Performance of a text-independent remote speaker recognition algorithm over communication channels with blind equalisation

Neville, Katrina; Jusak, Jusak; Hussain, Zahir; Lech, Margaret

https://researchrepository.rmit.edu.au/esploro/outputs/9921861943901341/filesAndLinks?institution=61RMIT_INST&index=null

---

Neville, K., Jusak, J., Hussain, Z., & Lech, M. (2005). Performance of a text-independent remote speaker recognition algorithm over communication channels with blind equalisation. Proceedings of the TENCON 2005 IEEE Region 10 Conference. <https://doi.org/10.1109/TENCON.2005.301283>

Published Version: <https://doi.org/10.1109/TENCON.2005.301283>

Repository homepage: <https://researchrepository.rmit.edu.au>

© 2007 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Downloaded On 2023/04/26 23:51:02 +1000

Performance of a Text-Independent Remote Speaker Recognition Algorithm over Communication Channels with Blind Equalisation

Katrina Neville, Jusak Jusak, *Student Member, IEEE*, Zahir M. Hussain, *Senior Member, IEEE* and Margaret Lech

School of Electrical and Computer Engineering, RMIT University, Melbourne, Victoria 3000, Australia.

E-mails: s2008178@student.rmit.edu.au; s3001898@student.rmit.edu.au; zmhussain@ieee.org; margaret.lech@rmit.edu.au

Abstract—In this work we will present a study of the reliability of a well-known speaker recognition algorithm when using speech sent over communication channels with Channel distortion and noise. The speech features used to test and train this system are the Mel-Frequency Cepstral Coefficients. For speaker recognition applications, channel deformations can lead to serious errors in recognition if the speech is transmitted, making the algorithm unreliable for usage in telephone banking or other applications requiring a high level of security. We will study the performance and reliability of this algorithm for text-independent speaker recognition with speech sent over a communication channel. We will be using blind equalisation techniques with QPSK modulation.

I. INTRODUCTION

Speaker recognition has attracted a lot of attention both in industry and academia in the last two decades particularly 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 and 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 problems can and do arise when speaker's need to be identified from their voices after being sent over a channel and compensation needs to come into effect so that the features extracted from a sample of speech are relevant and accurate.

Many feature extraction techniques are available for use in speaker recognition, they are classified into two groups, spectral based and non-spectral based features [3]. In this paper we have selected the spectral based Mel-cepstrum algorithm to be used in our system and we will study this system's performance when a decision needs to be made with voices sent through wired and wireless communication channels. The performance of this system will be determined by estimating the error, or distance between the reference cepstral coefficients and the received (equalised) coefficients.

Equalisation techniques have been extensively used in communication systems to remove intersymbol interference produced by dispersive channels [9], and has become increasingly important where full bandwidth utilisation of the channel is necessary. Conventional equalisation techniques rely on the

transmission of training signals which leads to a reduction in channel bandwidth and allocated resources. Thus, in the last few years, blind equalisation techniques have gained an increasing interest. The most popular and implemented blind adaptation algorithm is the constant modulus algorithm (CMA) proposed in [6] and developed independently by [7]. The main advantage of using a 'blind' system is apparent where the use of training signals is both unrealistic and costly to implement. The Constant Modulus Algorithm (CMA) has attracted the main research effort as a suitable blind wireless channel equaliser; due to its robustness over the violation of perfect blind equalisation (PBE) conditions [8].

II. SYSTEM MODEL

Our system was set up according to the following block diagram in fig 1.

Fig. 1. Speaker Recognition System Block Diagram.

Firstly the system is trained using 10 different speakers from the German emotional database Emo-DB [5] (For these experiments only neutral speech samples are used from this database, therefore change in emotions can'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, the silence segments are then removed and twenty Mel-Frequency Cepstral Coefficients are extracted and saved.

For the testing of the system clean speech from one of the speakers is converted into binary before being passed through a channel with impulse response $c = [0.04, -0.05, 0.07, -0.21, -0.5, 0.72, 0.36, 0, 0.21, 0.03, 0.07]$ and equalised using a Blind Equalisation algorithm, i.e., CMA. The binary data is then converted back to a speech file and processed as in the training phase and the Cepstral

Coefficients extracted and saved.

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 experiments we changed the Signal to Noise Ratio of the channel distorted speech to compare the sensitivity of the system with clean speech with no noise or distortion present, speech with a SNR of 20dB and SNR 30dB.

