Telematics and Informatics 31 (2014) 137–152

Contents lists available at SciVerse ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele

A low overhead scaled equalized harmonic-based voice authentication system

Sasan Adibi *

RMIT, School of Business IT and Logistics, 445 Swanston Street, Building 80, Level 8, Room 80, Melbourne, VIC 3000, Australia

ARTICLE INFO

Article history:
Received 1 May 2012
Accepted 7 February 2013
Available online 20 March 2013

Keywords:
Harmonic-based voice authentication
Amplitude equalizing
Frequency scaling
VoIP
UDP
Security
Performance
Delay
Jitter
Throughput
Labview

ABSTRACT

With the increasing trends of mobile interactions, voice authentication applications are in a higher demand, giving rise to new rounds of research activities. Authentication is an important security mechanism that requires the intended communication parties to present valid credentials to the communication network. In a stronger sense, all the involved parties are required to be authenticated to one another (mutual authentication).

In the voice authentication technique described in this paper, the voice characteristics of an intended individual wishing to take part in a communication channel will be classified and processed. This involves a low overhead voice authentication scheme, which features equalization and scaling of the voice frequency harmonics. The performance of this system is discussed in a Labview 8.5 visual development environment, following a complete security analysis.

1. Introduction

Voice authentication schemes are used in variety of scenarios, including; Party A-to-database, Party A-to-Party-B, and Party A-to-many (multicasting). The Party A-to-database scenario usually requires a simplex (one-way) or a half-duplex communication mechanism between Party A and the database, where Party A is either recording voice data and storing in the database and/or recalling voice data already saved in the database. The Party A-to-Party-B and Party A-to-many scenarios often require full-duplex (simultaneous bidirectional) communication capabilities. This paper is mostly biased towards the Part A-to-database scenario, which can easily be expanded to cover Party A-to-Party-B scenario as well. In both scenarios, each party is required to be authenticated based on the person's voice information and in all cases, mutual authentication is needed to reduce the chance of a Man-in-the-Middle (MitM) Attack. A MitM attack involves an illegitimate entity masquerading as a legitimate node, which may start a session or hijack an already progressed session and steal another legitimate entity's credentials. Once the credentials of a legitimate node are acquired, the illegitimate node can start a new session and pose as the legitimate entity and engage in illegal activities.

The contribution of this paper is in the design of a harmonic-based voice authentication system featuring a low overhead approach performing voice amplitude equalizing and frequency scaling. This system is being evaluated by LabView v8.5 followed by an analysis of the security system. The algorithms used in this paper are based on two patent applications published by the author (Sasan Adibi, PhD) (Adibi, 2011a,b), which will be discussed in this paper.

* Tel.: +61 3 9925 1542.
E-mail address: sasan.adibi@rmit.edu.au
URL: http://www.rmit.edu.au/browse?STYPE=PEOPLE&QRY=e03056

0736-5853/$ - see front matter
http://dx.doi.org/10.1016/j.tele.2013.02.004

The organization of this paper is as followed: Section 2 contains literature review of the current implementations on the areas of voice authentication techniques. Section 3 lays the foundation for the voice sampling procedures and processes. Section 4 aims at vocal waveform representations and classifications in the frequency domain. Section 5 is focused on the main contributions of this paper including; amplitude equalization and frequency-domain scaling, and the digital voice authentication-ID output format. Section 6 discusses the performance and security analyses, followed by conclusions and references.

2. Literature background

In this section, existing solutions on the voice authentication concept will be discussed. Before that, it is imperative to introduce a few fundamental common concepts used in most of the voice authentication techniques.

2.1. Speech cepstrum and homomorphic filtering

Homomorphic filtering is a generalized technique developed in the 1960s used in the signal processing techniques (mostly image and voice) (George et al., 1997). This filtering technique involves a nonlinear mapping between two different domains and the remapping of the original domain, similar to the Fourier conversion process, where the signal is converted to the frequency domain from the time domain and the inverse Fourier conversion that reconverts the frequency domain signal back to the time domain signal.

The cepstrum calculation is used in the homomorphic signal processing, which deals with a signal conversion combined with the signal convolution summing up to a linear separation. Power cepstrum is often used to represent human voices as vector representations. A cepstral calculation involves Fast Fourier Transform (FFT) calculations from a voice signal for voice recognition applications and transforming the voice data using the Mel (melody) scale. This sound processing system involving a Mel scale is called Mel Frequency Cepstrum (MFC), which is a short-term power spectrum representation of a sound spectrum based on linear cosine transform calculations. Mel Frequency Cepstral Coefficients (MFCCs) are collective coefficients of an MFC, which are normally calculated through the following steps (Mel-frequency cepstrum, 2012):

1. FFT application on the voice signal.
2. Mapping the power spectral density (PSD) onto the Mel scale using triangular overlapping windowing.
3. Calculating each Mel frequency PSD log values.
4. Performing Discrete Cosine Transform (DCT) on the results in step 3.
5. The resulted spectrum amplitudes are the MFCCs.

MFCC values are prone to room background noise, which requires modifications to withstand moderate amount of background noise levels.

In the cepstral coefficient calculation process, speech linear prediction coefficient calculation is involved. A MFCC system involved in such a speech linear prediction is often called; Linear Prediction-based Cepstral Coefficient (LPCC). Other types of cepstral coefficient calculation include; the Perceptual Linear Prediction (PLP) and Linear Filter Cepstral Coefficient (LFCC) calculations.

Both LPCC and MFCC are widely used for voice applications, which can form the LP-MFCC bundle (Linear Prediction-based, Mel Frequency Cepstral Coefficients) (Malegaonkar et al., 2008).

2.2. Text-independent speaker verification

Bimbot et al. (2004) consider a speaker verification system that involves a speaker training and test phases. The speech parameterization, called; the cepstral analysis, is used for speaker verification. To create the speaker model, the speech parameters are applied to the statistical modeling module and the results are normalized and evaluated (Bimbot et al., 2004). The speaker identification is based on two processes: Closed set and Open set. A closed set speaker identification is a references model of a nonexistent speaker and the nonexistent speaker does not match one of the existing models.

The speech parameterization, the modular representation of the training phase, and the speaker verification mechanism are shown in Figs. 1–3 respectively.

The speech verification is done using likelihood ratio detection (Maximum Likelihood concept). For the speaker recognition, Gaussian Mixture Models (GMM) and for additional temporal knowledge, Hidden Markov Model (HMM) are used.

2.3. Biometric authentication using voice

Nardini and Marzotti (2006) discuss a voice authentication biometric system. The voice capture and detection are the first steps in the process (segmentation), which are accompanied with zero crossing rate and peak pitch height used with an auto-correlation function. This way a voice signal is distinguished from other types of signals and periodic signals are distinguished from non-periodic signals. The next phase includes voice feature extraction mechanism involving two main techniques: MFCC and RelAtiveSpecTrAl; Perceptual Linear Prediction (RAS-TA_PLP) techniques. This involves voice signal sampling at 8 kHz,