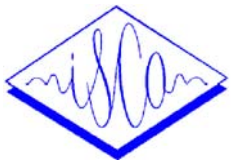

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A Distributed Speech Recognition System in Multi-user Environments

Kyu Jeong Han[†], *Naveen Srinivasamurthy*[‡] and *Shrikanth Narayanan*[†]

[†]Department of Electrical Engineering – Systems
University of Southern California
Los Angeles, California 90089
kyuhan@usc.edu and shri@sipi.usc.edu

[‡]Qualcomm Incorporated
San Diego, California 92121
naveens@qualcomm.com

Abstract

A typical distributed speech recognition (DSR) system is a configuration that distributes computational burden in signal processing and pattern recognition between a mobile unit and a remote recognition engine. For this system to be robust and practically acceptable, distortions caused by erroneous data transmission should be minimized. In this paper, the effects of multiple simultaneous users in a wireless network on speech recognition are considered. Specifically, multiple access interference (MAI) is shown to be a significant factor in the recognition performance degradation of a DSR system. From simulation results, both a minimum-mean-square-error (MMSE) detector and a de-correlating filter are shown to be effective in reducing MAI and improving recognition accuracy. In a CDMA system with 6 interferers at an SNR of 7 dB in an AWGN channel, there was a 30% absolute reduction in word-error-rate (WER) for a connected digit recognition task by using an MMSE detector to combat MAI. Almost same results were obtained by using a de-correlating filter.

1. Introduction

Provisioning automatic speech recognition (ASR) technology to mobile units, such as cellular phones or portable computers including personal digital assistants (PDAs), is one of the methods that help to meet increasing demands in a variety of wireless communication services. In other words, ASR on mobile units makes it possible to input various kinds of data - from phone numbers and names for storage to orders for business transactions - by speaking rather than by pressing keypads or by touching a screen with a pointing device. With this more flexible interface technology, more complex interactions with mobile units are possible even on the move, provided reliable recognition performance is guaranteed. However, performance reliability in complex ASR applications demands acoustic and language models that require greater computational resources and hence power resources not expendable at mobile units. Furthermore, in many cases, these models need to be continually updated to reflect dynamic application contents.

A distributed speech recognition (DSR) system provides a way to overcome the aforementioned challenges [1]. A typical DSR system has a client/server architecture, connected by wireless or wired networks, in order to distribute the computational burden of a speech recognizer between a mobile unit (client) and a remote recognition engine (server). In this system, speech utterances are acquired at a client and only features required for ASR are transmitted to a server. Hence, the client only performs feature extraction which is less complex than pattern recognition. On the other hand, the server (which is not hindered by complexity constraints) can aim for reliable recognition performance by using complex domain knowledge and increased computational resources.

To make a DSR system practically useful, the effects of distortions in source coding and transmission should be reduced. In this context, there have been a number of research efforts focusing on improving recognition performance in a DSR system. They have focused on reducing distortions from source coding (speech compression) [3] or channel errors [4]-[6]. In [3], it was emphasized that, since speech features were used for recognition rather than for playback, source coding system should be optimized for recognition. In [4] and [5], channel error protection and mitigation methods, superior to cyclic redundancy code (CRC) [1], were proposed. In [6], a modified Viterbi decoding for speech recognition based on reliability of the decoded data was investigated.

This paper addresses a new source for speech recognition degradation in a DSR system: *multiple access interference* (MAI). Multiple users sharing the same communication channel in mobile communication systems is known to result in MAI. In general communication systems, it is well known that MAI deteriorates bit-error-rate (BER) more severely as the number of users sharing a channel increase [7]. Given the significant degradation effect that MAI has had on general communication transmission, it is highly likely that MAI will also be a significantly undesirable factor causing degradation in ASR performance when multiple users share the same channel in a DSR system[1]. However, the effects of MAI on speech recognition have thus far not been investigated.

To cancel or suppress MAI, multi-user detection (MUD), which refers to the process of demodulating

[1] This significant degradation in recognition performance due to MAI is clearly demonstrated by recognition experiments in Section 4.

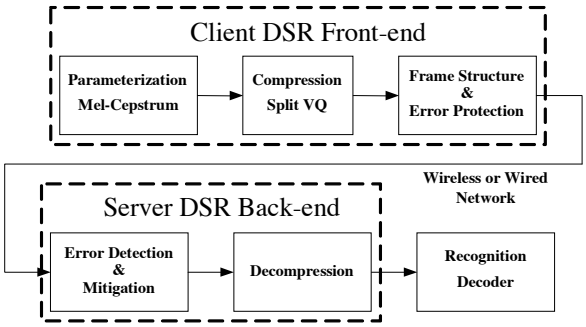

Fig. 1. *Structure of a DSR system.*

each user's data stream from a non-orthogonal multiplex, has been investigated in the field of communications [7]. Considering low complexity requirements needed in mobile devices, a minimum-mean-square-error (MMSE) detector and a de-correlating filter are the most attractive MUD schemes for real-time implementation in a DSR system. These have the advantages over optimal MUD schemes that they not only have less complex structures but also have only relatively small performance loss.

