Improving the Robustness of LPCC Feature Against Impulsive Noise by Applying the FOP Method

Pei Ding
Toshiba (China) Research and Development Center
5/F Tower W2, Oriental Plaza, No.1 East Chang An Avenue,
Dong Cheng District, Beijing 100738
dingpei@rdc.toshiba.com.cn

Abstract. Performance of an automatic speech recognition (ASR) system tends to be dramatically degraded in the presence of impulsive noise. In the previous work [1], we proposed flooring the observation probability (FOP) to compensate the adverse effect of impulsive noise on sensitive dimensions of Mel-frequency cepstral coefficient (MFCC) features. Linear prediction cepstral coefficient (LPCC) is another kind of widely used acoustic feature, and in this paper we study the performance of the FOP method when applied to LPCC features, including feature vector partition based upon noise sensitivity analysis of each feature dimension and flooring threshold calculation. Evaluation results confirm the efficiency of FOP method on LPCC feature. For example, the highest averaged error reduction rate (ERR) of 38.9% and 46.8% versus the baseline is obtained, respectively in simulated substitutive impulsive noise and machinegun noise environment.

Keywords: Robust Speech Recognition, Impulsive Noise, Observation Probability, Flooring Threshold, LPCC acoustic feature

1 Introduction

In recent years the considerable development of wireless and computer networks leads to a rapid growth of the speech recognition applications between the server and the client through such communication environments [2][3]. A representative application is interactive remote information inquiry. However, due to some inherent characteristic of unreliable data delivery, transmission errors and packet loss usually cause impulsive noise corruption, which consists of extremely non-stationary random short-time burst and tends to drastically degrade the recognition performance. Therefore, how to improve the robustness of ASR systems against impulsive noise is essential for such applications.

Several methods have been proposed with this issue. Some of them are based upon the idea that an impulsive noise introduces uncharacteristic discontinuity in correlated speech signal, so the outlying data can be detected and removed before it is fed into the recognizer [4][5]. In some other methods, the negative effect of impulsive noise is
compensated in decoding stage by skipping the unreliable frames or decreasing their contributions in likelihood calculation [6][7].

In our previous work [1], a FOP algorithm is proposed for impulsive noise compensation in recognition stage by flooring the observation probabilities of noise sensitive MFCC feature sub-vector at Gaussian mixture level. Feature partition and flooring threshold assignment are two critical steps for FOP method, which aim at eliminating the unreliable observation probability difference while at the fewest expense of likelihood distortion in robust feature dimensions and in clean speech.

LPCC is another kind of important and widely used features, especially in speaker recognition and verification applications. Unlike the spectrum-based representation in MFCC, the LPCC approximated the vocal tract spectral responds of the speakers with an all-poll modeling constraint. In this paper we apply the FOP method to LPCC feature to improve its robustness against impulsive noise, including feature partition based on noise sensitive dimension analysis and flooring threshold calculation for sensitive sub-vector. Evaluation results show that the FOP method efficiently improves the performance of LPCC feature in the presence of impulsive noise.

The rest of the paper is organized as follows. Section 2 reviews the theory of FOP algorithm. Section 3 describes the feature vector partition based upon the analysis of noise sensitivity dimension. Section 4 describes the calculation method of the flooring threshold. Section 5 describes the experiments and section 6 concludes the paper.

2 Description the FOP Method

2.1 Effect of Impulsive Noise on ASR

In a hidden Markov model (HMM) based ASR system, the input speech signals are converted into a series of feature vectors in front-end stage, and then the probabilities of these acoustic observations of each candidate model are calculated in decoding. The observation probability indicates the likelihood that the observation is a realization of the acoustic model, and the recognizer finally gives the state sequence with the highest probability, which can be realized by the recursive Viterbi algorithm:

\[
\delta_t(j) = \max_{i} [\delta_{t-1}(i) a_{ij}] b_j(X_t)
\]  

where \(\delta_t(j)\) is the maximum likelihood of observation feature vectors \(X_t\) to \(X_i\) given state \(j\) at time \(t\), \(a_{ij}\) is the transition probability from state \(i\) to state \(j\), and \(b_j(X_t)\) is the observation probability of \(X_t\) at state \(j\).

