A New Approach in Creating Decision Systems Used for Speaker Authentication

Vlad Andrei Cârstea, Robert Alexandru Dobre, Claudia Cristina Oprea(✉), and Radu Ovidiu Preda

Telecommunications Department, Politehnica University of Bucharest, Iuliu Maniu Blvd. 1-3, 69121 Bucharest, Romania
rdobre@elcom.pub.ro, {cristina,radu}@comm.pub.ro

Abstract. This paper discusses a new approach of a Decision System to be used in Speaker Authentication applications, with particular emphasis on security systems with a small programming data set. This decision system is based on the adjoining of a matched filter response of the two compared signals considering the position of the maximum onto the abscissa of the response graph and afterwards, the use of Kullback – Leibler divergence for comparing the Mel Frequency Cepstral Coefficients' statistical distribution of the password and input speech signal.

Keywords: Speaker authentication · Matched filter
Kullback – Leibler divergence · Mel Frequency Cepstral Coefficients
Security system

1 Introduction

This article aims to provide a new type of decision system for use in the Speaker Authentication domain, itself a subfield of Automatic Speech Recognition. According to [1], Speaker Authentication is the process in which the claimant identifies himself by speech which is recorded as a signal and is tested against the reference model or password (also a speech signal previously created) to verify the claim.

Most of the systems in this field did not use or aimed to create particular methods of analysis of the speech signal but borrowed existing methods from the parent domain of Automatic Speech Recognition. The literature offers algorithms like Gaussian Mixture Models (GMM) [2], Vector Quantization (VQ) [3] or Searching for Digital Watermarking (DW) [4, 5] among mainstay. This paper proposes a new approach. A better understanding of human voice in its biological, phonetic and information carrying aspects yield better methods to be used as decision in an authentication system, namely a matched filter used in conjunction with Kullback – Leibler Divergence.

2 Voice Apparatus

It is known that voice is a sound wave emanating from vibratory parts of our body. We consider voice only that continuous scale of frequencies of the elastic wave that can produce excitation to our auditory apparatus. We consider regular speech that

O. Fratu et al. (Eds.): FABULOUS 2017, LNICST 241, pp. 167–172, 2018.
https://doi.org/10.1007/978-3-319-92213-3_24

frequency band in which mutually intelligible messages can be sent between humans. The voice apparatus consists of the air pressure system, the vibratory and the resonating system. It creates different parameters of sound (pitch, loudness, tempo) and amplifies or reduces voice harmonics.

It is necessary to study the way voice exits our mouth for information transmission. According to [6], when speaking in any language we produce phonemes which are defined and we distinguish between them by their place of articulation which is the position of the mouth organs (lips, teeth, tongue etc.) at the moment of utterance and the by the manner of articulation which arises from all organs' positions obstructing the airflow column.

From this paragraph, it can be extrapolated that voiced and voiceless phonemes exist. Voiced phonemes are those whose sound wave is quasi-periodic and voiceless are those aperiodic sound waves.

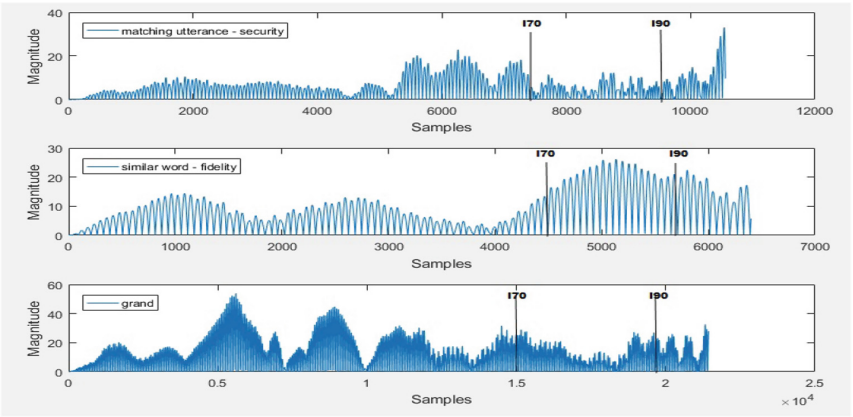

Fig. 1. The position of the maximum in the first case corresponds to a matching utterance of the word *security*. Note that the graph has maximum after the 90th interval. In the case of the second graph, we have a paronym to *security*, namely *fidelity*, so the maximum value is between 70th and 90th intervals. In the case of the different word, *grand*, the maximum is around the beginning of the response, before the 70th interval, thus the word is rejected.