The objective of this paper is to show that (i) MAI is a potential important factor resulting in recognition performance degradation in a DSR system and (ii) low complexity MUD schemes, an MMSE detector and a de-correlating filter, are effective in significantly improving recognition accuracy in MAI environments.

The rest of this paper is organized as follows: in Section 2, the structure of a DSR system considered in this paper is described. In Section 3, we explain MAI in a DSR system and mathematical representations for an MMSE detector and a de-correlating filter. In Section 4, simulation results are presented. Finally, conclusions are given in Section 5.

2. System Description

The structure of a typical DSR system is depicted as a block diagram in Fig. 1. It consists of three blocks, a client DSR front-end, a server DSR back-end and a recognizer. (In a typical configuration, the first block is implemented on a mobile unit and the rest on a remote recognition engine.) While general recognition systems have only two blocks, i.e., parameterization and recognizer blocks, a DSR system has additional source and channel coding/decoding blocks to transmit and receive features through wireless or wired networks.

First, in the client front-end, 13 mel-frequency cepstral coefficients (MFCCs, C1~C12 and C0) and log energy (logE) are acquired as features for ASR from speech signal through standard front end signal processing. Then, the features are compressed by vector quantization (VQ). In this paper, for compression of the features, we used the method provided by the ETSI standard Aurora [1]. Aurora uses split VQ to reduce the computational complexity of VQ encoding. (In split VQ, generally, a vector is split into several sub-vectors and each sub-vector is quantized separately by a VQ [2].) Consequently, each feature frame is converted into a bit-stream having 44 bits, i.e., a sequence of 7 binary indexes corresponding to code vectors. At the transmitting end, each bit is multiplied by a spreading code given distinctly to each user for multiple accessing and transmitted by a carrier signal to the receiving end in the server back-end through a channel. (There are three possible multiple access schemes, frequency division multiple access (FDMA), time division multiple access (TDMA) and code division multiple access (CDMA). They allocate distinct frequencies, time slots and spreading codes to each user, respectively, in order to have multiple users share a system efficiently. In this paper, we focus on a CDMA system since it has been lately chosen as a multiple access standard for 3rd generation wireless communication systems and is expected to be the most popular multiple access scheme). At the receiving end in the server back-end, matched-filtering with a corresponding spreading code demodulates the bit stream of each user from the non-orthogonal multiplex. Then, the bit stream is de-compressed by picking up code vectors corresponding to the received indexes. As a result, features, possibly distorted, are obtained for ASR. In the recognizer, pattern recognition is performed with these decoded features.

Note that, without any loss of generality, no error protection schemes are considered in this paper. We wish to focus on the effects of MAI on speech recognition performance, which are not largely affected by whether or not error protection schemes are used.

3. MAI in a DSR System

Consider the transmitting and receiving ends in a DSR system as a $K-$user binary communication system, employing normalized spreading codes, $s_1, s_2, \cdots, s_K$, and signaling through an additive white Gaussian noise (AWGN) channel [7]. A received signal can be modeled as

$$r(t) = S(t) + \sigma n(t) \tag{1}$$

where $n(t)$ is AWGN with unit power spectral density, σ^2 is channel noise variance and $S(t)$ is the superposition of K users' spread data streams, given by

$$S(t) = \sum_{k=1}^{K} \sum_{i=-M}^{M} b_k(i) s_k(t - iT) \tag{2}$$

with parameters and other quantities being defined as follows:

$(2M+1)$ frame length;
$b_k(i)$ i^{th} symbol of the k^{th} user (assumed to be ± 1);
T inverse of data rate.

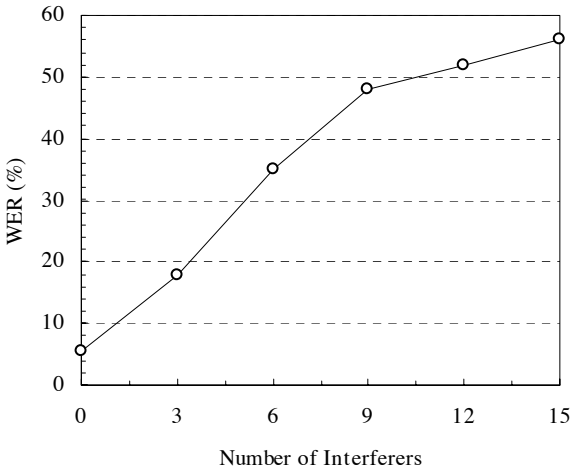

Fig. 2. *Performance of a DSR system for a connected digit recognition task according to the number of interferers (i.e., other users) in an error-free channel.*

