Digital Investigation xxx (2018) 1–8

Contents lists available at ScienceDirect

Digital Investigation

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

DFRWS 2018 Europe — Proceedings of the Fifth Annual DFRWS Europe

Speaker verification from codec distorted speech for forensic investigation through serial combination of classifiers

M.S. Athulya*, P.S. Sathidevi

Electronics and Communication Engineering Department, National Institute of Technology, Calicut, India

ARTICLE INFO

Article history:
Received 3 November 2017
Received in revised form
13 March 2018
Accepted 28 March 2018
Available online xxx

Keywords:
Codec distortion
Speaker verification
GMM-UBM
SVM
PNCC
Equal error rate

ABSTRACT

Forensic investigation often uses biometric evidence as important aids for identifying the culprits. Speech is one of the easily available biometrics in today's hi-tech world. But, most of the speech biometric evidence acquired for investigative purposes will usually be highly distorted. Among these distortions, most prominent is the distortion introduced by the speech codec. Speech codec may either remove or distort some of the speaker-specific features, and this may reduce the speaker verification accuracy. The effect of distortion on commonly used speaker-specific features namely Mel Frequency Cepstral Coefficients (MFCC) and Power Normalized Cepstral Coefficients (PNCC), due to Code Excited Linear Prediction (CELP) codec (the most widely used speech codec in today's mobile telephony), is quantified in this paper. The features which are least affected by the codec are experimentally determined as PNCC. But, when these PNCC coefficients are directly employed, speaker verification error rate obtained is 20% with Gaussian Mixture Model-Universal Background Model (GMM-UBM) classifier. To improve the verification accuracy, PNCCs are slightly modified, and these modified PNCCs (MPNCC) are used as the feature set for the speaker verification. With these modified PNCCs, the error rate is reduced to 15%. By fusing these MPNCCs with MFCC, the error rate is further reduced to 8.75%. A series combination of GMM-UBM and Support Vector Machine (SVM) classifiers is also proposed here to enhance the speaker verification accuracy further. The speaker verification error rates for different baseline classifiers are compared with that of the proposed serially combined GMM-UBM and SVM classifiers. The classifier fusion with the fused feature set largely reduced the error rates to 2.5% which is very much less than that of baseline classifiers with normal PNCC features. Hence, this system is a good candidate for investigative purposes.

© 2018 Elsevier Ltd. All rights reserved.

Introduction

As research in speech coding has shown drastic improvement for the past few years, it becomes equally important to develop methods for speaker verification from codec distorted speech. Speech codecs are meant to reduce the bit rate of speech transmission, by compression using different methods, without compromising on the quality of received speech. Two most important applications of speech codecs are in mobile telephony and Voice over IP (VoIP). Since previous research has shown serious degradation in speaker verification using speech distorted by codec effects (Nandan and Saha (2012); Gallardo et al. (2014b)), we cannot neglect its effect when considering any mobile phone or VoIP related speaker verification system, especially in forensic cases. Various algorithms are used for speech compression, and hence various standards are available. Wide classifications include waveform coders and vocoders. Standards with as low as 2.4 Kbits/s are available. CELP coder is a widely employed vocoder based speech codec with different bit rates, and hence the effect of distortion due to CELP codec on speaker verification is considered in this work.

Literature has mainly focussed on the analysis of the effect of codec distortions, bandwidth, sampling rate changes, etc. on speaker recognition (Silovsky et al. (2011); Imen et al. (2015); Gallardo et al. (2014a); Janicki (2012); Leis et al. (1997); Phythian et al. (1997); Wang and Lin (2007); Debyeche et al. (2010)). In (Phythian et al. (1997)), text-dependent speaker recognition is considered and the authors analyzed the deviations of many features like formant trajectories, pitch frequency, formant bandwidths, inter formant distances and formant frequencies due to

* Corresponding author.
E-mail address: athulya4nair@yahoo.com (M.S. Athulya).

https://doi.org/10.1016/j.diin.2018.03.005
1742-2876/© 2018 Elsevier Ltd. All rights reserved.

Please cite this article in press as: Athulya, M.S., Sathidevi, P.S., Speaker verification from codec distorted speech for forensic investigation through serial combination of classifiers, Digital Investigation (2018), https://doi.org/10.1016/j.diin.2018.03.005

codec distortions. They showed that the formant bandwidths for first formant (F1) are relatively unaffected. Changes in formant frequency due to codec impact does not depend on the bit rate of codecs (Guillemin and Watson (2006)). Analysis of the speaker recognition accuracy using speech distorted by codecs with different speech quality and bit rates revealed that the accuracy is not reduced with the quality and bitrate of the codecs (Petracca et al. (2006)). Speaker recognition accuracy using speech distorted by Mixed-excitation linear prediction (MELP) coder at 2.4 Kbits/s is more than that of using speech distorted with high bit rate GSM Adaptive Multi-Rate (GSM AMR 7.40 kb/s) coders (Petracca et al. (2006)).