III. THE MFCC

The spectral based features of the Mel-Frequency Cepstral Coefficients have been proven to provide an accurate depiction of the spectral information of the human vocal tract.

The Mel-Cepstral features are calculated by taking the cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale [4].

After preemphasising the speech using a first order high pass filter and windowing the speech segments using a Hamming window of 20ms length with a 10ms overlap, the Discrete Fourier Transform is taken of these 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 were calculated firstly by evenly spacing the triangular filters on the Mel-Scale and then using equation 1 to convert these values back to the linear scale.

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

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

Fig. 2. Triangular Mel-Scale Filterbank.

The Cepstral Coefficients were lastly calculated from the log-energy outputs of these filters by the equation [4]:

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

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

For our experiments 20 Cepstral coefficients were extracted from each frame of speech and the average of these coefficients were used in our recognition phase.

IV. THE CMA

A model of baseband communication systems is considered, where subscript n denotes a discrete time index in baud-spaced quantities. Through this system, a sequence of symbol x_n with T symbol period is transmitted through a wireless channel c of a length N_c . On the other end of the system, a linear filter is applied to equalise the received signal which is already corrupted by inter-symbol interference (ISI) and white Gaussian noise. Thus, the system output of the model may be expressed as: $y_n = \mathbf{r}^H(n)\mathbf{f}(n) + \mathbf{w}^H(n)\mathbf{f}(n) = \mathbf{x}^H(n)\mathbf{C}\mathbf{f}(n) + \mathbf{w}^H(n)\mathbf{f}(n)$, where $x(n) = [x_n, x_{n-1}, \dots, x_{n-N_x+1}]^t$ denotes a finite-length source symbol vector, each symbol is assumed to be generated from a sub-Gaussian random variable with zero mean, independent and identically distributed (i.i.d) with variance $\sigma_x^2 \equiv E\{|x_n|^2\}$, $\mathbf{C}$ symbolise a $N_x \times N_f$ time-decimated channel convolution, $\mathbf{f}$ is representing a column vector of fractionally-spaced equaliser coefficients with a length N_f . A column vector $\mathbf{r}(n)$ and $\mathbf{w}(n)$ be the symbol of time-decimated of N_f received samples and white Gaussian noise, respectively. A Hermitian operator is denoted by $(\cdot)^H$, and matrix transposition is symbolised by $(\cdot)^t$.

Utilising stochastic gradient descent strategy, the equaliser coefficients are updated according to the following algorithm:

$$\mathbf{f}(n+1) = \mathbf{f}(n) + \mu \mathbf{r}(n) \psi_{\text{cma}}(y_n), \quad (3)$$

where μ is a small constant called the step-size, and $\psi_{\text{cma}}(y_n)$ is the CMA error function.

The error function of CMA is described as [6]:

$$\psi_{\text{cma}}(y_n) = y_n^* (\gamma - |y_n|^2), \quad (4)$$

where γ is a dispersion constant defined as

$$\gamma = E[|x(n)|^4] / E[|x(n)|^2]^2. \quad (5)$$

Under a perfect blind equalisation (PBE) condition [8], equalisers minimising the CM cost function can perfectly recover the original source symbols for some system delay δ ($0 \leq \delta \leq N_x - 1$) and phase shift θ ($\theta \in [0, 2\pi]$), such that $y_n = e^{j\theta} x_{n-\delta}$.

A. Simulation Results and Discussion

The reliability of our speaker recognition system was evaluated and the following results in Tables I and II were achieved.

TABLE I
PERCENTAGE OF SPEAKERS RECOGNISED FROM CHANNEL EQUALISED
SPEECH AT SIGNAL TO NOISE RATIO OF 20DB.

Speaker	1	2	3	4	5	6	7	8	9	10
correct	100	66	100	33	33	100	33	100	100	0

Without channel equalisation the recognition rate was 0% for all speakers; the channel distortion was such that no meaningful features could be successfully extracted. Tables I and II show the percentage of times that the speaker's voices

TABLE II
PERCENTAGE OF SPEAKERS RECOGNISED FROM CHANNEL EQUALISED
SPEECH AT SIGNAL TO NOISE RATIO OF 30dB.

