Distorted Speech Rejection for Automatic Speech Recognition in Wireless Communication

Joon-Hyuk CHANG\textsuperscript{1)} and Nam Soo KIM\textsuperscript{1)}, Nonmembers

SUMMARY This letter introduces a pre-rejection technique for wireless channel distorted speech with application to automatic speech recognition (ASR). Based on analysis of distorted speech signals over a wireless communication channel, we propose a method to reject the channel distorted speech with a small computational load. From a number of simulation results, we can discover that the pre-rejection algorithm enhances the robustness of speech recognition operation.

Key words: pre-rejection, ASR, wireless communication

1. Introduction

Progress in automatic speech recognition (ASR) technology has enabled the development of more practical and more successful speech recognition applications. Actually, the processing and computation subparts of an ASR application can be located not in the individual terminal but in the central ASR server. Having a recognizer residing in a central server enables large-scale and more powerful computing machines to perform recognition, facilitating more sophisticated and elaborate ASR applications. However, this provides a chance for the speech signal to be affected by the wireless channel distortions.

In previous studies, much work has been focused on solving the problems of acoustic robustness in mobile and cellular environments since the speech recognition performance degrades drastically when the speech signal is distorted by the wireless network and noisy environments. One possible method to enhance the robustness of an ASR system is to employ the conventional rejection algorithms \cite{1}, \cite{2}. However, since these rejection algorithms are usually done after feature extraction and recognition phases, much computation is required. Moreover, little attention has been paid to the transmission channel characteristics and as a result the rejection capability is somewhat restricted. For these reasons, we propose a pre-rejection technique aiming at defeating the distorted speech signal in noisy wireless environments with a low computational load. A large amount of speech data is collected in the wireless communication environment and analyzed. From the analysis, we can discover that the speech signal is seriously distorted and the distortion is mainly classified into three different types when the communication channel is in several bad conditions which will be illustrated in Sect. 2.

2. Speech Signal in Wireless Communication

In real wireless environments, the speech signal is seriously distorted because of not only the source environment \cite{3} but also the various channel conditions \cite{4}, \cite{5}. Generally, a mobile terminal is connected to the radio base station through the wireless channel. At this point, it is possible for the errors to be introduced in the bit-stream representing the signal due to the physical phenomena of channel or interference noise during transmission. For that reason, it is considered meaningful to describe the propagation from the transmitter to receiver in terms of individual ray path \cite{5}. In an urban environment a variety of ray paths exist and they involve the phenomena of reflection, refraction, and scattering from the building, vehicle, people and so on. These paths make it possible to receive signals even when the transmitter and receiver are not placed within the line of sight (LOS) with each other.

Based on the former surveys, we know that there are two principal reasons which give rise to the distortion in the observed signal \cite{5}. First, a shadow fading often happens in the wireless channel when the subscriber is in moving environments. Specifically, when moving farther down the street, the subscriber may pass by buildings of different height, vacant lots, street intersection and so on. As a result of shadowing by buildings and other objects, the envelope of the transmitted signal within individual small areas varies from one place to another in an apparently random manner. Second, the envelope or magnitude of the received signal voltage is averaged over a distance on the order of 10 m, and the result is referred to as the small-area average. The small-area average varies over a shorter time scale than the shadow fading. This phenomenon is referred to as fast fading. In addition to the fading effects of the wireless channel, the signal can be distorted while being passed through a cascade of multiple communication systems.

In order to investigate the effects of wireless channel to the transmitted speech signal, we carried out an experiment by the use of an interactive voice response (IVR) system. A large amount of speech data, which had been transmitted through the code division multiple access (CDMA) channel, was recorded in the IVR system. Speech data was produced in a variety of source environments including indoor, outdoor, street, stopped car, moving car and so on. We performed speech recognition on the collected speech database.
After analyzing the speech data which made some recognition errors, we could find that the representative types of channel distortions are summarized as follows:

**TYPE 1**: unrecognizable signal attenuation

**TYPE 2**: waveform clipping

**TYPE 3**: waveform truncation or omission

Typical examples are shown in Fig. 1. It can be concluded that the bad environments in wireless communication are likely to make the speech signal distorted and then degrade the performance of the ASR system. When the above distorted speech signals were applied to the ASR system, they were mostly rejected by the rejection module in the ASR or falsely recognized. However, since the conventional rejection algorithm is performed after the feature analysis, probability computation and search procedures, heavy computation is consumed. To reduce the computational amount, we propose a novel pre-rejection algorithm which is easy to implement and can be performed with very little computation.

### 3. Pre-Rejection Algorithm for Speech Recognition

It is well-known that the speech signals have periodic structures due to the periodic excitation of the vocal cord. For that reason, the periodicity of a signal is considered an efficient measure to discriminate speech. Particularly, we use the normalized pitch correlation which is a useful measure to determine the pitch period and the degree of voicing inherent in the input speech signal [6]. We are motivated by the fact that the periodic characteristic is easily broken when the speech signal is seriously distorted during the wireless communication. This measure can be useful in distinguishing non-distorted speech signal from the distorted speech signal.

![Fig. 1](image1.png)  
*Fig. 1* A typical example of channel distorted speech signal in wireless communication. (a) TYPE 1 (b) TYPE 2 (c) TYPE 3.

