Pre-rejection of distorted speech for speech recognition in wireless communication channel

Joon-Hyuk Chang, Dong Jin Seo, Young-Joon Kim and Nam Soo Kim
School of Electrical Engineering and INMC, Seoul National University
Kwanak P.O.Box 34, Seoul 151-742, Korea,
Tel: +82-2-880-1824 Fax: +82-2-878-1452
E-mail: {changjh@, djseo@hi., kjun@hi., nkim@}{snu.ac.kr

Abstract— This paper introduces a pre-rejection technique for wireless channel distorted speech with application to automatic speech recognition (ASR). We investigate the characteristics of distorted speech over wireless communication channel with a survey on the physical phenomena of the propagation channel. Based on the analysis of the channel distorted speech signal, 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.

I. INTRODUCTION

Nowadays, the progress in automatic speech recognition (ASR) technology has enabled the development of more practical and more successful speech recognition applications. This progress makes it feasible for the ASR to become a common feature and services for the current and future portable terminal devices deployed in mobile or wireless networks. Actually, the processing and computation routines 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 computers 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 efforts made by other researchers, a significant work has been devoted to 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 algorithm. Various studies have been conducted aiming to improve the robustness of a speech recognition system based on the rejection technique [1], [2]. However, since the algorithms are usually done after feature extraction and recognition phases, much computation is required. Moreover, little attention is paid to the transmission channel characteristics and as a result the rejection capability is restricted. For these reasons, we propose a pre-rejection technique aiming to defeat 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 a bad condition such as fading.

In this paper, we present a novel pre-rejection technique for rejecting the distorted speech signal communicated through the wireless channel. From the ASR experiments, we can see that the proposed method yields a good performance of rejection with a low complexity.

II. SPEECH SIGNAL IN WIRELESS COMMUNICATION

In real wireless environments, the speech signal is seriously distorted because of not only the source environment [3] but also the various channel conditions [4]. Here, we review the characteristics of the propagation channel to learn the actual environments about wireless communication. Generally, the mobile terminal is connected to the radio base station through the wireless channel. At this point, it is possible that errors have been introduced in the bit-stream representing the signal due to physical phenomena of the channel or interference noise in the transmission. For that reason, it is considered meaningful to describe the propagation from the transmitter to receiver in terms of individual ray path [5]. In an urban environment a variety of ray paths exist and they involve the phenomena of reflection, diffraction, 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. 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
an 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. Fig. 1 represents the frequency responses of the fast and shadow fading channels in wireless communication, respectively. 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. 2. It can be concluded that the bad environments in wireless communication such as the fading are likely to make the speech signal distorted and then degrade the performance of the ASR system. Actually, when the above distorted speech signals are applied into the ASR system, they are 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, high computational complexity is inevitable. To reduce the computational amount, we propose a novel pre-rejection algorithm which is compact and can be performed with very little computation.

III. 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 periodicity of 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.

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 the subframe (10 ms) of the coding frame, 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

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

In Fig. 3, the pitch lag contours of the channel-distorted speech signals given in Fig. 2 are displayed. The normalized pitch correlation is defined such that

$$R_p(m) = R_m(\hat{\tau})$$  \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. 4.

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

\[
\bar{R}_p(m) = \lambda_R \bar{R}_p(m-1) + (1 - \lambda_R) \mu_{R,p}
\]

where \( \mu_{R,p} \) is the averaged normalized pitch correlation of the five latest subframes and \( \lambda_R(=0.8) \) is the long-term smoothing parameter. A decision rule is made on each subframe such that

\[
\begin{align*}
\text{if } \bar{R}_p(m) > \tau_{\mu_1} & \quad \text{FLAG} = 1 \\
\text{else if } \bar{R}_p(m) > \tau_{\mu_2} & \quad \text{and } \sigma_p < \tau_{\sigma} \quad \text{FLAG} = 1 \\
\text{else} & \quad \text{FLAG} = 0
\end{align*}
\]

where \( \sigma_p \) is the standard deviation of the pitch lag computed over the five latest subframes and \( \tau_{\mu_1}(=0.63), \tau_{\mu_2}(=0.45), \tau_{\sigma}(=1.30) \) denote the specified thresholds. Based on the value of \( \text{FLAG} \), we set the counter for pitch continuity as follows:

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

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. Given two hypotheses, \( H_0 \) and \( H_1 \), which respectively indicate distorted and non-distorted speech signal, the final decision rule is described as

\[
C_{pc} \begin{cases} H_1 & \text{if } C_{pc} H_1 \geq \eta \\ H_0 & \text{else}
\end{cases}
\]

in which \( \eta \) is the threshold of decision.

IV. EXPERIMENTAL RESULTS

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 persons’ names. A total of 1100 names were used for the recognition list and 2345 files were recorded. Each file was applied to the ASR 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).

The speech recognition system developed in our group was based on the hidden Markov models (HMMs) [8]. The speech signal was sampled at 8 kHz and segmented into 30 ms frames. Each name was characterized by the triphone and the respective phone was modeled by a three-state left-to-right HMM without skips and two mixture components for each state. Silence models were placed on both ends of the name models to provide a flexible word boundary. The feature vector used in our experiments consisted of eight Mel-frequency cepstrum coefficient, log energy and their first-order
differences. After the described speech recognition process, the falsely recognized speech files were divided into the non-
distorted and distorted speech classes manually.

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 trade-off between \( P_d \)
and \( P_f \) of the pre-rejection algorithm is given in Fig. 5. As
shown in Fig. 5, the pre-rejection algorithm produced \( P_d > \)
60% for the distorted speech signal when \( P_f \) is fixed at 5%.

![Receiver operating characteristic of the proposed pre-rejection algorithm.](image)

**V. CONCLUSIONS**

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.

**REFERENCES**


vocabulary utterances in unconstrained speech”, Proc. of IEEE Int. Conf.
its variability in VHF and UHF land-mobile radio service,” Review of
the Electrical Communication Laboratory, Vol. 16, No. 9-10, Sep.-Oct.
1968.
Hall, 2000.