HIGHLY ACCURATE NON-INTRUSIVE SPEECH FORENSICS FOR CODEC IDENTIFICATIONS FROM OBSERVED DECODED SIGNALS

Frank Jenner, Andres Kwasinski

Department of Computer Engineering
Rochester Institute of Technology, Rochester, NY 14623
{faj6816, axkeec}@rit.edu

ABSTRACT

The ability to detect a particular speech codec from only the decoded audio has several useful forensic and system performance improvement applications. This paper presents a novel scheme for non-intrusive identification of speech codecs. The identification approach is based upon comparing a profile of a set of noise spectra and a time-domain histogram from the decoded speech to those from the candidate codecs. The presented results show a very high accuracy in identifying speech contemporary codecs from a diverse set of types and encoding rates. The presented codec identification scheme has a very low misidentification rate, including in the high coding rate regime where it improves on previous works by achieving perfect identification. This performance is achieved while reducing the duration of the analysis window of speech from 2 minutes to only 4 seconds.

Index Terms— Speech forensics, speech coding, vocoder, identification, processing traces.

1. INTRODUCTION

Decades of research into speech coding for telecommunications networks has resulted in the development of numerous codecs. Since the performance of other speech processing subsystems in the network varies with the codec being used, identifying this codec from the decoded speech presents numerous valuable applications. For example, the identification of the speech codec would allow for more accurate online voice quality estimation in VoIP Networks [1], would also allow for speech recognition algorithms that are more accurate by being tailored to the codec in use [2] and even enable targeted content delivery (e.g. a consumer navigating a voice menu system during a call to their cable internet provider might be presented with an advertisement for VoIP phone service if the system detects that the user is calling from a cellular or PSTN network). Furthermore, since usually each communication network or application utilizes its own standardized codec (e.g. a call through a GSM cellular network would use the AMR codec and a Skype VoIP call would use Silk), the identification of the speech codec from decoded speech has obvious key forensic use to identify call provenance and to detect doctored speech sequences and instances of caller ID spoofing.

Despite the numerous applications, little research can be found in this area. A method to identify the provenance of speech calls can be found in [3]. Since the focus of [3] is in identifying the call provenance, and not necessarily the specific codec used, the presented technique is based on identifying artifacts specific to each possible path (VoIP, landline or cellular networks) such as signs of packet drops for VoIP, and noise profiles differentiating landline and cellular networks. The work in [4] presents a technique for the identification of speech codecs that is based on distinguishing the noise spectrum characteristic of each codec. This noise spectrum is calculated from the difference between the actual spectrum and a harmonic spectrum, resulting from the harmonic/noise decomposition from the multiband excitation (MBE) vocoder model [5]. This approach shows accurate results in identifying many speech codecs, but it does so at the cost of a using a very long analysis window of about 2 minutes of audio. In addition, the accuracy of this approach is degraded when trying to identify codecs operating at high coding rates.

This paper presents a novel technique that improves upon the approach in [4] and results in highly accurate identification of the codec with which an unknown signal has been processed. The technique is based on the decoded speech stream only, and does not require access to the channel bitstream or knowledge of the original speech sample. Because most speech compression schemes are lossy, the presented technique is based on identifying the traces of the lossy signal processing operations that are present in the decoded speech. The results presented in this paper will show very high accuracies in identifying the speech codec, including those operating at high coding rates. This is achieved even when using a short analysis window of only 4 seconds.

2. IDENTIFICATION OF SPEECH CODEC FROM SIGNAL PROCESSING TRACES

The goal of the developed technique is to process a sequence of decoded speech so as to identify the codec used. The result of such processing should be, when analyzing speech samples from the same codec, largely independent of speaker characteristics, such as pitch range and formant structure. Conversely, the output of such processing should vary significantly when analyzing speech samples from different codecs, even if the original input signal to the encoders was identical.

