A Robust Speech Biometric System for Vehicle Access

Sunil Kumar Kopparapu
TCS Innovation Labs - Mumbai
Tata Consultancy Services
Thane (West), Maharastra 400601
Email: SunilKumar.Kopparapu@TCS.Com

***Abstract*—Gaining key-less access to one's own vehicle using biometrics is a feature that is gaining popularity and has been implemented in concept and high level entry cars. But the state of the art in speech biometric especially as an embedded solution and the environmental conditions that the owner of the vehicle might be in (highway, noisy street, and garage) present challenges to the speech based verification system. We propose a multi authentication speaker verification system to enable the owner of the vehicle gain access to his vehicle. The idea is to have a small footprint embedded speaker verification system installed inside the car and a robust speaker verification system (more accurate) installed on a server. The output of both the channels is fused to verify the user.**

I. Introduction

On one hand vehicle theft is a common urban problem all over the globe while on the other hand people are expecting better and more comfortable driving experience including key-less access to their vehicles [1] [2]. In the United States alone, a motor vehicle is stolen every 28.8 seconds [3]. While it is believed that most of the theft of vehicles is possible because of slackness of the owner in terms of not locking the vehicle or parking it in a isolated area [4], it is also important to provide significant technological mechanisms that aid in not only making the vehicle theft safe but also provide the owner of the vehicle the comfort of having access to the vehicle.

Speech biometrics has been an active area of research for several years now but continues to hold interest [5] [6] because of the practical utility of being able to identify a speaker based on voice in several applications. Many voice based automated self help systems need the speaker to identify themselves before being serviced, for example banks and financial institutions.

In this paper, we propose a speech biometric based vehicle access system. In essence, the proposed mechanism allows key-less access to the vehicle using the owners voice to authenticate the owner and provide access to the vehicle. The contribution of the paper is in terms of proposing a robust access mechanism based on voice; the system uses a multi parameter, multi path, multi model architecture mechanism that provides robustness to the speaker identification mechanism. The rest of the paper is structured as follows, in Section II we describe in detail speech biometric followed by a brief description of speech signal processing (Section III), the proposed robust speaker identification platform for vehicles in Section IV and conclude in Section V.

II. Speech Biometrics

Speech Biometrics is a mechanism of verifying the identity of the speaker based on their voice characteristics. All speech biometric systems are learning systems, namely they have to be trained in order for them to be able to identify or recognize the speaker. Essentially there is a training phase which requires the speaker to register with the system by providing his voice samples and there is a test phase where the system identifies the speaker based on his previous reference voice samples. In a very broad sense speech biometrics can be classified as a speaker verification process or a speaker identification process. In the speaker verification mode, a person identifies himself as being $\mathcal{X}$ and the system verifies the identity through a binary decision, namely, you are $\mathcal{X}$ (YES) or you are not $\mathcal{X}$ (NO). Speaker verification tests the hypothesis that a certain individual $\mathcal{X}$ is the speaker of a given utterance (*Are you who you claim to be?*, namely, a test with two possible outcomes YES or NO). On the other hand in the speaker identification mode, the system determines if the speaker of a given utterance is among a set of $\mathcal{N}$ registered speakers or is an unregistered speaker (*Do I know you?*, $\mathcal{N}+1$ possible outcomes). This conceptual difference causes differences in accuracy, execution time, scalability and applicability between the verification and the identification process. It is not difficult to notice that for identification process both accuracy and execution time critically depend on the size of the set of registered speakers: the larger $\mathcal{N}$ the set size, the lower accuracy and longer the execution time; while for the verification process there is no significant affect.

Depending on the flexibility of what the speaker can say and the type of application (access to confidential data or access to public information) the speaker biometric system can be classified as (a) text dependent or (b) text independent. In a text dependent system the speaker is required to choose a fixed word or sentence at the time of training or registration; the same fixed word or sentence has to be spoken at the time of verification or identification. On the other hand a text independent system allows the speaker to speak a word or a phrase or a sentence of his choice when he wishes to be identified or verified. Clearly, a text dependent speaker verification or

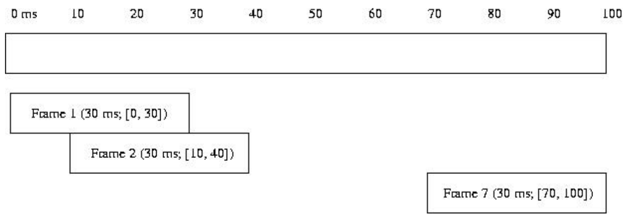

Fig. 1. If Speech signal duration is T (= 100 ms), then total total number of frames (P) is given by (T - F_{size})/F_{shift} = (100 - 30)/10 = 7, provided frame size (F_{size}): 30 ms and frame shift (F_{shift}): 10 ms. Typically, F_{shift} = F_{size}/3, F_{size}/2

identification system is more prone to be attacked than a text independent system but on the other hand a text independent system is hard to build than a text dependent system the former requires longer training and testing utterances to achieve good performance [7].