Because the observation probability represents the similarity between the feature vector and a candidate model, the sudden mismatch introduced by impulsive noise tends to cause abnormal distribution of the observation probabilities. Fig. 1(b) and (C) compare the log probabilities of the noise corrupted observations given the state sequence obtained by running the baseline recognizer on clean speech and the state sequence obtained by running the recognizer on noisy speech. The feature used in Figure 1(b) is 13-dimension static LPCC denoted as \([L_0, \cdots, L_{12}]\) and in Figure 1(c) the sensitive feature sub-vector \([L_0, L_2, L_4, L_5, L_7, L_9]\)
Fig. 1. Effect of impulsive noise on observation probability. (a) Speech Spectrogram: partial frames are corrupted by impulsive noise; (b) Log probabilities of corrupted observations in all LPCC dimensions: \([L_1, \cdots, L_{13}]\); (c) Log probabilities of corrupted observations in partial LPCC dimensions: \([L_1, L_2, L_4, L_5, L_7, L_8]\).

In Figure 1(b) we can observe that in noisy period the observation probabilities in the expected state sequence are much smaller than the actual one, i.e. the incorrect path here acquires unreliable probability difference. We hope the recognizer can translate the noisy observations into the expected state sequence as it decodes the clean counterpart, but the relative probability difference between the two sequences prevents the token from propagating in the expected path. In Figure 1(c) we can find that the unreliable probability difference is concentrated on sub-vector \([L_1, L_2, L_4, L_5, L_7, L_8]\). It is obvious that FOP with a proper threshold strictly on such sensitive sub-vector can directly reduce the probability difference during noise frames while maintaining adequate discriminating capability in clean period. Thus, the expected path will recover the priority of being chosen in decoding.
2.2 FOP at Gaussian Mixture Level

We assume that the probability density function (PDF) in each state $j$ can be adequately modeled using mixtures of $M$ Gaussian distributions with diagonal covariance matrix:

$$b_j(X) = \sum_{k=1}^{M} w_k N(X; \mu_k, \Sigma_k)$$

(2)

where $X = [x_1, x_2, \ldots, x_N]$ is the $N$-dimension feature vector, $N(X; \mu_k, \Sigma_k)$ is the multivariate Gaussian distribution, $\mu_k = [\mu_{k1}, \mu_{k2}, \ldots, \mu_{ki}]$, $\Sigma_k = \begin{bmatrix} \sigma_{k1}^2 & \cdots & 0 \\ 0 & \cdots & \sigma_{ki}^2 \end{bmatrix}$, and $w_k$ represent the mean, variance and weight of the $k$th Gaussian component. According to the noise sensitivity of each feature dimension, the observation vector $X$ can be divided into $L$ sub-vectors:

$$X = [x_1, x_2, \ldots, x_L] = [X_1, X_2, \ldots, X_L] .$$

(3)

Since the independence of each feature dimension is satisfied at the Gaussian mixture level, we can separate the distribution of each feature sub-vector from the total PDF in this level and floor it with a proper threshold:

$$b'_j(X) = \sum_{l=1}^{L} \prod_{i=1}^{N} N(X_i; \mu_{li}, \Sigma_{li}),$$

(4)

where $N_j(X_i; \mu_{li}, \Sigma_{li})$ is the floored Gaussian distribution of $X_i$ with threshold $T_l$:

$$N_j(X_i; \mu_{li}, \Sigma_{li}) = \begin{cases} N(X_i; \mu_{li}, \Sigma_{li}) & \text{if } N(X_i; \mu_{li}, \Sigma_{li}) \geq T_l \\ T_l & \text{otherwise} \end{cases} .$$

(5)

The floored observation probability $b'_j(X)$ directly replaces $b_j(X)$ in Viterbi decoding described in Eq.(1). Thus, the unreliable probability difference caused by impulsive noise is effectively eliminated.

3 Noise Sensitivity Analysis and Feature Partition

We suppose that the $i$th-dimension LPCC feature of clean speech follows a Gaussian distribution:

$$p(x_i) = N(x_i; \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right),$$

(6)

where $\mu_i$ and $\sigma_i^2$ are roughly estimated by averaging the mean and variance parameters of all speech HMMs. We make the global analysis of noise sensitivity for each LPCC dimension based upon the log likelihood distance to noise, which is normalized by speech variance:

$$D_i = \log\left( \frac{N(\mu_i; \mu_i, \sigma_i^2)}{N(\mu_i; \mu_i, \sigma_i^2)} \right) = \frac{(\mu_{ni} - \mu_i)^2}{2\sigma_i^2},$$

(7)

where $\mu_{ni}$ is the average of the $i$th noise LPCC.
The $D_i$ with larger value indicates the higher probability drop when encountering outlying noises, and the more sensitive this feature dimension is. Figure 2 shows the likelihood distance in log scale of both static LPCC and delta LPCC. We can find that delta LPCC is relatively more robust and the most six sensitive LPCC dimensions sorted by $D_i$ are $[L_1, L_2, L_5, L_8, L_4, L_7]$. Figure 1(c) confirms the efficiency of noise sensitivity analysis. Thus, the noise sensitive feature sub-vector, whose observation probability is strictly floored, should selected among the above 6 dimensions.