The presented novel technique augments the identifica-

tion criteria in [4], which performs codec identification using only one feature of the input speech (the noise spectrum computed by subtracting the decoded speech spectrum from the spectrum of the MBE modeled speech signal), by creating a multidimentional profile whose features include noise spectra from multiple speech models. As will be seen, this approach significantly improves the identification performance. Furthermore, because the noise spectrum does not carry information about the time domain representation of the input signal, a time domain amplitude histogram is also added to the profile in order to capture possible quantization patterns imparted by the codec. The profile is therefore a set of properties computed from the input signal that attempt to provide a fingerprint of certain input signal characteristics and signal processing traces that are related to a specific source codec.

The presented codec identification technique is shown in Fig. 1. First, the input speech, $x(t)$, previously encoded and decoded with an unknown codec, is reprocessed by each of the candidate codecs by undergoing a new encoding-decoding processing. Speech codecs typically employ some sort of underlying speech synthesis model. For example, the MBE vocoder model generates speech by providing pitch information, and voicing decision, phase, and amplitude for each harmonic band in a given frame of speech. More common vocoders typically provide information about the pitch, formant envelope, and excitation signal for the frame of speech. In either case, the output of a given vocoder is generated directly from applying such parameters to some form of human speech synthesis model. In a sense, then, reprocessing the speech with a codec eliminates signal components that are not part of human speech in accordance with that codec's underlying speech synthesis model, and outputs a perfect speech signal as conformed to the model. In addition, the reprocessing of the decoded speech will result in a noise spectrum with a more clearly defined features for the codec truly originally used and a more non-descriptive noise spectrum, with less defined features, for the rest of the candidate codecs.

Each of the speech sequences at the output of each codec reprocessing stage, $x_1(t), x_2(t), \ldots, x_m(t)$, is then windowed into frames of 256 samples (32 ms at an 8 kHz narrowband sampling rate) with a 50% overlap. It is important to select a frame length that is both long enough to provide an adequate number of samples for analysis and short enough for the audio signal to remain fairly stationary. A segment of between 10 and 40 ms is generally regarded to be an appropriate frame length for speech signal analysis [6]. The speech is then processed by a voicing activity detector to determine whether each frame is voiced, unvoiced, or silence. In general, most of the effort in speech coding is spent on the modeling and compression of voiced speech, as voiced speech accounts for more of the bandwidth of the coded signal. Significantly, the coding of voiced speech accounts for the more sophisticated techniques, including those frequency domain processing that would leave detectable traces in the noise spectrum. Unvoiced speech is usually synthesized by passing random noise through a simple filter, and silence might be neglected altogether or be encoded into some form of very low bit rate background comfort noise. Thus, it is expected that any distinguishing characteristics in the output signal from a speech codec will be prevalent in the voiced frames, and we discard any unvoiced or silent frames from further analysis.

Next, the voiced frames from each of the audio streams $x_1^v(t), x_2^v(t), \ldots, x_m^v(t)$ are transformed into the frequency domain by means of a Fast Fourier transform (FFT). Each frame is first multiplied by a hamming window and zero padded to $L = 4096$ points in order to increase the frequency resolution of the resulting magnitude spectra, $|X_j|, j = 1, \ldots, m$. Each of these spectra is subtracted from the magnitude spectrum of the corresponding frame in the original input signal, $x(t)$ to form the noise spectra $|X_{N,j}| = |X| - |X_j|$. These noise spectra represent the spectral differences between a perfectly modeled rendition of speech and the actual input signal, thereby helping to reveal some of the artifacts imparted on the original speech from the codec. As more frames are processed, the noise spectra are accumulated for up to k voiced frames to capture a variety of voiced phonemes. These aggregated noise spectra, call them $\hat{X}_{N,j}$, collectively form a profile that characterizes the codec present in the original input signal in conjunction with the codec used during reprocessing.

