NON INTRUSIVE CODEC IDENTIFICATION ALGORITHM

Dushyant Sharma, Patrick A. Naylor, Nikolay D. Gaubitch and Mike Brookes

Centre for Law Enforcement Audio Research (CLEAR)
Imperial College London

ABSTRACT

We present a non-intrusive data driven method for codec detection and identification in the presence of background noise. The method uses a number of speech features which are then used to train a CART classifier. We demonstrate the performance of the method using several different noise types over a wide range of SNRs. Our results show that we can identify a codec and its bit rate to an accuracy of 92% and we are able to detect the presence of a codec with an accuracy of 97% at -5 dB SNR.

***Index Terms*—** Speech CODEC Detection, Automatic Diagnosis, Quality of Service

1. INTRODUCTION

The past two decades has seen the standardisation of a number of speech coding and encoding (CODEC) algorithms and with the large volume of users of telecommunication systems, the challenge is to develop algorithms that minimise the data throughput in terms of bits per second while maximising the quality of the signal. Many of the speech CODECs operating with narrowband signals therefore have a low perceptual quality due to the high levels of compression that must be applied. The type of CODEC used in the transmission channel has a dominating effect on speech quality in the absence of channel artefacts [1]. Additionally, in the field of law enforcement audio processing, it is often required to validate the collection mechanism where an important issue is the identification of the CODEC used in the transmission channel. The presence of a particular CODEC has been shown to have adverse effects on many speech processing systems. A study by Besacier et al. [2] has shown that the presence of a GSM coding significantly degrades the performance of speaker identification and verification. The identification is required to be non-intrusive, since the original source speech signal is not usually available. Previous studies on CODEC identification include a study by Ludwig et al. [3] which presents a multi-dimensional Gaussian classifier that is able to detect low and high bit rate CODECs with an accuracy of 97%. A more recent study presents a spectral harmonic decomposition (SHD) based CODEC detection method which is able to detect five types of CODEC with an accuracy better than 92% [4]. This method uses the average long term noise spectrum from the SHD of the input signal as the feature and a simple cross correlation based classifier is used to assign the signal to one of the noise templates. It should be noted that previous studies have concentrated on CODEC identification in clean speech.

In this paper we present a non-intrusive CODEC detection method that is robust to additive noise. The method is tested with 180 conditions (car, babble and hum noise) at realistic SNR's (-5 to 15 dB range) and nine coding systems. Also, we present the signal properties that are useful in CODEC class detection. Our method is able to detect the bit rate of the system as well as the type of CODEC. The remainder of this paper is organised as follows: in Section 2 we present the algorithm along with the features used in the method. In Section 3 the database and evaluation criteria are presented and the experimental results are given in Section 4. Finally, conclusions are drawn in Section 5.

2. ALGORITHM OVERVIEW

The algorithm presented here is a data driven, non-intrusive approach to CODEC detection and identification based on the feature extraction framework described in [5, 6]. The algorithm begins by segmenting the sampled speech signal $s(n)$ into non-overlapping frames of 20 ms duration using a Hanning window. Then for each of the N frames of the signal, 85 features are extracted (ψ_n). The method then computes the statistics from the local features, resulting in a single vector of 340 features per signal ($\mathbf{\Psi}$). These are then used to train a classification tree (CART).

2.1. Frame Based Feature Extraction

The first step is to extract the 85 dimensional per frame feature vector ($\boldsymbol{\psi}$), which includes the 10th order Linear Prediction Coding (LPC) coefficients as well as the spectral flatness, spectral centroid and spectral dynamics of the magnitude response of the LPC spectrum. The energy per frame, zero-crossing rate and the Hilbert envelope is calculated per frame and its variance and dynamic range are included. The Importance weighted Signal to Noise Ratio (iSNR) is calculated according to [5] and included along with the pitch period estimate using the PEFAC algorithm [7, 8]. The rate of change

of all features are also included (except for spectral dynamics feature). We also include 12th order Mel Frequency Cepstral Coeffcients using FFTs along with the acceleration and velocity coefficients. Finally, an estimate of the long term acoustic channel is derived using a blind channel estimation (BCE) algorithm [9] and the spectral flatness, dynamics and centroid of the channel estimate are included.

