide whether the current frame o_t should be skipped.

If we skip the current frame, denoting the partial path score for skipping the current frame as $\delta_t^s(j,k)$.

$$\delta_t^s(j,k) = \max_i a_{ij}\delta_{t-1}(i,k-1).$$

If we do not skip the current frame, denoting the partial path score for not skipping the current frame o_t as $\delta_t^n(j,k)$.

$$\delta_t^n(j,k) = \max_i a_{ij}\delta_{t-1}(i,k)b_j(o_t).$$

Combining the above two equations, we have

$$\begin{aligned} &\delta_t(j,k) \\ &\quad = \max(\delta_t^s(j,k), \delta_t^n(j,k)), \\ &\quad = \max_i [a_{ij} \max(\delta_{t-1}(i,k-1), \delta_{t-1}(i,k)b_j(o_t))]\,. \end{aligned} \tag{5}$$

We called this the frame-skipping Viterbi algorithm (FSVA) and can be easily be implemented as an extension to standard Viterbi algorithm by expanding the state space. The detail of this can be found in [2].

2.1.3. K Solutions

Let's denote the ending node of the network as V. If we look at the partial path at time T at node V, $\delta_T(V,k)$, we actually compute all the best robust paths $\hat{Q}_1^T(1), \ldots, \hat{Q}_1^T(K)$ and robust likelihoods $\phi_1(\,O_1^T|\hat{Q}_1^T(K)), \ldots, \phi_K(\,O_1^T|\hat{Q}_1^T(K))$ for skipping 0 to K frames.

Based on the result from Eqn. 3, we have

$$\begin{aligned} \phi_{K+1}(\,O_1^T|\hat{Q}_1^T(K+1)) &\ge \phi_{K+1}(\,O_1^T|Q), \quad \forall Q \\ &\ge \phi_{K+1}(\,O_1^T|\hat{Q}_1^T(K)), \\ &\ge \phi_K(\,O_1^T|\hat{Q}_1^T(K)). \end{aligned}$$

This relationship is found to be useful for finding the current number of skips.

2.2. Limitation

We note that the above formulation is based on the knowledge of K, the number of skip. Determining its appropriate value is crucial to the performance of FSVA because skipping too few frames will limit the effect of FSVA and skipping too many frames will cause degradation. In this paper, we propose a method of automatic skip determination called log likelihood ratio thresholding (LLRT) and showed that the appropriate number of skips can be determined automatically on a sentence by sentence basis.

3. AUTOMATIC SKIP DETERMINATION USING LOG LIKELIHOOD RATIO (LLRT)

There are a number of ways to select the appropriate number of skips. If the "amount" of corruption is known a'priori, we can intuitively skip fixed number of frames or a fixed percentage of frames. However, in more realistic situation, it is desirable to have a selection algorithm to decide the number of skips for each sentence without any prior knowledge of the environment.

In Section 2.1.3, we explained that when we search for the solution for N skips, the solutions for $1, \ldots, N-1$ skips are also computed. We can then interpret the finding the appropriate number of skip $\hat{K}$ as picking the best out of the N solutions. Since the robust log likelihood increases as K increases, an intuitive solution is to compute the relative change in the robust likelihood and compare it with a threshold c given by,

$$\phi_K(\, O_1^T|\hat{Q}_1^T(K)) - \phi_{K+1}(\, O_1^T|\hat{Q}_1^T(K+1)) > c. \qquad (6)$$

Since $\phi_K(\, O_1^T|\hat{Q}_1^T(K))$ is the robust likelihood of skipping K frames, we denote the left-hand-side of Eqn. 6 as $LLR(K, K+1)$ and Eqn 6 as log likelihood ratio (LLR). Fig. 1 shows a typical plot of $LLR(K, K+1)$ against K, we will observe that there is a general decreasing trend of $LLR(K, K+1)$. The value of c has an intuitive meaning as the minimum likelihood of a clean frame and this can be a function of the model.