However, by analyzing the input signal only in the frequency domain, we neglect the possibility of extracting valuable traces of the signal processing that manifest themselves in the time domain. Codecs may be limited in the range or set of sample amplitudes that can occur at the output. For example, the output samples from ITU-T G.711 codecs are quantized to set of 256 discrete amplitudes from among the usual 16-bit linear PCM space used for representation in memory. Thus, in our approach, a histogram of the input speech sample amplitudes (including from voiced, unvoiced, and silent frames) is also collected and used as a feature of the profile for the input signal.

Before being able to identify codecs from signals of unknown origin, it is necessary to first generate a set of training profiles from known codecs against which to compare. In this research, we are interested in detecting the following m diverse collection of codecs: G.711, G.726, G.728, G.729, iLBC, AMR Narrowband, and Silk. To create the training profiles, 100 randomly selected speech samples from the "training" partition of the Texas Instruments/MIT (TIMIT) speech corpus [7] were processed by each of the m codecs, and then profiled using the previously described strategy.

To determine the codec that is present in an unknown signal, the new signal must be profiled, and then its profile compared to each of the training profiles. This comparison is performed through the use of several normalized cross-correlations between corresponding features in each profile.

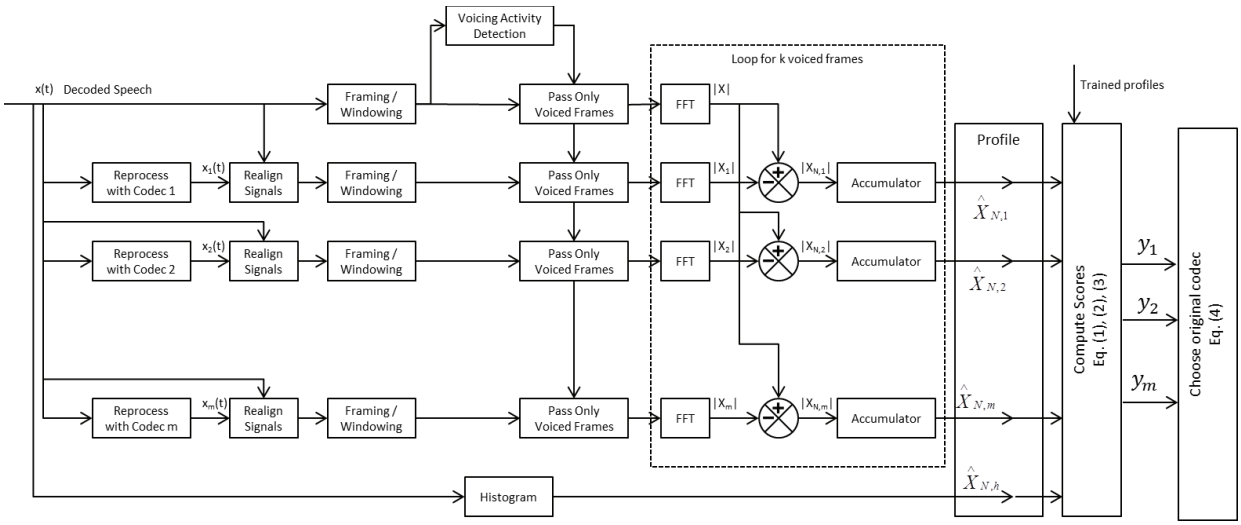

Fig. 1. Profiling strategy for audio samples

The normalized cross-correlation function is shown in (1). This equation demonstrates the comparison between a noise spectrum, $\hat{X}_N$, from the unknown signal's profile and the corresponding noise spectrum, $\bar{X}_N$, from a trained profile, although it is also used to compare the histograms between profiles.

$$\rho = \frac{\left\langle \bar{X}_N, \hat{X}_N \right\rangle}{\|\bar{X}_N\| \|\hat{X}_N\|} = \frac{\sum_{i=1}^{L} \bar{X}_N[i] \hat{X}_N[i]^*}{\sqrt{\sum_{i=1}^{L} |\bar{X}_N[i]|^2} \sqrt{\sum_{i=1}^{L} |\hat{X}_N[i]|^2}} \quad (1)$$

