PERFORMANCE ANALYSIS OF COMPRESSED-DOMAIN AUTOMATIC SPEAKER RECOGNITION AS A FUNCTION OF SPEECH CODING TECHNIQUE AND BIT RATE

M. Petracca, A. Servetti, J.C. De Martin

Dipartimento di Automatica e Informatica – Politecnico di Torino
Corso Duca degli Abruzzi, 24 — I-10129 Torino, Italy
E-mail: [matteo.petracca|servetti|demartin]@polito.it

ABSTRACT

Compressed-domain automatic speaker recognition is based on the analysis of the compressed parameters of speech coders. The objective is to perform low-complexity on-line speaker recognition for VoIP in the compressed domain, without the need to decode or resynthesize the speech bitstream. In this paper, we present initial results in determining the recognition accuracy that can be achieved with five widely used speech coding standards. Experiments with a database of 14 speakers obtain a recognition ratio close to 100% after the analysis of 30 seconds of active speech for most of the considered speech coders and rates. In particular, the results show that performance does not strictly depend on coding rate or codec speech quality.

1. INTRODUCTION

The Internet is rapidly evolving into a universal communication network that carries all types of traffic, including voice, video and data. Among them, the most important trend over the past few years was arguably the rapid growth of voice over IP (VoIP) services. In the coming years, with the continue increase in use of VoIP telephony, there will also be increased interest in the availability of online speaker recognition systems for providing various interactive voice services via VoIP phones. Additionally, fast and scalable processing of VoIP packets for speaker identification will be a requirement for law enforcement agencies when wiretapping and eavesdropping on VoIP provider high traffic networks would be necessary.

However, traditional automatic speaker recognition (ASR) cannot be directly applied to live VoIP calls because it operates on the uncompressed (PCM) speech waveform while voice travels the IP networks mostly in a compressed format. Before transmission, in fact, the sender applies compression standards to reduce the amount of information that must be sent to the other party. As a consequence the data has to be decompressed to obtain an appoximation of the original voice signal waveform before traditional speaker recognition methods can be applied. This time- and resource- consuming process is therefore unsuitable for an implementation in VoIP apparatuses or network sniffers where a large number of calls should be monitored simultaneously.

In this paper, we consider an alternative approach for performing online speaker recognition from live streams of compressed voice packets. This method has been previously presented as *compressed-domain automatic speaker recognition* (CD-ASR) in [1] [2] where voice feature vectors are made up of compressed bitstream values from coded speech frames. In [1] a tentative implementation limited to the GSM Adaptive Multi-Rate (AMR) standard at 12.2 kb/s showed that, in some circumstances, speaker recognition in the compressed domain is possible (for that particular coder) after the analysis of about 20 seconds of active speech. The objective of this paper is to investigate if CD-ASR is applicable in a broader context to othe compressed speech formats, or, within the GSM AMR standard, to other coding bitrates. In particular we adapt the speaker recognition algorithm to widely used speech codecs for VoIP telephony that differ not only in the bitrate, but also in the compression technique. We consider, in fact, a low-bitrate LPC based mixed excitation (MELP) vocoder, some analysis-by-synthesis algorithms with multipulse (G.723) or codebook (GSM AMR, G.729) based excitation model, and a coder that does not employ inter-frame prediction (iLBC).

The rest of this paper is organized as follows. An overview of ASR approaches is presented in Section 2 where, besides traditional systems that uses clean voice waveforms as input, we describe other approaches that work with coded speech. CD-ASR is then discussed in Section 3. In Section 4 we investigate the recognition rate achieved in our experiments with various speech codecs at different bit rates.

2. OVERVIEW OF AUTOMATIC SPEAKER RECOGNITION APPROACHES

Figure 1 illustrates the encoding, transmission and decoding chain for VoIP communications. Within this context, the four

The work was supported in part by Motorola Electronics S.p.A., MDB Development Center, Turin, Italy.