2.2. Statistical Description

The mean, variance, skewness and kurtosis of the 85 per-frame features provides a statistical description of the features for each signal. This results in a 340 dimensional global feature vector ($\mathbf{\Psi}$) per audio file.

2.3. CART Classifier

The global features ($\mathbf{\Psi}$) and their corresponding class labels ($\mathbf{\Theta}$) are used to train a classification and regression tree (CART) [10]. The CART method is an recursive partitioning tree based method and has the desirable property of being human interpretable and low computational complexity in the test mode. The initial CART model is further pruned by 10-fold cross-validation over the training partition of the database with the objective of removing branches giving small reductions in the misclassification rate.

3. METHODOLOGY

In order to evaluate the performance of our algorithm, a database was constructed, as described in Section 3.1. The partitioning of the database into a test and training set is described in Section 3.2, followed by an explanation of the metrics used for evaluation in Section 3.3.

3.1. Database

The database used for the experiments in this paper is based on speech from 48 speakers from the TIMIT database [11]. The TIMIT database contains speech from American English speakers representing various accents. The test and training partitions each contain 24 distinct speakers, without any overlap of speech. The database consists of 180 conditions, representing nine coding conditions:

- Linear PCM - uncompressed data transmission.
- GSM Full Rate (GSM-FR) [12] - representing baseline mobile transmission.
- GSM Adaptive Mulit Rate (AMR) at 4.5kbs, 5.4 kbs, 7.4kbs and 12.4kbs [13] - representing commonly occurring mobile transmission bit rates.
- ITU-T G.711 [14] - representing typical infrastructure routed transmission.
- GSM transcoding (TRANS) - an example transcoding scenario. GSM to GSM communication that has to be routed through infrastructure, which is typically using a G711 CODEC (GSM-G711-GSM).
- MP3 (16 kbs) - represents speech recorded by portable recorders.

The speech is grouped into 20 base conditions, of which 5 conditions represent clean speech and the remaining 15 conditions are additive noise conditions. The noise types include car, babble and hum at signal-to-noise ratios (SNRs) of 15, 10, 5, 0, -5 dB. The base conditions are then processed by the nine coding systems described above.

3.1.1. Training

The database is partitioned into a test and train partition, each containing 3360 audio files representing all conditions. The noise sources in the test and train partitions are separate (separate recordings) to ensure the classifier is not trained with the same noise source as that in the test set. A 50% cross validation is used to evaluate the performance of our method (the test and training partitions contain different speakers, speech material and noise files).

3.2. Evaluation

The performance of the CODEC detection classification is assessed with three key metrics:

- Hit Rate - this is the percentage of files correctly classified.

$$HR_x = \frac{\sum_{i=1}^{N} I(\tilde{\mathbf{\Theta}}_i^x, \mathbf{\Theta}_i^x)}{N} \times 100, \qquad (1)$$

  where $\tilde{\mathbf{\Theta}}_i^x$ is the estimated class label according to detection criteria x and $\mathbf{\Theta}_i^x$ is the actual class label for the i^{th} file. The total number of files in the test set is N and $I(a, b)$ is an index function defined as:

$$I(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

- False Positives Rate - this is the percentage of files that have been classified as belonging to the class of interest when the ground truth suggests a different class:

$$FPR_x = \frac{\sum_{i=1}^{N} I(\tilde{\mathbf{\Theta}}_i^{x=1}, \mathbf{\Theta}_i^{x=0})}{N} \times 100, \qquad (3)$$

  where $\tilde{\mathbf{\Theta}}_i^{x=1}$ is the estimated class label when the algorithm estimates the detection class to be present and $\mathbf{\Theta}_i^{x=0}$ is the ground truth class label when the detection class label if false.