III. Speech Signal Processing

For any speech recognition or speaker identification process, the speech signal has to be processed and suitable features are to be extracted, the most commonly used features are (a) Mel frequency Cepstral coefficients (MFCC) or (b) linear predictive coefficients (LPC). To extract these features, in general, the speech signal has to be preprocessed.

We describe a process of extracting the popular MFCC speech parameters. In general, the Mel Frequency Cepstral Coefficients (MFCCs) are computed as follows [8]. Let $x[n]$ be a discrete speech signal which is divided into P frames (see Figure 1) each of length N samples with an overlap of $N/2$ samples such that

$$\{\vec{x}_1[n], \vec{x}_2[n] \cdots \vec{x}_p[n] \cdots \vec{x}_P[n]\}$$

where $\vec{x}_p[n]$ denotes the p^{th} frame of the speech signal $x[n]$ and is

$$\vec{x}_p[n] = \left\{ x \left[p \left(\frac{N}{2} - 1 \right) + i \right] \right\}_{i=0}^{N-1} \tag{1}$$

Now the speech signal $x[n]$ can be represented in matrix notation as

$$\hat{X} \overset{def}{=} [\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_p, \cdots, \vec{x}_P]$$

where

$$\vec{x}_p = \begin{bmatrix} x\left[(p-1) * \frac{N}{2}\right] \\ x\left[(p-1) * \frac{N}{2} + 1\right] \\ \vdots \\ x\left[(p-1) * \frac{N}{2} + N - 1\right] \end{bmatrix}$$

Note that the size of the matrix $\hat{X}$ is $N \times P$.

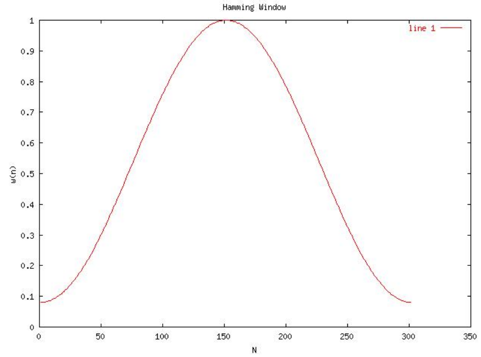

Fig. 2. Hamming Window

A. Windowing and DFT

In speech signal processing, in order to compute the MFCCs of the p^{th} frame, $\vec{x}_p$ is multiplied with a hamming window (see Figure 2)

$$w[n] = 0.54 - 0.46 \cos\left(\frac{n\pi}{N}\right)$$

followed by the discrete Fourier transform (DFT) as shown in (2).

$$X_p(k) = \sum_{n=0}^{N-1} x_p[n] w[n] \exp^{-j \frac{2\pi k n}{N}} \tag{2}$$

for $k = 0, 1, \cdots, N-1$. If f_s is the sampling rate of the speech signal $x[n]$ then k corresponds to the frequency $f(k) = k f_s / N$.

Let $\vec{X}_p = [X_p(0),\ X_p(1),\ \cdots,\ X_p(N-1)]^T$ represent the DFT of $\vec{x}_p$, then, $X = [\vec{X}_1,\ \vec{X}_2,\ \cdots\ \vec{X}_p,\ \cdots,\ \vec{X}_P]$ represents the DFT of the windowed p^{th} frame of the speech signal $x[n]$. Note that the size of X is $N \times P$ and is known as STFT (short time Fourier transform) matrix.

B. Mel Frequency Filter Bank

The modulus of Fourier transform is extracted and the magnitude spectrum is obtained as $|X|$ which is a matrix of size $N \times P$. The magnitude spectrum is warped according to the Mel scale in order to adapt the frequency resolution to the properties of the human ear [9]. It should be noted that the relation between the Mel frequency and the linear frequency is given by

$$m_f = 2595 log(1 + f/700)$$

where m_f is the Mel frequency and f is the linear frequency [10]. The inverse relationship between f and m_f is given by

$$f = 700(\exp^{m_f/2595} - 1)$$

Then the magnitude spectrum $|X|$ is segmented into a number of critical bands by means of a Mel filter bank which typically consists of a series of, say, F overlapping triangular filters defined by their center frequencies $f_c(m)$. The parameters that define a Mel filter bank are (a) number of Mel filters, (b) minimum frequency and (c) maximum frequency. For speech, in general, it is suggested in [11] that the minimum frequency be greater than 100 Hz. Furthermore, by setting the minimum frequency above 50/60Hz, we get rid of the hum resulting from the AC power, if present. [11] also suggests that the maximum frequency be less than the Nyquist frequency. Furthermore, there is not much information above 6800 Hz.