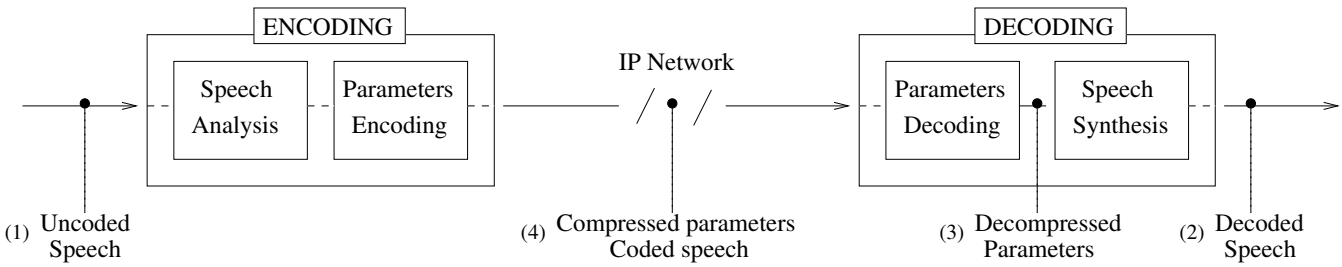

Fig. 1. In VoIP communications, the sender applies encoding standards to reduce the amount of information that is sent through the IP network. Hence, speech data traverses the network in a coded format and it has to be decoded and resynthesized at the receiver to obtain a voice signal similar to the original waveform.

mostly used ASR approaches may work, with different level of complexity and performance, at the sender with uncoded speech (1), at the receiver with decoded speech (2), at the receiver with decompressed parameters (3), in the IP network with coded speech and compressed parameters (4).

In the first, most traditional, case input material is a digitalized PCM representation of the voice waveform (i.e., *uncoded speech*). This signal is Fourier transformed into the frequency domain where the magnitude spectrum from a short-time frame of speech is extracted. The spectrum is then pre-emphasized and processed by a simulated mel-scale filterbank. Finally, the log-scaled output energy of each individual filter is cosine transformed to produce the cepstral coefficients. This processing may occur every 10 ms, producing 100 feature vectors per second that are then used in a classification algorithm such as the Gaussian Mixture Model - Universal Background Model (GMM-UBM) as presented in [3].

In the recent years however, due to the widespread use of digital speech communication systems, there has been an increasing necessity of a second automatic speaker recognition approach that uses *decoded speech*. The effect of speech coding/decoding on speaker and language recognition tasks has been analyzed for several coders and a wide range of bit rates (e.g., GSM at 12.2 kb/s, G.729 at 8 kb/s, and G.723.1 at 5.3 kb/s) [4]. These studies showed that straightforward application of traditional GMM-based speaker verification on the re-synthesized speech generally degrades with coder bit rate, relative to an uncoded baseline.

A third alternative, the parametric approach, was investigated to reduce the computational load related to the synthesis process [5]. In the parametric approach, the goal is to perform speaker recognition using a feature vector consisting of *decompressed parameters* representing both the all-pole spectrum and the corresponding prediction residual.

More recentely, the fourth compressed-domain ASR approach started exploring the possibility of working directly in the compressed domain with *coded speech and compressed parameters*, so that no decoding is applied, thus lowering the computational requirements with respect to previous mentioned approaches.

Moreover, in the specific context of CD-ASR applied to live VoIP calls, some works investigated the recognition accuracy achievable using techniques able to easily scale in terms of CPU, disk access, and memory use for many data streams. Drawbacks of traditional approaches such as CPU intensive operations (i.e., Fourier transform, mel-scale filters, cosine transforms) and memory consuming algorithms (i.e., gaussian mixture models, neural networks) are rejected in favor of lightweight clustering algorithms [2] or medium-term statistical analysis [1]. One of the benefit from the this tentative idea that we are trying to investigate would be its low memory requirement when applied over many data streams simultaneously. This is because the large volumes of data arriving in a stream may render some traditional algorithms inefficient. Using aggregation techniques, that is the process of computing statistical measures such as means and variance that summarize the incoming stream, we aim instead at keeping constant the amount of data to be processed with respect to the length of the analysis window.

3. COMPRESSED DOMAIN ASR

In the literature there have been several studies on the choice of acoustic features in speaker recognition tasks. Average fundamental frequency has been found to be a useful discriminating feature, as have gain measurements and long-term speech spectra, and cepstral coefficients.

In the approach under investigation, the feature space is instead derived from bitstream values of compressed speech. In this particular case our study extends over various speech codecs the results in [1] where CD-ASR was only applied to the bitstream generated by the widely used GSM AMR speech coder at 12.2 kb/s, the default speech coder for GSM 2+ and WCDMA third generation wireless systems [6]. Although compressed speech parameters are non-linearly related to the more physically meaningful features, each compressed voice packet explicitly carries a set of important voice characteristics (e.g., voice tract filter model parameters, pitch delay, amplitude) that can be used to create a voice feature vector for the speaker.