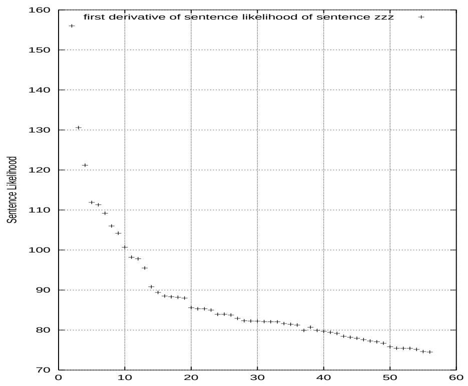

Fig. 1. $LLR(K-1, K)$ against the number of skips,K, in sentence "zero zero zero" which is corrupted by Gaussian noise.

With the results explained in Section 2.1.3, implementing the LLRT is straightforward.

4. EXPERIMENT

We performed a number of experiments that aims to 1) evaluate the effectiveness of LLRT for determining the correct number of skip, and 2) test the usefulness of FSVA under different noise conditions.

Before we describe the experiments and results in Sec 4.3, we first describe the baseline experiment and the details on training and evaluation in Sec. 4.1 as well as our procedure of noise addition in Section 4.2.

4.1. Baseline Experiment

All experiments were performed using the TI-digits corpus. FSVA was evaluated under continuous digit recognition (CDR) with the same HMM topologies used AURORA baseline experiment [4]. The acoustic features included 12 cepstral coefficients, their energy and the corresponding delta and acceleration coefficients. Only adult sentences of the training and testing sets were used resulting in a total of 8668 digit utterances for training and 8668 for test. Each utterance was a digit string with a maximum length of 7 digits and up to 3 seconds long. We followed the training procedure as suggested in HTK Handbook. The baseline results were 99.72% in isolated word recognition and 98.90% in CDR. [1]

4.2. Procedure of Noise Addition

In our experiments, short-time noise could be characterized by different attributes.

- **Signal to noise ratio (SNR)** can be computed as the ratio between the power of speech signal and noise in a sentence.
- **Rate of corruption (C)** is measured as the percentage of frames corrupted.
- **Length of corruption (L)** is the duration of the noise. We measure in number of frames.

To simulate the situation in which we have no prior knowledge of noise spectrum, we randomly picked one of the seven noises for corruption for each sentence. The seven different noises included 4 telephone ring tones plus 1 whistling noise found in [5], and two message sound clips in ICQ application. After selecting the noise, we then randomly selected frames (frame length 20ms) from this noise and add it to speech portions of the utterance in predefined SNR ,S, and rate of corruption ,C. This noise addition ensured that the noise are unexpected and unknown.

4.3. Experiments

We evaluated the use of LLRT in skip determination under different noise conditions. Tables 1, 2 and 3 are the results of using FSVA together with LLRT under different noise conditions. The performances were obtained using the best LLRT threshold in each of the conditione.

As shown in Tables 1, lower SNR caused more degradation to ASR performance but FSVA is effective because it might be easier to identify and skip the noisy frames. When $SNR > 0dB$, although FSVA was not very effective, it still did not degrade performance indicating that FSVA can be used at all times.

In Table 2 , we varied the rate of corruption between 20% to 40% respectively at $SNR = -10dB$ and the length of noise $L = 1$. As the rate of corruption increased, the ASR performance degraded and so was the effectiveness of FSVA because fewer clean frames were available in the decision process. However, at all rate of corruption, FSVA together with LLRT consistently improved performance.

The results for varying the length of corruption were shown in Table 3 from 1 to 8 frames in which we kept $C = 20\%$ and $SNR = -10dB$. Shorter length of corruption caused more damage to the utterances if the rate of corruption was kept constant because it affected more phonetic units making recognition harder.

[1] The baseline result of CDR can be over 99% if we tuned the insertion penalty of the recognizer.

SNR	BL	LLRT(optimal)
∞	98.90	98.99(102)
10	98.62	98.67(106)
0	97.57	97.99(102)
-10	84.04	91.89(94)

Table 1. Performance of LLRT in noise with varying SNR and $C = 20\%, L = 1$. Key : BL : No frames are skipped. LLRT(100) : number of skips is determined by LLRT using threshold optimal value of c (The number inside the bracket shows c). $SNR = \infty dB$ illustrate the performance of strategies in clean utterances.