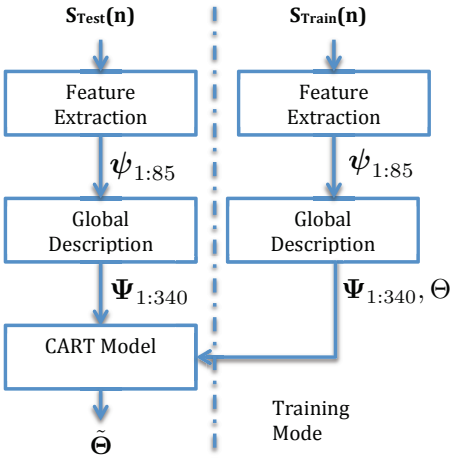

Fig. 1. Algorithm framework with training mode shown in the right partition of the figure and test mode in the left half.

- False Negatives Rate - this is the percentage of files where the class of interest is present in the ground truth but the algorithm failed to detect it:

$$FNR_x = \frac{\sum_{i=1}^{N} I(\tilde{\Theta}_i^{x=0}, \Theta_i^{x=1})}{N} \times 100, \quad (4)$$

4. RESULTS

In this section, the performance of our non-intrusive method is assessed on the database described in Section 3. A number of classifiers have been tested as shown in Table 1 .

4.1. CODEC Type and Bit Rate Detection (Θ_{All})

In this configuration, the CART model is trained and tested with each individual CODEC type. The objective of the system is to identify the type of CODEC present along with the bit rate used for encoding the signal. This is the most challenging objective and our method achieves a hit rate of nearly 92% in this scenario. The average power of the LPC residual per frame is the most important feature for this type of classifier. The classification tree in this mode uses 29 of the 340 global features to perform the classification.

4.2. CODEC Type Detection (Θ_{Class})

The model in this configuration is tasked with identifying the class of CODEC used, that is one of (PCM, GSM, G711, MP3, Transcoding or AMR). The bit rate of the system is not identified and this gives a hit rate of nearly 95%. This compares favourably with previous CODEC type detection algorithms presented [4]. The most important feature here is the kurtosis of the rate of change of the pitch period, with the

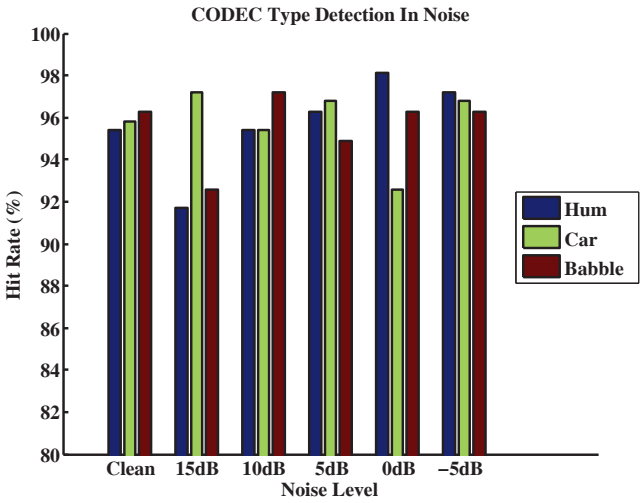

Fig. 2. CODEC type detection (Θ_{Class}) with different noise types (hum, car, babble) and SNR in terms of the hit rate.

classifier using 17 global features in total. Figure 2 shows the performance of this classifier for different test noise types and SNR. The performance can be seen to be very similar for all conditions, suggesting that the method is robust to additive noise with hit rates greater than 91% and a standard deviation of 1.7%.

4.3. Single CODEC Detection (Θ_x)

In addition to identifying the type of CODEC present, it is sometimes beneficial to test for the presence or absence of a single CODEC. Table 1 shows the performance of the possible binary classifiers for our database. The classifier in each case is trained to identify just one CODEC (Θ_x, $_x = [PCM, G711, GSM - FR, AMR, MP3, TRANS]$). In this configuration the performance of the algorithm is very good due the easier classification task with hit rates better than 97% in all cases. Also, the false positive rate is generally lower than the false negative rate. The mean and variance are important global descriptors and LPC and BCE are important per-frame features, which the CART model using between 6 and 14 features in all.