The normalized cross-correlations are then used to compute comparison metrics as shown in (2). The values $y_i, i = 1, \ldots, m$ are a score measuring how closely the unknown profile matches the training profile of the i-th codec. These scores y_i are calculated as the weighted sum of values $\rho_{i,j}, j = 1, \ldots, m, h$. The values $\rho_{i,j}, j = 1, \ldots, m, h$ are the normalized cross-correlation of the j-th profile feature (where $1, \ldots, m$ are the noise spectra from the m voice models, one for each candidate codec, and $h = m + 1$ is the histogram) between the unknown profile and the training profile of the i-th codec.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,m} & \rho_{1,h} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,m} & \rho_{2,h} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{m,1} & \rho_{m,2} & \cdots & \rho_{m,m} & \rho_{m,h} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \\ a_h \end{bmatrix} \quad (2)$$

The weights a_j in (2) take into consideration that the different profile features are not equally effective at correctly identifying a particular codec. Through extensive empirical testing that measured the accuracy of the codec detection scheme using different weights, we observed that the histogram is the most effective feature in successfully identifying codecs and that all m noise spectra were approximately equally effective between themselves. Thus, the histogram was assigned half of the overall weight, while the other half was uniformly distributed among the m noise spectra, as shown in (3).

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \\ a_h \end{bmatrix} = \begin{bmatrix} 1/2m \\ 1/2m \\ \vdots \\ 1/2m \\ 1/2 \end{bmatrix} \quad (3)$$

Finally, the codec originally used is determined from the scores y_i through the decision

$$\text{Codec Index} = \arg\max_i y_i : y_i \in \{y_1, y_2, \ldots, y_m\} \quad (4)$$

3. RESULTS

To test the proposed identification scheme, speech segments from all 168 speakers in the "test" partition of the TIMIT speech corpus were used. This ensures that all speakers and sentences used in testing are mutually exclusive from those that were used to create the training profiles. Each speaker's speech clips are truncated or concatenated as necessary to form a speech sequence that contains exactly k voiced frames. For the majority of our experiments, an analysis length of $k = 160$ voiced frames is used. For the particular voicing activity detection algorithm used in our research, this corresponds to an average of 3.94 seconds of audio per speaker. These audio clips are processed by several codecs at different bitrate settings to form the set of input signals (168 for each source codec). Each input signal is then profiled and compared to the training profiles following the procedure explained in the previous Section. The set of candidate codecs was chosen to be diverse in the underlined technique used and the coding rate. The candidate codecs where: G.711 (waveform μ-law companding at 64 kbit/s), G.726 (waveform ADPCM, with rates 16, 24, 32 and 40 kbit/s), G.728 (low delay

			Classified As						
			G.711	G.726	G.728	G.729	iLBC	AMR	Silk
Source Codec	G.711	μ-law	**100.00%**	-	-	-	-	-	-
	G.726	40 kbit/s	-	**100.00%**	-	-	-	-	-
		32 kbit/s	-	**100.00%**	-	-	-	-	-
		24 kbit/s	-	**100.00%**	-	-	-	-	-
		16 kbit/s	-	**100.00%**	-	-	-	-	-
	G.728	16 kbit/s	-	-	**100.00%**	-	-	-	-
	G.729	11.8 kbit/s	-	-	-	**96.43%**	0.60%	-	2.98%
		8 kbit/s	-	-	-	**100.00%**	-	-	-
		6.4 kbit/s	-	-	-	**100.00%**	-	-	-
	iLBC	15.2 kbit/s	-	-	-	11.90%	**88.10%**	-	-
		13.33 kbit/s	-	-	-	14.29%	**85.71%**	-	-
	AMR	12.2 kbit/s	-	-	-	-	-	**100.00%**	-
		10.2 kbit/s	-	-	-	-	-	**100.00%**	-
		7.95 kbit/s	-	-	-	-	-	**100.00%**	-
		7.4 kbit/s	-	-	-	-	-	**100.00%**	-
		6.7 kbit/s	-	-	-	-	-	**100.00%**	-
		5.9 kbit/s	-	-	-	-	-	**100.00%**	-
		5.15 kbit/s	-	-	-	-	-	**100.00%**	-
		4.75 kbit/s	-	-	-	-	-	**100.00%**	-
	Silk	VBR	-	-	-	21.43%	-	-	**78.57%**