A few methods have been reported to improve the speaker recognition accuracy from codec distorted speech. In (Stauffer and Lawson (2009)), by applying a speech coder called Speex to data compressed by GSM codec, the authors tried to improve the recognition rates. Methods for extracting the features directly from the coded speech were proposed in (Petracca et al. (2006); Quatieri et al. (2000); Dan et al. (2008); Peláez-Moreno et al. (2001)). But, the drawback of these methods is that we need to know the type of the codec apriori. The verification error rates are found to be almost unaltered when the training and testing data are compressed by the same codecs (Jiang et al. (2009)). When mismatched conditions occur, accuracy reduces (Jiang et al. (2009); Vuppala et al. (2010)). An affine transform based codec compensation was shown to give better performance for 8 kb/s ITU-T G.729 codec distorted test speech (Mudrowsky et al. (2010)). Moreover, a combination of feature sets gave better results than those from individual feature sets. Extracting features from the perceptually less relevant portion of the speech was considered in (Khan et al. (2010)). But, they obtained a high equal error rate (EER) of 23%. Authors claim that for investigation purposes, this could limit the focus of investigators to a few individuals. In (Yessad and Amrouche (2013)), a score fusion strategy was used to improve the verification accuracy.

In this work, a different approach is employed to improve the speaker verification accuracy for investigative purposes. A forensic crime scenario where a speech sample is obtained as evidence and many persons are suspected as the speaker, is considered here. In such a scenario, the correct identification of speaker can be made easy if it is possible to reduce the number of suspects. A series combination of GMM-UBM and SVM classifiers is implemented to bring down the number of suspected persons to a very small number. In GMM-UBM classification, only a few top scored speakers are taken. Only these speakers are now considered as the suspects, and again a classification using SVM classifier is performed. If SVM classification alone is used, the number of error points will be more since many wrong speakers will also be included. But, here by introducing GMM-UBM classifier as the first stage, the number of wrong speakers given to SVM classifier is reduced. This improves the classification performance of SVM classifier and reduces the overall error rate. A slightly modified version of recently developed PNCC features (Kim and Stern (2016)) is also proposed in this work. Even though PNCC features were developed for suppressing noise effect in noisy speech, it can be used as robust features for codec distorted speech (Wang and Wang (2016)). In the proposed system, we used a fusion of modified PNCC features and MFCC features to get better performance. The proposed system greatly reduces the EER as compared to other competent systems.

Most of the available speaker recognition systems need more than 10 s test data and prior knowledge of the codec. But, in the present work, prior knowledge of the codec is not required, and hence clean uncompressed data is used for training. In most of the previous works on speaker recognition, the test data selected were of duration more than 10 s. In our work, since we are dealing with forensic speaker recognition, we have to consider short-duration speech to mimic the real scenario. In some previous works on forensic speaker recognition, speech segments of about 2 s alone were considered (Enzinger and Morrison (2015); Kanagasundaram et al. (2011)). One of these works dealt with real forensic data, and the test utterance was of duration 2.08 s (Enzinger and Morrison (2015)). Hence, by considering the non-availability of long duration forensic speech in most of the cases, the duration of the test data is fixed as 2−3 s in the proposed work. The speaker recognition performance rate using speech distorted by CELP codec is found to be less than that of other codecs (Vuppala et al. (2010)). Therefore our work focusses on designing an efficient speaker verification system using CELP distorted speech so that it will guarantee a good performance with other codecs also.

Related work

A background study of the basic methods used in this work is presented here. In this work, Modified Power Normalized Cepstral Coefficients (MPNCC) fused with MFCC features are used as features for speaker verification, and GMM-UBM classifier followed by SVM classifier is used for classification purpose.

Modified Power Normalized Cepstral Coefficients

In (Kim and Stern (2016)), PNCC features were developed as robust features for reducing the effects of noise in speaker recognition. Its relevance to codec distortions is not explored much. The main difference of PNCC features from MFCC features is that in place of log nonlinearity used in MFCC, PNCC uses power-law nonlinearity.

PNCC in brief consists of the following stages (Kim and Stern (2016)):

- An initial processing stage in which we obtain the spectral power of pre-emphasized speech by taking its short time fourier transform (STFT) using a 25.6 ms hamming window and then weighing the STFT output magnitude square by the frequency response of a 40 channel gammatone filter bank.

$$P[m,l] = \sum_{k=0}^{(K/2)-1} \left|X\left[m, e^{j\omega_k}\right] H_l\left(e^{j\omega_k}\right)\right|^2 \quad (1)$$

- where m and l represent frame and filter indices, X, the speech spectrum and H, the filter response.
- In the next step, a medium time-power calculation is done using the equation,

$$\tilde{Q}[m,l] = \frac{1}{2M+1} \sum_{m'=m-M}^{m+M} P\left[m', l\right] \quad (2)$$

- This power is actually the moving average of $P[m,l]$. This is done due to the fact that noise power is a slowly varying factor as compared to speech power. This medium-time power is used only for the estimation and compensation of noise which modifies the short-time power $P[m,l]$
- Next is asymmetric noise filtering which is one of the most important steps in PNCC feature extraction. Here, an asymmetric nonlinear filter is used to get an estimate of the time-varying noise floor, and this estimate is subtracted from the instantaneous power. Temporal masking added to this enhances the onset of incoming power envelope.

Please cite this article in press as: Athulya, M.S., Sathidevi, P.S., Speaker verification from codec distorted speech for forensic investigation through serial combination of classifiers, Digital Investigation (2018), https://doi.org/10.1016/j.diin.2018.03.005

Download English Version:

https://daneshyari.com/en/article/6884413

Download Persian Version:

https://daneshyari.com/article/6884413

Daneshyari.com