5. CONCLUSIONS

In this paper, we have described a non-intrusive CODEC detection and identification algorithm that is able to identify both the CODEC used in a communication channel and its operational bit rate with an accuracy of 92%. The algorithm was tested using a 50% cross validation on a database with 180 test conditions comprising three noise types (car, babble and hum) at five SNRs processed through nine CODECs.

Detection Criterion	HR (%)	FPR (%)	FNR (%)	NFeatures	Best Per-Frame Feature	Global Descriptor
Θ_{All}	91.9	-	-	29	LPC Residual Power	Mean
Θ_{Class}	94.9	-	-	17	d/dt(Pitch)	Kurtosis
Θ_{PCM}	98.4	0.5	1.1	6	d/dt(Spectral Dynamics of BCE)	Variance
$\Theta_{G.711}$	97.1	1.0	1.9	8	d/dt(6th LPC Coefficient)	Variance
Θ_{GSM-FR}	97.0	1.8	1.2	14	LPC Residual Power	Skewness
Θ_{AMR}	99.0	0.8	0.2	9	d/dt(10th LPC Coefficient)	Mean
Θ_{MP3}	98.0	1.1	0.9	7	Spectral Flatness of BCE	Mean
Θ_{TRANS}	98.2	0.6	1.2	7	LPC Residual Power	Mean

Table 1. Classification results for 50% cross validation on database (Section 3.1). The Hit Rate (HR), False Positives Rate (FPR) and False Negative Rate (FNR) are given as a % of the total number of files in the test set. The decision criteria are shown in the first column and the most important per-frame feature and associated statistical descriptor as the last two columns. The number of global features deployed in the final CART model is also shown (NFeatures).

Also, the features most important for each detection criteria were presented.

It was found that the LPC, blind channel estimate (BCE) and pitch period based features were generally most important, with the mean and variance being important global descriptors. Our method has been shown to be robust to additive noise, with standard deviation in hit rate being 1.7% over the SNR range. Additionally, the algorithm has been tested with various binary classification criteria, where the average hit rate is higher than 97%, comparable to results from previous studies on clean speech.

6. REFERENCES

[1] T. Ludwig and U. Heute, "Detection of digital transmissionsystems for voice quality measurements," in *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, 2001, pp. 1699 – 1702.

[2] L. Besacier nad S. Grassi, A. Dufaux, M. Ansorge, and F. Pellandini, "GSM coding and speaker recognition," in *Proc. ICASSP*, 2000, vol. 2, pp. 1085–1088.

[3] T. Ludwig, "Comfort noise detection and GSM-FR codec detection for speech -quality evaluations in telephone networks," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 309–312.

[4] K. Sholz, L. Leutelt, and U. Heute, "Speech-codec detection by spectral harmonic-plus-noise decomposition," in *Proc. Asilomar Conference on Signals, Systems and Computers*, 2004, pp. 2295–2299.

[5] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *Proc. European Signal Processing Conference (EUSIPCO)*, Denmark, Aug. 2010.

[6] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1948–1956, Nov. 2006.

[7] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2011.

[8] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997.

[9] Nikolay D. Gaubitch, Mike Brookes, and Patrick A. Naylor, "Blind channel identification in speech using the long-term average speech spectrum," in *Proc. European Signal Processing Conference (EUSIPCO)*, Glasgow, Aug. 2009.

[10] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, CRC Press, 1984.

[11] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Dec. 1988.

[12] European Telecommunications Standards Institute (ETSI), "GSM 06.10: Full rate (FR) speech transcoding," 1995.

[13] European Telecommunications Standards Institute (ETSI), "GSM 06.90: Adaptive multi-rate (AMR) speech transcoding," 1998.

[14] International Telecommunications Union (ITU-T), "Pulse code modulation (PCM) of voice frequencies," Nov. 1998.