A 20 ms frame of speech signal $s(n)$ sampled at the rate of 8000 Hz (160 samples) is processed. A normalized pitch correlation, obtained from the linear prediction (LP) analysis on each subframe (10 ms), is computed [7] such that

$$R_m(\tau) = \frac{\sum_{n=0}^{N-1} s(n) \cdot s(n + \tau)}{\sqrt{\sum_{n=0}^{N-1} s^2(n + \tau)}}$$  \hspace{1cm} (1)$$

where $m$ is the subframe index, $N$ is the window size and $\tau$ is the time shift, respectively. Using (1), the open loop pitch lag is determined by

$$\hat{p} = \arg \max_{\tau} R_m(\tau).$$  \hspace{1cm} (2)$$

In Fig. 2, the pitch lag contours of the channel-distorted speech signals given in Fig. 1 are displayed. The normalized pitch correlation is defined such that

$$R_p(m) = R_m(\hat{p})$$  \hspace{1cm} (3)$$

where $\hat{p}$ is the open loop pitch lag determined by (2). The normalized pitch correlations computed from the above three distorted signals are illustrated in Fig. 3. For surveying the pitch parameter continuity on frames, we use buffers 5 previous normalized pitch correlation $R_p$ and pitch lag $P_p$ in which $R_p = [R_p(m-4), R_p(m-3), \cdots, R_p(m)]$ and $P_p = [p(m-4), p(m-3), \cdots, p(m)]$ where $R_p(m)$ and $p(m)$ are the normalized pitch correlation and pitch, respectively, computed in subframe $m$.

Our approach to pre-rejection is based on the pitch continuity of the input signal. First a running average of the normalized pitch correlation is computed such that

$$R_p(m) = \lambda_R R_p(m-1) + (1 - \lambda_R) \mu R_p$$  \hspace{1cm} (4)$$

where $\mu R_p$ is the averaged normalized pitch correlation of $R_p$ and $\lambda_R$ (= 0.8) is the long-term smoothing parameter. A decision rule is made on each subframe such that

\[ \text{if } R_p(m) > \tau_{\mu 1} \text{ then } \text{FLAG} = 1 \]
else if \[ \bar{R}_P(m) > \tau_{\mu 2} \text{ and } \sigma_P < \tau_{\sigma} \]
then \( \text{FLAG} = 1 \)
else \( \text{FLAG} = 0 \)

where \( \sigma_P \) is the standard deviation of the pitch lags computed over \( P \) and \( \tau_{\mu 1}(=0.63), \tau_{\mu 2}(=0.45), \tau_{\sigma}(=1.30) \) denote the specified thresholds which are experimentally optimized values. The purpose of these parameters is to consider the quasi-stationary periodic nature of the non-distorted speech.

Based on the determined value of \( \text{FLAG} \), we set the counter for pitch continuity as follows:

\[
\begin{align*}
\text{if} & \left( \text{FLAG} = 1 \text{ and } \text{VAD} = 1 \right) \\
\text{then} & \quad C_{pc} = C_{pc} + 1 \\
\text{else} & \quad C_{pc} = C_{pc}
\end{align*}
\]

in which \( \text{VAD} \) denotes the binary decision for the frame status as to whether speech signal exists (=1) or not (=0), which is received by the end-point procedure of the ASR. Specifically, we adopt the VAD routine of the selectable mode vocoder (SMV) which only gives us the information to distinguish the active speech frame and silence frames [6]. Given two hypotheses, \( H_0 \) and \( H_1 \), which respectively indicate distorted (TYPE 1, 2, 3) and non-distorted speech signal, the final decision rule by the pre-rejector is described as

\[
C_{pc} \begin{cases} H_1 & \text{if } C_{pc} < \eta \\ H_0 & \text{otherwise} \end{cases} \tag{5}
\]

in which \( \eta \) is the threshold of decision. In actual situation, \( \eta \) can be chosen carefully by considering the tradeoff between detection and false-alarm accuracies. It is important to reset \( C_{pc} \) in (5) whenever a sufficiently long pause period is detected by the VAD.

4. Simulation

In this section, the proposed pre-rejection algorithm is evaluated over a speech database. Since this work deals with pre-rejection of the distorted speech signal, we made reference decisions for the recorded speech signals which had been transmitted through the CDMA communication channels. The CDMA database was collected in actual field conditions including indoor, outdoor, car and so on using the IVR system. People were asked to spontaneously answer the listed Korean persons' names which consisted of 2-3 syllables.

A total of 1100 names were used for the recognition list and 2345 files were recorded. Each file was applied to the ASR system which was based on the hidden Markov models (HMMs) [8] and then classified into two categories automatically. One was the set of files successfully recognized by the ASR and the other was the falsely recognized class. The falsely recognized database was further divided into the non-distorted and distorted speech signals manually. Finally, 2126 files were classified to the non-distorted speech class (truly or falsely recognized by the ASR) and 219 files were decided to belong to the distorted speech class (falsely recognized by the ASR). After the ASR procedure and the manual classification, all of the 2345 files were applied to the proposed pre-rejection algorithm to determine whether the speech file is channel distorted or not.

Let us represent the probability of detection as \( P_d \) which indicates the ratio of detecting the distorted signal while \( P_f \) denotes the false alarm probability. The receiver operating characteristic (ROC), which shows the tradeoff between \( P_d \) and \( P_f \) of the pre-rejection algorithm is given in Fig. 4. As shown in Fig. 4, the pre-rejection algorithm produced \( P_d > 60\% \) for the distorted speech signal while \( P_f \) is fixed below 5%.

5. Conclusion

We have proposed a pre-rejection technique for wireless channel distorted speech as a pre-processing stage of speech
recognition. Our approach is based on the quasi-stationary periodic nature of the non-distorted speech, and it can be implemented without a heavy computational burden.

Acknowledgement

This work was supported by Brain Korea 21 Project.

References