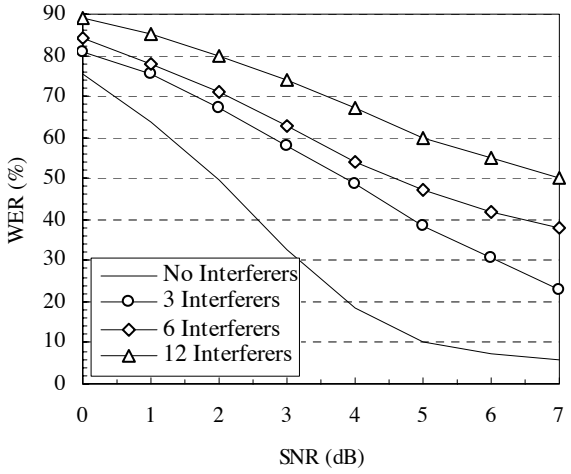

Fig. 3. *Performance of a DSR system for a connected digit recognition task in various multi-user environments and SNRs. (An AWGN channel is assumed.)*

It is assumed that s_k is supported only in an interval $[0, T]$. In this synchronous CDMA case, it is easily seen that a sufficient statistic for demodulating the i^{th} bits of the K users is given by a $K-$dimensional vector $\mathcal{Y}$ whose k^{th} component is

$$y_k(i) = \int_{iT}^{(i+1)T} s_k(t - iT) r(t) dt, \quad k = 1, 2, \cdots, K. \quad (3)$$

This sufficient statistics vector $\mathcal{Y}$ can be written as

$$\mathcal{Y} = \underline{\mathcal{R}} b + \sigma n, \quad (4)$$

where $\underline{\mathcal{R}}$ denotes the cross-correlation matrix of the spreading codes, $s_1, s_2, \cdots, s_K$,

$$\underline{\mathcal{R}}_{k,l} = \langle s_k, s_l \rangle, \quad (5)$$

b is a $K-$dimensional vector whose k^{th} component is b_k and n is a random vector, $\mathcal{N}(0, \underline{\mathcal{R}})$, independent of b. As seen in Eqs. $(3)-(5)$, MAI occurs as the sum of the multiplications of other users' transmitted symbols and cross-correlation values between different spreading codes unless $\underline{\mathcal{R}}$ is $\underline{\mathcal{I}}$, an identity matrix, i.e., the spreading codes are mutually orthogonal. Therefore, MAI increases in proportion to the number of users.

To cancel or suppress MAI, in this paper, an MMSE detector and a de-correlating filter are applied after matched-filtering at the receiving end. Both are linear transformations for $\mathcal{Y}$, the former of which chooses a linear filter to minimize average mean-square value between $\mathcal{Y}$ and a filter output and the latter chooses $\underline{\mathcal{R}}^{-1}$ as a linear filter. The MMSE detector demodulates data stream by multiplying $\underline{\mathcal{M}} = (\underline{\mathcal{R}} + \sigma^2 \underline{\mathcal{I}})^{-1}$ with $\mathcal{Y}$, i.e.,

$$\hat{b}_k = \text{sgn}\left(\left(\underline{\mathcal{M}} \mathcal{Y} \right)_k \right), \quad (6)$$

and so does the de-correlating filter by multiplying $\underline{\mathcal{R}}^{-1}$ with $\mathcal{Y}$, i.e.,

$$\hat{b}_k = \text{sgn}\left(\left(\underline{\mathcal{R}}^{-1} \mathcal{Y} \right)_k \right), \quad (7)$$

where $(\ )_k$ represents the k^{th} component of a column vector inside the parenthesis and $\text{sgn}(\cdot)$ represents a signum function.

4. Simulation Results

In this paper, recognition experiments are carried out using the TIDigits connected digit database. The database consists of 8440 utterances from 55 male and 55 female adult speakers for training. For testing, 4004 different utterances from 52 male and 52 female adult speakers are used. The HMM for each digit consists of 5 states with observations in each state modeled by 4 mixture Gaussian density functions. The baseline performance for the connected digit recognition task was 96.69% WER. To discern users, normalized Gold sequences with the length of 63 were used as spreading codes.

Firstly, Fig. 2 shows the recognition performance of a DSR system as a function of the number of interferers (i.e., other users) under the assumption of error-free channel. Note that the WER of no interference case is about 2% (absolute) greater when compared to the baseline performance because of distortion due to source coding. We can see that WER increases in proportion to the number of interferers. Specifically, observe that WER increases by more than 10% even with only 3 interferers. (It was 5.74% with no interferers.)