Table 1. Results for analysis $k = 160$ voiced frames (3.94 sec.)

CELP vocoder at 16 kbit/s), G.729 (CS-ACELP vocoder at 11.8, 8 and 6.4 kbit/s), iLBC (vocoder at 15.2 kbit/s for 20 ms frame and 13.33 kbit/s for 30 ms frame), AMR (vocoder at 12.2, 10.2, 7.95, 7.4, 6.7, 5.9, 5.15 and 4.75 kbit/s) and Silk (Skype's variable bit rate audio codec).

Results are shown in Table 1. Each row illustrates the distribution of how the identification strategy classifies the 168 speech inputs when initially processed using the codec on the left-hand side of the table. Figures in bold print indicate the percentage of those input sequences that were correctly identified. The remainder indicate the proportions of inputs that are misclassified.

Overall, the results are very favorable, with the majority of the codecs being correctly identified for all 168 test inputs. In all cases the presented scheme shows more accurate results than those reported in [4], despite the fact that our scheme uses approximately 4 seconds of speech, while [4] requires about 2 minutes. Of particular note are the results for high bit rate encoding. In the case of the G.726 (ADPCM) where our technique achieves accuracy of 100 % at all bit rates, while the results in [4] are 86 % at 32 kbit/s and 24 % at 40 kbit/s. As seen in Table 1, the codec that presents room for future improvement is Silk. We believe that the accuracy of our approach will be improved by considering the dynamic variable bit rate operation.

4. CONCLUSION

In this paper we have presented a novel highly accurate technique for the identification from the decoded speech of the codec used. The technique does not require access to the original uncoded speech and is based on identifying the traces left from the signal processing operation performed during encoding and decoding. The identification scheme operates by creating a multidimentional profile that includes noise spectra from multiple speech models and a time domain amplitude histogram. This profile is compared against reference profiles from the candidate codecs. The results show that the proposed technique is highly accurate, with 100 % correct identification for most of the candidate codecs, which represent a diverse set including waveform codecs, vocodecs and low and high coding bit rate operation. Furthermore, the accurate performance is achieved when using input speech sequences with much shorter duration than previously reported techniques. The proposed technique outperforms previously reported techniques, specially for high source coding rate.

5. REFERENCES

[1] L. Sun and E. C. Ifeachor, "Voice quality prediction models and their application in voip networks", *IEEE Trans. on Multimedia*, 8(4):809-820, Aug. 2006.'

[2] L. Besacier, C. Bergamini, D. Vaufreydaz, and E. Castelli, "The effect of speech and audio compression on speech recognition performance", *IEEE Workshop on Multimedia Signal Processing*, pp. 301-306, 2001.'

[3] V. A. Balasubramaniyan, A. Poonawalla, M. Ahamad, M. T. Hunter, and P. Traynor, "PinDr0p : Using Single-Ended Audio Features To Determine Call Provenance Categories and Subject Descriptors", *ACM Conference on Computer and Communications Security*, pp. pp. 109-120, 2010.'

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

[5] D.W. Griffith and J.S. Lim, "Multiband excitation vocoder", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 36(8):1223-1235, Aug. 1988.'

[6] L. Hanzo, F. C. Somerville, and J. Woodard, "Voice and Audio Compression for Wireless Communications", IEEE Press, 2nd. edition, 2007.

[7] DARPA TIMIT - "Acoustic-phonetic continuous speech corpus cd-rom"