Speaker	1	2	3	4	5	6	7	8	9	10
correct	100	100	100	100	66	66	33	100	66	33

were recognised during our experiments after equalisation occurred and for Signal to Noise ratios of 20dB and 30dB. It can be seen that, as expected, the recognition accuracy for almost all speakers improves as Signal to Noise Ratio increases. With a Signal to Noise ratio of 30dB all speakers are correctly recognised at least once during our experiments. Figures 3 and 4 show the amount of difference between the cepstral coefficients for two frames of speech for speaker one. The utterance used for this comparison was identical in each case and the figures show the Cepstral coefficients for clean speech and then channel compensated speech at Signal to Noise Ratios of 20dB and 30dB.

Fig. 3. Amount of error between the Cepstral Coefficients in the twenty-fifth frame of speech for speaker 1. solid line (Clean speech), dashed line (channel equalised with SNR of 20dB), dotted line (channel equalised with SNR 30dB)

Figs. 3 and 4 clearly shows that Communication channels have a great impact on the speech information being sent and therefore the features extracted from this speech. The error in the Cepstral coefficients for the clean speech and both the speech samples sent over the channel are obvious and substantial for both the frames analysed. Even though there is substantial differences between the Cepstral coefficients for the channel distorted and equalised speech and the clean un-corrupted speech, the CMA channel compensation seems to effectively reduce error between the two distorted speech samples, and towards the end of the samples the error is reduced to zero between the channel compensated speech. This is a good result as usually there would be a substantial difference in the speech spectral information and the resulting coefficients when signal to noise ratio is changed from 30dB to 20dB within the channel. Another point to note about the CMA algorithm from the

Fig. 4. Amount of error between the Cepstral Coefficients in the eighty-ninth frame of speech for speaker 1. solid line (Clean speech), dashed line (channel equalised with SNR of 20dB), dotted line (channel equalised with SNR 30dB)

results obtained is that CMA takes a number of frames before it can achieve optimum channel equalisation. From our results that number is approximately thirty five frames before the algorithm can lock onto the channel parameters and converge to accurately equalise the distorted data. Fig. 5 shows the convergence occurring.

Fig. 5. Differences between the fifth Cepstral coefficient for: solid line (channel equalised with SNR of 20dB), dashed line (channel equalised with SNR 30dB) over the entire speech segment

While fig. 5 shows only one coefficient, the fifth, over the entire speech segment the results are the same for the other 19 coefficients as well, they all seem to converge around the thirty fifth frame.

V. CONCLUSION

In this paper we examined the reliability of the MFCC algorithm when applied on speech transmitted through a communication channel with channel equalisation for remote speaker identification.

We found that by using the Blind Equalisation Algorithm CMA that the relative error between the extracted features from our speech is reduced between speech samples that have been sent over a communication channel with varying Signal

to Noise Ratios.

It was also found that the CMA algorithm takes some time before it's able to converge accurately to the channel parameters so as to effectively compensate for the channel distortion within the speech information. This could open up some more areas to investigate further to improve on the accuracy of our speaker recognition system, such as possibly introducing some form of compensation technique in the speaker recognition testing system itself to mitigate the effects on the accuracy of the system that the first thirty or so frames of speech have before convergence occurs.

More accurate results would also be achieved if a text-dependent rather than text-independent system was experimented on using these techniques since text-independent speaker recognition always increases the risk of error being achieved in the recognition phase.

REFERENCES

- [1] S. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Waveform for Automatic Speaker Identification and Verification," *Journal of the Acoustical Society of America*, vol. 55, No.6, pp. 1304-1312, 1974.
- [2] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology," *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '02)*, vol. 4, pp. 4072-4075, 2002.
- [3] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice*, Prentice Hall, New Jersey, 2002.
- [4] S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-28, No. 4 pp. 357-366, 1980.
- [5] *Emo-DB*, Retrieved (June 3rd, 2005), from <http://pascal.kgw.tu-berlin.de/emodb/index-1024.html>
- [6] D. Godard, "Self-recovering equalization and carrier tracking in two dimensional data communication systems," *IEEE Trans. Commun.*, vol. 28, pp. 1867-1875, 1980.
- [7] J. R. Treichler and B. G. Agee, "A new approach to multipath correction of constant modulus signals," *IEEE Trans. Acoust., Speech. Signal Processing*, vol. ASSP-31, no. 2, pp. 459-472, 1983.
- [8] C. R. Johnson, Jr., P. Schniter, T. J. Endres, J. D. Behm, D. R. Brown and R. A. Casas, "Blind equalization using the constant modulus criterion: A review," *Proceedings of IEEE*, vol. 86, no. 10, pp. 1927-1950, Oct. 1998.
- [9] B. Sklar, *Digital Communications*, 2nd ed, Prentice Hall, New Jersey, 2001.