3 Matched Filters

The matched filter implemented in this paper is the first decision test for a speech signal which is compared to the stored password. It acts as a trigger for the more refined statistical Kullback – Leibler divergence method. Of interest is the position of the maxima on the abscissa of the matched filter response for reasons which shall be further detailed.

A matched filter is typically a linear filter that cross-correlates a known signal with a received signal which has noise added to it and finds if in the received signal, the original one is present. Of course, the goal is not to find a signal in a noisy one but to

find if a signal is present in another and thus consider the second signal as noise which adds difference in information to the searched signal (any other phonemes/words than the searched one contains is considered added noise).

Based on [7], a filter matched to the signal $s(t)$ will have the impulse response:

$$h(t) = \mathrm{k} \cdot s(\mathrm{T} - t) \ , \tag{1}$$

where k and T are arbitrary constants.

As mentioned, it is not of interest if the two utterance signals are perfectly correlated between them because this would prove impossible with respect to the complexity of even a short vocal signal. The matched filter is primarily used to analyze the signal in time domain and to detect if the words in the two signals are the same, if they are paronyms or if they are totally different. If the filter response is analyzed, it shall be observed that the autocorrelation of the known signal with itself has on the response graph (here always restricted to the support of the known signal) the maximum value with its abscissa as the last value of the support. Now for two different utterances, it was observed that the value of the maximum is of no use, for example it can happen to be higher for different words than for similar utterances. For this, the system does not analyze quantitatively the matched filter response but qualitatively. We are not interested in the value of the maximum but in its position onto the abscissa and thus the ability of the graph to tend to the autocorrelation one.

For this decision system, it was decided to analyze the position of the maximum by dividing the abscissa of the response in 100 equally spaced intervals and to formulate an answer based on which interval the maximum is situated in. For example, if the maximum is in the first 70 intervals, the filtered signals are majorly different and the system rejects the speech signal. No further decision methods are taken. If the signal is between the intervals 70 to 90, there is a fair possibility that the signals match but decision is taken by a further test. If the maximum is situated in an interval greater than 90 it can be stated that the two utterances match and the divergence test has a great probability of confirming this match. Some situations are illustrated in Fig. 1.

4 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients are the constituents of the Mel Frequency Cepstrum which characterize the power spectrum of a sound wave, over a short time interval. They are linked directly with the sound properties of pitch, loudness and duration. Based on [8] it can be concluded that they are used to extract important linguistic features definitory for each voice in performing automatic speech recognition, in general, and speaker authentication, in particular. They are centered on the information carrying part of the voice disregarding noisy parts and other non-useful signals. Voice is formed and influenced by the vocal tract and the mouth. The shapes of all these organs are represented mathematically as the envelope of the short time power spectrum. Because of the distinctive perception of sound humans have, we need to use the Mel scale, a scale which links our perception of frequency with the real frequency scale in a non-linear manner.

5 Kullback – Leibler Divergence

The Kullback – Leibler Divergence is the second, more refined, test in this decision system. It analyzes the speech and the password subjectively because ultimately this system is used by humans, if we develop powerful objective decision methods we will not be able to obtain a positive system response because our utterances even if we hear them as similar are physically, profoundly different sound waves in terms of frequency, amplitude, duration, envelope of the wave. The Kullback – Leibler divergence is not applied to the signals but to the Mel Frequency Cepstral Coefficients extracted from the vocal waves. The password is formed from five utterances of the same word of which the median one, the third is considered as a reference signal.

5.1 Deriving the Kullback – Leibler Divergence

This divergence is the measurement of discrimination between two probability density functions (PDF). It measures how one PDF diverges from another expected PDF. Regarding information theory, an important measured parameter is entropy which give us an idea of how many units of information are necessary to encode the message. Starting from this, the divergence adds in the logarithmic comparison, another distribution and returning to us the expected logarithmic difference between the first and second distribution.

In terms of information theory, it tells us the number of information units we expect to lose from the first distribution if it is approximated by the second. For two gaussian distributions, P and Q, the Kullback – Leibler divergence is a scalar with:

$$KL(P,Q) = \ln\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{2}$$

After the Mel Frequency Cepstral Coefficients of each of the five password constituent signals are computed, and of the speech signal to be compared, if we look at the histogram of magnitude versus number of coefficients, they tend to respect a normal distribution. Thus, a Gaussian is fitted for each signal. For the password, the divergence of the four signals versus the reference signal is computed. Finally, the divergence of the speech signal versus the reference signal is computed and we look to see if it is in the range of the password divergences. If the matched filter did not reject the signal and so it arrived up to this point into the algorithm, we have three decision options. If the maximum was above the 90^{th} interval and the divergences correspond, the response is "positive". If the maximum was above the 90^{th} or 70^{th} interval but the divergences did not match the response is "reject". If the maximum was above the 70^{th} interval and the divergences correspond we repeat three times the process, then decide if "positive" or "reject".