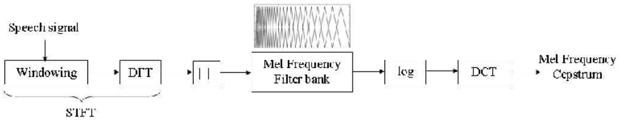

Fig. 3. Computation of Mel Frequency Cepstral Coefficients

The Mel filter bank, $M(m,k)$ [12] is given by $M(m,k)$

$$= \begin{cases} 0 & \text{for } f_k < f_c(m-1) \\ \frac{f_k - f_c(m-1)}{f_c(m) - f_c(m) - f_c(m-1)} & \text{for } f_c(m-1) \leq f(k) < f_c(m) \\ \frac{f_k - f_c(m+1)}{f_c(m) - f_c(m) - f_c(m+1)} & \text{for } f_c(m) \leq f(k) < f_c(m+1) \\ 0 & \text{for } f_k \geq f_c(m+1) \end{cases}$$

The Mel filter bank $M(m,k)$ is an $F \times N$ matrix.

C. Log Mel Spectrum

The logarithm of the filter bank outputs (Mel spectrum) is given in (3).

$$L(m,p) = ln \left\{ \sum_{k=0}^{N-1} M(m,k) |X(k,p)| \right\} \tag{3}$$

where $m = 1, 2, \cdots, F$ and $p = 1, 2, \cdots, P$. The filter bank output, which is the product of the Mel filter bank, M and the magnitude spectrum, $|X|$ is a $F \times P$ matrix.

D. Mel Frequency Cepstrum

A discrete cosine transform of $L(m,p)$ results in the MFCC vector.

$$D(r,p) = \sum_{m=1}^{F} L(m,p) \cos \left\{ \frac{r(2m-1)\pi}{2F} \right\} \tag{4}$$

where $r = 1, 2, \cdots, F$ and $D(r,p)$ is the r^{th} MFCC of the p^{th} frame. The MFCC of all the P frames of the speech signal are obtained as a matrix Φ

$$\Phi\{\hat{X}\} = D = [\Phi_1, \Phi_2, \cdots, \Phi_p, \cdots \Phi_P] \tag{5}$$

Note that the p^{th} column of the matrix Φ represents the MFCC of the speech signal, $x[n]$, corresponding to the p^{th} frame.

The outline of the computation of Mel frequency cepstral coefficients (speech parameters) is shown in Figure 3. These MFCC features are used to train the speech biometric system to verify or identify a speaker. Typically, hidden Markov models (HMM) are used to build statistical models for a speaker or dynamic time warping (DTW) to build deterministic models.

IV. Proposed Approach

Gaining key-less access to one's own vehicle[1] using biometrics is a feature that is gaining popularity and has been implemented in concept and high level entry cars. But the state of the art in speech biometric especially as an embedded solution and the environmental conditions that the owner of the vehicle might be in (highway, noisy street, and garage) present challenges to the speech based verification system. We propose a *multi authentication* speaker verification system to enable the owner of the vehicle gain access to his vehicle with small false rejection ratio and small false acceptance ratio.

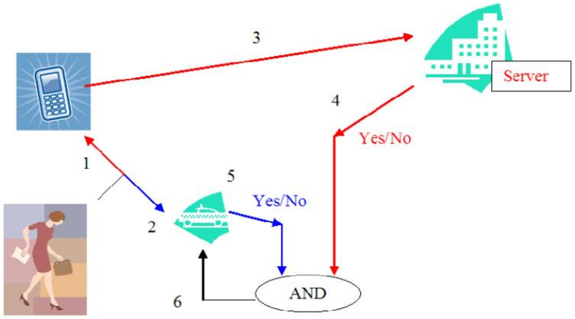

Fig. 4. High Level Architecture for Vehicle Access

The essential idea is to have a small footprint embedded speaker verification system installed inside the vehicle and a robust speaker verification system (more accurate) installed on a remote server. When wanting to gain access to the car, the car owner speaks simultaneously into his mobile phone and to the microphone mounted discretely on the vehicle. The same speech signal (spoken by the person wishing to gain access to his vehicle) is authenticated by the less powerful speaker verification system inside the car and simultaneously the speech via the mobile phone is delivered to the speaker verification server for verification. Only when both the speaker verification system on board the car and the system on the server verify the identity of the user is the user allowed access to the car. The high level architecture of the system is depicted in Figure 4.