4 Threshold Calculation for Noise Sensitive Feature Sub-vector

In Eq.(5) the flooring threshold $T_i$ is very crucial to FOP algorithm. Using a higher threshold that strictly flooring the $b_i(X)$ can more efficiently eliminate the unreliable probability scores in noise period, but also impairs more discriminating information in clean period. FOP with a lower threshold results in the opposite effect. Therefore, the optimum threshold should make the best compromise between discriminative capability and the robustness to outlying feature. In this paper we propose calculating the thresholds as follows:

For a $R$-dimension sensitive sub-vector $X_f=[x_{j_1}, \cdots, x_{j_R}]$, the threshold $T$ can be described by the following two equations:

$$N(X_f;\mu_f, \Sigma_f)|_{x_f \in C} = T \quad \text{(8)}$$

$$\int_{H} N(X_f;\mu_f, \Sigma_f)dx_f = P \quad \text{(9)}$$

where $H$ is the integration space enclosed by the contour $C$ given by Eq.(8), $\mu_f$ and $\Sigma_f$ are the averaged mean and variance, respectively. $T$ should satisfy that the confidence of $H$ under $N(X_f;\mu_f, \Sigma_f)$ distribution is $P$. Fig. 3 gives a demonstration of threshold calculation projected in two dimensions: $x_p$ and $x_q$. 

![Fig. 2. Noise sensitivity analysis for each LPCC dimension](image)
Considering the project of Eq.(8) in the $i$th dimension:

$$N(X_i;\mu_i,\Sigma_i)\big|_{x_{i}^2\sigma_i^2=0} = \frac{1}{\sqrt{2\pi \sigma_i}} \exp\left(-\frac{d_i^2}{2\sigma_i^2}\right) = T, \quad i = 1, 2, \cdots, R,$$

where $d_i$ is the radius of the hyper-ellipse $C$ in the $i$th dimension of $X_i$, we can get

$$d_i/\sigma_i = \sqrt{2 \log(G/T)} \quad \forall i = 1, 2, \cdots R$$

where $G = \prod_{i=1}^{R} \sqrt{2\pi \sigma_i}$.  

If we use a tangential hyper-rectangle $\tilde{H}$ to substitute $H$ in Eq.(9), the integration can be approximated by the production of the confidence of $[\mu_i - d_i, \mu_i + d_i]$ in each dimension:

$$\int_{\tilde{H}} N(X_i;\mu_i,\Sigma_i) dX_i = \prod_{i=1}^{R} \int_{-d_i}^{d_i} N(x_i;0,\sigma_i^2) dx_i = \prod_{i=1}^{R} \int_{-d_i}^{d_i} N(\xi_i;0,1) d\xi_i$$

$$= \prod_{i=1}^{R} \int_{-\sqrt{\log(G/T)}}^{\sqrt{\log(G/T)}} N(\xi_i;0,1) d\xi_i = (2 \cdot \Phi(\sqrt{2 \log(G/T)}) - 1)^R = P$$

where $\xi_i = x_i/\sigma_i$, and $\Phi(.)$ is defined as:

$$\Phi(x) = \int_{-\infty}^{x} N(x;0,1) dx = \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{1}{n!} \left(-\frac{1}{2}\right)^n x^{2n+1}$$

From Eq.(12), we can find that the project of $H$ in each dimension, $[\mu_i - d_i, \mu_i + d_i]$, has the identical confidence:

$$\int_{-d_i}^{d_i} N(x_i;0,\sigma_i^2) dx_i = 2 \cdot \Phi(\sqrt{2 \log(G/T)}) - 1 = \sqrt{P} \triangleq P_i.$$

We can calculate the corresponding threshold $T$ by Eq.(14) after assigning the $P_i$, which is related to the training data and the acoustic feature. In our experiments, for those mentioned sensitive LPCC we find $P_i = 97.9\% = 2 \cdot \Phi(2.3078) - 1$ is suitable.
5 Experiments

5.1 Experiment Configurations

The proposed method is evaluated on the TI-Digit database in simulated substitutive impulsive noise and additive machinegun noise environments. Substitutive impulsive noise environment[6] is generated by randomly replacing 25 ms speech frames by white noise with the same width under different occurrence rates (OR). In this noisy experiment, because the recognition performance degrades mainly owing to the OR of the impulsive noise, we fixed the noise to the same energy level of clean speech. Machinegun noise contaminated speech is generated by artificially adding 62.5ms noise segments with different ORs and signal-to-impulsive noise ratios (SINR) to clean speech [4]. The definition of SINR is as follows:

\[
\text{SINR}(dB) = 10\log_{10}(\frac{P_{\text{signal}}}{OR \times P_{\text{input}}})
\]

where \(P_{\text{input}}\) denotes the average power of each impulse and \(P_{\text{signal}}\) the signal power.