6 System Performances

Let us consider some graphs of performance of this decision system which was implemented in a security system such that, from this security perspective we shall look upon them. For each of the graphs below, 20 consecutive utterances were performed and their results were rounded. In the first, the password and the utterances were of the same user, in the second there is a corroboration of the results taken from three different users trying to crack the password in different scenarios. It is worth to mention that these results were obtained from low security passwords, words like *open, password, sharply* and still they yielded good results for a five signal formed password and 20 utterances. It is especially good to see that the passwords with predominant voiceless phonemes (*sharply*) gave a 10% increase in the strength of the password and decision results as it can be seen in Fig. 2.

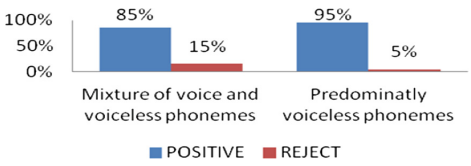

Fig. 2. The results obtained when the authorized user uttered his password.

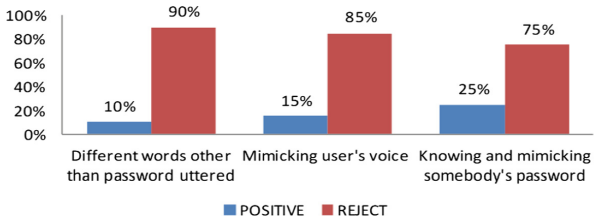

Fig. 3. Here three cases of password breaching by an impostor are presented. Of course the parameter relevant here is the percentage of rejection in each case, which is 75% even if the impostor knows the password yet he still needs to utter it – here lies the strength of the system, as it is not enough to know the password but it also need to be reproduced.

Table 1. Performance comparison between various systems. All results are rounded.

System	Performances [%]	Observations
Ours	85%–95%	Depends on phoneme composition
GMM	80%–87%	Depends on model order 8–16
VQ	57%–95%	Depends on codeblocks (i.e. speakers) 1–8
DW	51%–94%	Depends on test database

Regarding other systems' performances especially with respect to ours in Fig. 2, we analyzed the systems mentioned in sources [2, 3, 5] under tests as much as appropriate to ours namely small input database (5 utterances of 3 s each), small password length, around 3 s, Mel Frequency Cepstral Coefficients when present (Fig. 3 and Table 1).

7 Conclusions

The decision system is versatile, with the "reject" percentage highly influenced by the matched filter interval choice, which can easily be adapted to any language and calibrated after more testing, the choice in this paper was more to explain the concept and so they can be further and further narrowed for better results. Still by knowing both the password and how the user's voice sounds, the ratio of "positive"/"reject" was only 1/4 and this for a low security password, as mentioned. With further tests and more calibration, the performances of a security system with this decision system can be much improved. The authors recommend for this decision system, if used in a security system, as it was the original intent, to use as passwords, speech signals which are profoundly voiceless, for they are more difficult to be breached.

Acknowledgment. This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI - UEFISCDI, project number PN-III-P2-2.1-PED-2016-1465, within PNCDI III.

References

1. Beigi, H.: Speaker recognition. In: Yang, J. (ed.) Biometrics, p. 7. InTech (2011)
2. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. **3**, 72–83 (1995)
3. Hasan, R., Jamil, M., Rabbani, G., Rahman, S.: Speaker identification using Mel Frequency Cepstral Coefficients. In: 3rd International Conference on Electrical and Computer Engineering. ICECE, Dhaka (2004)
4. Faundez-Zanuy, M., Hagmuller, M., Kubin, G.: Speaker verification security improvement by means of speech watermarking. Speech Comun. **48**, 1608–1619 (2006)
5. Nematollahi, M.A., Akhaee, M.A., Al-Haddad, S.A.R., Gamboa-Rosales, H.: Semi-fragile digital speech watermarking for online speaker recognition. EURASIP J. Audio Speech Music Process. (2015). https://doi.org/10.1186/s13636-015-0074-5
6. Ladeforged, P., Johnson, K.: A Course in Phonetics, 6th edn. Cengage Learning. Inc., Boston (2010)
7. Turin, G.L.: An introduction to matched filters. IRE Trans. Inf. Theory **6**, 311–329 (1960)
8. Quatieri, Th.F.: Discrete Time Speech Signal Processing. Prentice-Hall, Upper Saddle River (2002)