Recognition of Modulated Speech over OFDMA

Katrina Neville, Fawaz Al-Qahtani, Zahir M. Hussain, and Margaret Lech
School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia
s2008178@student.rmit.edu.au ; s3115478@student.rmit.edu.au ; zmhusain@ieee.org ; margaret.lech@rmit.edu.au

Abstract- This paper presents an analysis of the error associated with speech after it has been processed using QAM modulation and sent through an IFFT/FFT OFDM transmission system.

The main focus will be on how this type of modulation and transmission system will affect the unique features of speech used for speaker recognition and verification systems.

Channel deformations can lead to serious errors in speech feature, which would result in recognition and verification algorithms becoming unreliable for secure applications.

We will use the well-known Mel-Frequency Cepstral feature extraction algorithm to extract the speaker dependent features before and after the speech is modulated and sent over a noisy communication channel for comparison in our study.

I. INTRODUCTION

Speaker recognition and verification has attracted a lot of attention in recent times with the increased need for security of sensitive information. Substantial research has been done on the accuracy of automatic systems in recognising speakers from certain unique features within the individual’s voice as well as verifying who they say they are. It has been found in many cases that automatic systems can be more accurate in recognising someone from their voices than even humans can [1]. Even so, the accuracy of these systems can only be as good as the quality of the voices being used in testing and training the systems and with the increasing use of channels being used in the transmitting of voices, the encoding schemes and quality of the channel is a huge issue which cannot be ignored since problems can and do arise with the slightest distortion in the speech being used.

Recently, frequency diversity through orthogonal frequency division multiplexing (OFDM) has received massive attention as a candidate for optimal multimedia transmission. OFDM proved to be an effective frequency diversity technique that mitigates inter-symbol interference (ISI) in frequency selective fading channels and enables high data rate transmission over wireless channels. OFDM enhances the system performance in terms of lower error rate and higher channel capacity. In many real-time applications of speaker recognition or verification (e.g., wireless telephone banking), a high data rate with low ISI are essential, hence OFDM is a strong candidate for secure wireless recognition.

In most practical OFDM systems, IFFT/FFT is used for modulation. However, this technique requires a cyclic prefix (CP) to prevent inter-symbol interference (ISI). The cyclic prefix will reduce data transmission rate. The wavelet transform, although a bit more complicated in implementation, proved to be more promising in this connection as there is no need for CP.

In this paper we will study the performance of the mel-cepstrum speaker recognition algorithm (in terms of mean-square error) if speech is QAM/PSK modulated and sent using OFDM over a frequency-selective wireless channel.

II. SYSTEM CONFIGURATION

The system we used for experiments included a text-independent speaker recognition system which was set up according to the following block diagram in fig 1.

![Figure 1: Speaker recognition over OFDM.](image)

In the tests using the speaker recognition system the following procedures were followed:

Firstly, the system was trained using speakers from the German emotional database Emo-DB [5] (For these experiments neutral speech samples were used, therefore changes in emotions couldn't interfere with the results obtained).

To train the system, firstly the clean speech is preprocessed by pre-emphasising the speech using a first order high pass filter of the form

\[ y(z) = 1 - \frac{15}{16} z^{-1} \quad (1) \]

The silence segments are then removed and twenty mel-frequency cepstral coefficients are extracted and saved.

For system testing clean speech from the speakers are compressed using the DCT then the DCT coefficients are converted into a binary sequence before being modulated then passed through IFFT for transmission over a channel. The binary data is then converted back to a speech file and
processed the same as the training phase and the cepstral coefficients are extracted.

The recognition phase involves a comparison of the testing and training coefficients using the minimum distance classifier method to decide which speaker out of all the training set is most likely to be the test speaker.

In our experiment we also used the Mean Squared Error algorithm to determine the amount of error caused by the transmission system by comparing the transmitted speech to the same speech but not transmitted.

III. SPEECH FEATURE EXTRACTION: THE MFCC

The speaker-dependent features of human speech can be extracted and represented using the mel-frequency cepstral coefficients [4]. These coefficients are calculated by taking the cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale [3].

After pre-emphasis and speech/silence detection the speech segments are windowed using a Hamming window of 20 ms length with a 10 ms overlap [2], the Discrete Fourier Transform (DFS) is taken of these windowed segments. The magnitude of the Fourier Transform is then passed into a filter-bank comprising of twenty triangular filters. The start and end points of these filters are calculated by evenly spacing the triangular filters on the mel-scale and then using the following equation to convert these values back to the linear scale:

$$\text{Mel}(f) = 2595 \log_{10}(1 + \frac{f}{700})$$

The resulting filters used in our experiments are shown in Fig. 2.

The cepstral coefficients are calculated from the log-energy outputs of these filters by the equation [3]:

$$\text{MFCC}_i = \sum_{k=1}^{20} X_k \cos[i(k - \frac{1}{2}) \frac{\pi}{20}]$$

where $i$ is the number of the coefficients and $X_k$ is the log energy output of the $k^{th}$ filter.

In this paper 20 cepstral coefficients were extracted from each frame of speech and the average of these coefficients were used in recognising which speakers spoke them.

IV. MODULATION AND OFDM

After 4QAM modulation, an inverse discrete Fourier Transform (IDFT) is applied to obtain the signal in time domain. Following the IDFT a cyclic prefix is applied and the period of cyclic prefix must have a length greater than the channel delay spread to reduce the effect of Interblock Interference (IBI) caused by the dispersive Rayleigh fading environment [6]. We consider here the channels are frequency selective.

V. SIMULATION RESULTS

The first test we did was to see the amount of error between the clean speech files and the same speech files after they had been transmitted through the system. We completed five trials with all speakers; the results for three of the trials are as shown below in figures 3, 4 and 5.

The Modulation scheme 4QAM (QPSK) is used at a SNR of 30 dB. The FFT size was 8, and the channel impulse response was 6 taps generated as complex white Gaussian noise with unit power. Perfect channel estimation is assumed at the receiver. SNR was calculated per QAM symbol.
results from firstly using clean speech and then using the same utterances but after the speech has been transmitted, the results of these experiments are shown in Figures 6 and 7.

These figures show that by using 4QAM modulation and transmitting through an OFDM system over a frequency selective channel, the amount of error created in the speech, as little as it seems, is enough to cause enough error in the Mel-Frequency Cepstral Coefficients to cause a speaker recognition system to lose some accuracy.

The accuracy can be seen to decrease on most speakers, particularly with speakers 2, 7 and 9. Recognition was less accurate on most speakers, showing that the minor change in the speech files, analysed using Mean Squared Error between the clean and transmitted speech time waveforms (figures 3, 4 and 5), was enough to cause the Cepstral coefficients to differ.

Hence, for dependable speaker recognition over frequency-selective wireless, text-dependent approach could be and improvement.

VI. CONCLUSIONS

We investigated the possibility of text-independent speaker recognition using Mel-cepstral approach over a frequency selective channel with Gaussian noise. 4QAM modulation has been used with FFT/IFFT OFDM. Even a small amount of noise (SNR=30 dB) would lead to a considerable error. In this case text-dependent could be of use rather than text – independent recognition.

REFERENCES