Next, Fig. 3 shows the recognition performance of a DSR system with 0, 3, 6 and 12 interferers for various SNRs in an AWGN channel. The performance of no interferers can be regarded as a baseline to evaluate the other cases. We can see more precisely that WER increases in proportion to the number of interferers. Also,

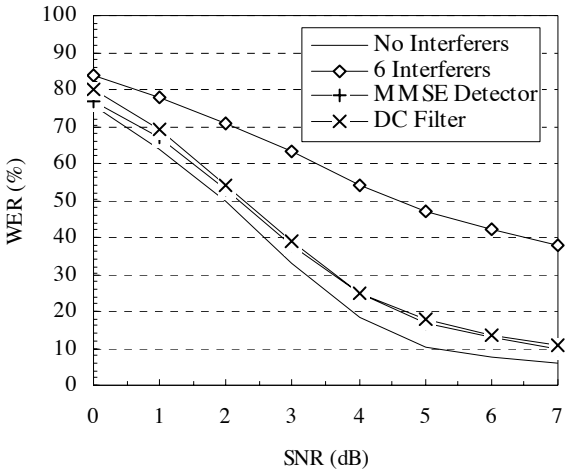

Fig. 4. *Performance of a DSR system for a connected digit recognition task with no interferers, 6 interferers and 6 suppressed interferers by an MMSE detector and a de-correlating filter. (In the legend, 'DC filter' means a de-correlating filter.)*

we can see that MAI is clearly a dominant factor in degrading recognition performance especially at high SNRs while channel errors dominate at low SNRs. Note that difference between the WERs of 0 and 3 interferer cases at an SNR of 7 dB is around 20% while it is less than 15% in Fig. 2. From this observation, we can conclude that MAI is more pronounced in an AWGN channel than in an error-free channel.

From the results of Fig. 2 and 3, we can conclude that MAI is a potentially significant factor that can cause significant degradation in the recognition performance of a DSR system and it is especially pronounced at high SNRs.

In Fig. 4, the recognition performance of a DSR system with 0, 6 interferers, and 6 suppressed interferers using an MMSE detector and a de-correlating filter for various SNRs are presented. We can see that the WER with interferers is reduced by the use of MUD schemes. Especially, for the case of 6 interferers, the WER at an SNR of 7 dB is reduced by about 30% (absolute) by the use of the MMSE detector and almost same results are obtained by the use of the de-correlating filter. From this observation, we can conclude that both the MMSE detector and the de-correlating filter are effective in multi-user environments for improving recognition accuracy in a DSR system as they were for reducing BER in general communication systems. Note that the recognition performance improvement of the de-correlating filter is almost same as that of the MMSE detector although the de-correlating filter has a less complex structure than the MMSE detector as shown in Section 3. This means that we can use the de-correlating filter as a MUD scheme in a DSR system rather than the MMSE detector, an optimal linear filter in the sense of minimizing average mean-squared error, for practical use in this environment.

5. Conclusions

We investigated the performance of a DSR system in multi-user environments and applied an MMSE detector and a de-correlating filter to tackle MAI. As mentioned in Section 1, the MMSE detector and the de-correlating filter were considered for MUD in a DSR system because linear multi-user detectors not only have less complex structures than optimal ones but also permit real-time implementation. From simulation results, MAI is shown to be a significant factor contributing towards recognition performance degradation in a DSR system. Additionally, the MMSE detector and the de-correlating filter are shown to efficiently suppress MAI and improve recognition performance in a DSR system.

Although, in this paper, only MAI in a DSR system under CDMA was considered, we can readily expect that MAI, which manifests as the collision of time slots allocated to different users, can also be overcome by using TDMA based recovery mechanisms. The comparison between MAI effects on a DSR system with CDMA and TDMA across various numbers of interferers and channel environments remains as a topic for future work.

6. References

[1] The European Telecommunications Standards Institute (ETSI) Standard, *Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms*. ETSI ES 201 108 V1.1.3, Sept. 2003.

[2] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp. 3-14, Jan. 1993.

[3] N. Srinivasamurthy, A. Ortega and S. Narayanan, "Efficient scalable encoding for distributed speech recognition," *IEEE Trans. Speech and Audio Processing*, Submitted 2004.

[4] V. Weerackody, W. Reichl and A. Potamianos, "An error-protected speech recognition system for wireless communications," *IEEE Trans. Wireless Comm.*, vol. 1, no. 2, pp. 282-291, April 2002.

[5] Z. Tan and P. Dalsgaard, "Channel error protection scheme for distributed speech recognition," in *Proc. ICSLP 2002*, Denver, CO, U.S.A., pp. 2225-2228, Sep. 2002.

[6] A. Bernard and A. Alwan, "Joint channel decoding – Viterbi recognition for wireless applications," in *Proc. EUROSPEECH 2001*, Aalborg, Denmark, pp. 2703-2706, Nov. 2001.

[7] S. Verdu, *Multiuser Detection*. Cambridge, UK: Cambridge University Press, 1998.