The training set and testing set includes 500 and 100 clean utterances, respectively. Digits and silence were respectively modeled by 10-state and 3-state whole word HMMs with 4 diagonal Gaussian mixtures in each state. The acoustic features were 13 LPCC+13ΔLPCC, i.e. \([L_1,...,L_{13},\Delta L_1,...,\Delta L_{13}]\). For static LPCC, the probability degradation is mainly concentrated on \([L_1,...,L_8,\Delta L_1,...,\Delta L_8]\), we evaluate several schemes which selected among the above six dimensions to form the sensitive sub-vector \(X_1\) and calculate the threshold \(T_1\) by Eq.(14) under \(P_i = 97.9\%\).

Table 1: Feature partition schemes and threshold assignment

<table>
<thead>
<tr>
<th>S1</th>
<th>Sub-vector (X_1)</th>
<th>(T_1)</th>
<th>Sub-vector (X_2)</th>
<th>(T_2)</th>
<th>Sub-vector (X_3)</th>
<th>(T_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>(L_1, L_2, L_3)</td>
<td>-6.1</td>
<td>(L_4, L_5, L_6, L_7, L_8)</td>
<td>-(\infty)</td>
<td>(\Delta L_1, \Delta L_2)</td>
<td>-20.0</td>
</tr>
<tr>
<td>S2</td>
<td>(L_1, L_2, L_3, L_4)</td>
<td>-7.5</td>
<td>(L_5, L_6, L_7, L_8, L_9)</td>
<td>-(\infty)</td>
<td>(\Delta L_1, \Delta L_2)</td>
<td>-20.0</td>
</tr>
<tr>
<td>S3</td>
<td>(L_1, L_2, L_3, L_4, L_5)</td>
<td>-8.8</td>
<td>(L_6, L_7, L_8, L_9, L_{10})</td>
<td>-(\infty)</td>
<td>(\Delta L_1, \Delta L_2)</td>
<td>-20.0</td>
</tr>
<tr>
<td>S4</td>
<td>(L_1, L_2, L_3, L_4, L_5, L_6)</td>
<td>-10.2</td>
<td>(L_7, L_8, L_9, L_{10}, L_{11})</td>
<td>-(\infty)</td>
<td>(\Delta L_1, \Delta L_2)</td>
<td>-20.0</td>
</tr>
</tbody>
</table>

The feature partition schemes (S1, S2, S3 and S4) and the corresponding threshold assignment are showed in Table 1 (for simplicity, all threshold values are expressed in log scale). The feature sub-vector \(X_2\) is robust to impulsive noise, so we do not floor its observation probability by setting threshold to \(-\infty\). The dynamic LPCC is robust, so we make the 13 ΔLPCC as the third sub-vector \(X_3\) and a very loose empirical threshold \(T_3 = -33.0\) is assigned just to slightly floor its observation probabilities.

5.2 Evaluation Results and Analysis

Figure 4 shows the recognition accuracy of the baseline and the FOP method under different feature partition schemes in substitutive impulsive noise environment. It can be seen that in baseline the LPCC feature can obtain very high recognition accuracy in clean environment (OR=0%), but its performance drops rapidly as the OR increases from 5% to 50%. The proposed FOP method can efficiently improve the robustness of
LPCC feature. The error reduction rate (ERR) of $S_1$, $S_2$, $S_3$, and $S_4$ versus the baseline averaged over all ORs are 38.9%, 34.8%, 35.5% and 29.8%, respectively.

Figure 5 compares the performance of baseline and the best FOP scheme, i.e. $S_3$ in additive machinegun noise. It can be seen that performance of baseline LPCC degrades dramatically in the presence of impulsive noise, especially at low SINRs and
high ORs. It is very obviously that the S4 FOP scheme is much more robust in additive machinegun noise. For example, at -10dB, -5dB and 0dB, the ERR of S4 versus the baseline averaged over all ORs are 41.6%, 46.7% and 52.1%, respectively.

6 Conclusions

In this paper we apply the FOP method to LPCC feature to improve its robustness against impulsive noise. The feature vector is partitioned into three sub-vectors according to the noise sensitivity analysis and a flooring threshold is calculated for the sensitive sub-vector. Experimental results confirm the efficiency of FOP method on LPCC features. Compared with MFCC feature, LPCC feature is more sensitive to impulsive noise. In future work it is promising to use feature recovery methods to further improve the robustness of LPCC feature.

References