The user who wishes to gain access to the car speaks the pass phrase simultaneously into the mobile phone (1 in Figure 4) and the car (2 in Figure 4). An embedded speaker verification system in the car verifies the identity and flags it as verified (YES) or not verified (NO) (5 in Figure 4); simultaneously via 1, 3 the speaker is verified at the server and a verified or not is marked (4 in Figure 4). Both the verifications are combined and a decision to given access to the car is invoked if both server and the embedded verification system return a Yes (6 in Figure 4). The server verification path is shown in red while the embedded verification path is shown in blue in Figure 4.

The speaker verification system on-board the car is based on using dynamic time warping (DTW) and cepstral LPC (CC-LPC) coefficients which allows for a small footprint verification system while the server based system is based on the more reliable Mel Frequency Cepstral Coefficients (MFCC) and uses hidden Markov models (HMM) to verify the user. This configuration of multi parameter (LPC, MFCC), multi path (on-board the vehicle, on server), multi model (HMM, DTW) architecture leads to a robust verification system as against using a speaker verification system on-board the car alone. It should be noted that in our approach,

[1] we will use vehicle and car interchangeably in this paper

- speaker verification rather than speaker identification makes sense assuming that the car is generally owned by an individual
- we extract two sets of parameters; one set is based on the modeling of speech production system as a linear finite impulse response (FIR) filter and popularly called the LPC parameters, while the second set is based on the modeling of the auditory perception of speech which leads to the MFCC parameters, this leads to better modeling of the speech. In our approach we use both these parameters extracted from the spoken speech which gives the overall system the required robustness to perform the verification more efficiently.
- Further in speech literature there are two different methods of comparing speech signals (a) one is a deterministic approach based on dynamic programming called DTW and (b) the other is a statistical approach where the speech is modeled as a hidden Markov model (HMM). In our proposed approach we use both these to compare and hence verify the validity of the speaker.

The use of multiple parameters (LPC, MFCC); multiple models (DTW, HMM; deterministic, statistical) give the speaker verification system robustness for verifying the identity of a person[2].

V. Conclusion

In this paper, we have proposed a speech biometric system for key-less access to vehicles. We have proposed a framework that allows the speaker authentication to be robust. In essence the authentication of speaker happens at two different places simultaneously, the speaker if verified both locally (in vehicle) and remotely (on a secure speaker authentication server) providing an extra mechanism of security. Additionally, the proposed framework is robust because of the use of two different types of speech feature (MFCC, LPC) from the speech signal and also use of different models (HMM based and DTW based) to verify the authenticity of the speaker.

References

[1] Michael Biermann, Tobias Hoppe, Jana Dittmann, and Claus Vielhauer, "Vehicle systems: comfort & security enhancement of face/speech fusion with compensational biometric modalities," in *Proceedings of the 10th ACM workshop on Multimedia and security*, New York, NY, USA, 2008, pp. 185–194, ACM.

[2] A Khare, A Sinha, B Bhowmick, KSC Kumar, H Gosh, SS Wattamar, and SK Kopparapu, "Multimodal interaction in modern automobiles," in *Multimodal interfaces for automotive applications*, Sanibel Island, Florida, 2009.

[3] WWW, ," http://www.rmiia.org/Auto/Auto_theft/Statistics.htm, Last Accessed July 2009.

[4] Karnataka Police, ," http://karnatakastatepolice.org/karnatakastatepolice/vehicle_theft.htm, Last Accessed July 2009.

[5] Jian-Da Wu and Bing-Fu Lin, "Speaker identification based on the frame linear predictive coding spectrum technique," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8056–8063, 2009.

[6] Homayoon Beigi, *Fundamentals of Speaker Recognition*, Springer, 2009.

[7] Microsoft Research, ," http://research.microsoft.com/en-us/um/people/zhang/speaker%20verification/default.htm, Last Accessed July 2009.

[8] Laxminarayana M and Sunil Kumar Kopparapu, "Effect of Noise-in-speech on MFCC Parameters," in *9th WSEAS International Conference on Signal, Speech and Image Processing (SSIP '09)*, Budapest, Hungary, September 3-5, 2009, 2009.

[9] Sirko Molau, Michael Pitz, Ralf Schl Uter, and Hermann Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," *Proc. Int. Conf. on Acoustic, Speech and Signal Processing*, pp. 73 – 76, 2001.

[10] Thomas F. Quatieri, "Discrete-time speech signal processing: Principles and practice," *Pearson Education*, vol. II, pp. 686, 713, 1989.

[11] CMU, "http:// cmusphinx.sourceforge.net/ sphinx4/ javadoc/ edu/ cmu/ sphinx/ frontend/ frequencywarp/ melfrequencyfilterbank.html," .

[12] Sigurdur Sigurdsson, Kaare Brandt Petersen, and Tue Lehn Schiler, "Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music," *Conference Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR)*, Vicoria, Canada, 2006.

[2] We plan to include experimental results in the final version of the paper