C	BL	LLRT(optimal)
20%	84.04	91.89(94)
30%	69.21	82.96(94)
40%	56.39	71.15(94)

Table 2. Performance of LLRT in noise with varying rate of corruption and $SNR = -10dB, L = 1$. Key : C : rate of corruption ,BL : No frames are skipped. LLRT(optimal) : number of skips is determined by LLRT with the optimal value of c (The number inside the bracket shows c)

In the above experiments, the optimal LLRT thresholds were used. It is very important to see whether this is necessary. Figure 2 shows the average performance over different noisy conditions tabulated in Tables 1, 2 and 3 by using a single fixed threshold is used for all conditions. The dashed line is the baseline performance. The solid line is the accuracy obtained using various LLRT thresholds. The dotted line is the optimal performance using the optimal value of c. The plot shows that there is a wide range of value of c that can improve recognition performance over the baseline. More importantly, the average performance using any values from the range is not much worse than the optimal performance obtained by using different LLRT thresholds for different environments. For example, we could use any threshold from range $[90, 100]$ for all conditions which gives only slight difference from the optimal performance. As we have discussed in Sec. 3, the value c has the meaning of the minimum likelihood for a clean frame and is dependent on the models. However, the above results suggest that by tuning the value of c in a single condition, we can obtain a reasonable range of value to use under different noise environments.

5. CONCLUSION AND FUTURE WORK

In our previous work, we have introduced the FSVA to handle corruption when speech frames are replaced by Gaussian noise. This paper extends the previous work to unknown, unexpected short-time additive noise under different SNRs, rates of corruption and lengths of corruption. We show that FSVA always outperforms the baseline and in no case that it degrades performance. Furthermore, we have introduced an algorithm called log likelihood ratio thresholding (LLRT) that can automatically determine which the appropriate number of skips. By using the correct range of values of log likelihood ratio, we could obtain good result in the different conditions of short-time noise.

As we mentioned in Sec. 3, finding the solution for K skips also generates the solution for smaller number of skips. However, this requires performing FSVA at a high value of K to make sure that we can find a k that the log-likelihood ratio is larger than the threshold. Thus, one of our future work is to incorporate LLRT into the dynamic programming algorithm of FSVA such that we can efficiently determine the number of skips required.

length	BL	LLRT(optimal)
1	84.04	91.89(94)
2	87.33	91.99(100)
3	90.24	94.07(94)
4	93.30	96.22(92)
5	95.17	97.09(94)
6	95.69	97.31(92)
7	95.49	97.22(94)
8	95.38	96.85(94)

Table 3. Performance of LLRT in noise with varying length and $SNR = -10dB, C = 20\%$. The third column describes the evaluation result of LLRT with the optimal vealue of c (The number inside the bracket shows c).

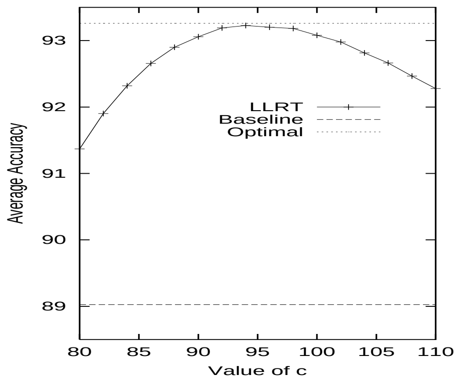

Fig. 2. Average acurracy of all conditions against value of c

6. REFERENCES

[1] J. Ming and F. J. Smith, "Union : a model for partial temporal corruption of speech," *Computer Speech and Language*, vol. 15, 2001.

[2] M. Siu and Y. C. Chan, "Robust speech recognition against packet loss," in *Eurospeech*, 2001.

[3] M.J.F Gales, *Model-based techniques for noise robust speech recognition*, Ph.D. thesis, University of Cambridge, September 1995.

[4] H.G. Hirsch and D. Pearse, "The aurora experimental framework for the performance evaluation of speech recognitions system under noisy conditions," in *ASR 2000 Automatic Speech Recogntion : Challenges for the Next Millenium.*, 2000.

[5] J. Ming, "Ji. ming : Noisy speech recognition," Can be downloaded at http://www.cs.qub.ac.uk/ J.Ming/Html/Robust.htm.