In [1], the discriminant power of GSM AMR compressed parameters was studied and the parameters with the best per-

Speech Coder	Speech Spectra						Excitation		Gain		General	
GSM AMR 12.2 kb/s	LSFq1	LSFq2	LSFq3	LSFq4	LSFq5	-	Ac	AcR	G1	G2	-	-
GSM AMR 7.40 kb/s	LSFq1	LSFq2	LSFq3	-	-	-	Ac	AcR	G1	-	-	-
GSM AMR 6.70 kb/s	LSFq1	LSFq2	LSFq3	-	-	-	Ac	AcR	G1	-	-	-
GSM AMR 4.75 kb/s	LSFq1	LSFq2	LSFq3	-	-	-	Ac	AcR	G1	-	-	-
G.729 8.00 kb/s	LSPq1	LSPq2	LSPq3	-	-	-	Ac	AcR	G1	G2	-	-
G.723 6.3 kb/s	LSPq1	LSPq2	LSPq3	-	-	-	Ac	AcR	G1	-	-	-
G.723 5.3 kb/s	LSPq1	LSPq2	LSPq3	-	-	-	Ac	AcR	G1	-	-	-
MELP 2.4 kb/s	LSPq1	LSPq2	LSPq3	LSFq4	-	-	Ac	-	G1	G2	M	V
iLBC 15.20 kb/s	LSFq1-1	LSFq1-2	LSFq1-3	LSFq2-1	LSFq2-2	LSFq2-3	Cb1-3		Ga1-3		M	-
iLBC 13.33 kb/s	LSFq1-1	LSFq1-2	LSFq1-3	-	-	-	Cb1-3		Ga1-3		M	-

Table 1. Bitstream parameters for the various compressed speech formats: LSFq (n-th quantized parameter for line spectral frequencies), LSPq (n-th quantized parameter for line spectral pairs), Ac (adaptive codebook index), AcR (relative adaptive codebook index), G (gain value), Ga (n-th stage gain value), Cb (n-th stage codebook index), M (magnitude value for LPC residuals), V (voicing detector value).

formance were selected. These can be classified in three main groups: speech spectrum (i.e., line spectral frequencies), excitation (i.e., adaptive codebook index) and gain related features. The recognition algorithm was then based on the coefficient of variation (CoV) and skewness (SKEW) of all the selected parameters. The squared Euclidean distances from the test samples to each speaker reference model were used as the identification criterion. Results that appear to be promising at least for some applications were achieved with the following linear combination of COV (δ) and SKEW (ξ):

$$d(X, Y_i) = \alpha\, d(\delta_X, \delta_{Y_i}) + (1 - \alpha)\, d(\xi_X, \xi_{Y_i}), \qquad (1)$$

where $d(a, b)$ is the squared Euclidean distance between a and b, X is the test vector to be classified, Y_i is the model vector for speaker i, and α is an experimentally derived optimal weighting parameter ($\alpha = 0.48$). This metric achieved perfect recognition with at least twenty seconds of active speech in the limited case of a speech corpora of fourteen speakers recorded under normal room noise conditions.

4. ANALYSIS OF RECOGNITION ACCURACY IN THE COMPRESSED DOMAIN

Automatic speaker recognition in the compressed domain was previously applied to a specific speech encoding algorithm and bitstream format, i.e., GSM AMR at 12.2 kb/s. In this paper we validate the effectiveness of the compressed domain approach on different encoding algorithms by considering a wide variety of compressed speech formats and coding rates. In particular we analyze the performance of the recognition system on: GSM AMR [6] at 12.2 kb/s, 7.40 kb/s (IS-641 compatible), 6.70 kb/s (PDC-EFR compatible), and 4.75 kb/s; G.729 [7] at 8 Kb/s; G.723 [8] at 6.3 kb/s and 5.3 kb/s; iLBC [9] at 15.20 kb/s and 13.33 kb/s; MELP [10] at 2.4 kb/s.

If we consider the compressed parameters of each coding format, we can find a relation to the ones used for the GSM AMR case. Obviously the number of parameters and their bit size change from coder to coder. However, as explained in section 3, we can divide the whole set of compressed features in three main groups, with each group containing the parameters with the same physical meaning. Table 1 shows the parameters of each speech coder in the different groups (i.e., speech spectrum, excitation and gain related parameters). The number of compressed parameters is variable among different encoders (due to the dissimilar voice compression model employed), but it also varies among different encoding rates for the same speech format.

The recognition system is common to all the encoders and it is based, as for the GSM AMR, on CoV and Skewness of the compressed parameter values as encoded in the bitstream. Equation (1) is used to combine the relative distance between the test sample and the reference model for the two statistical measures. The speaker reference models are estimated using the statistic of speech parameters after 90 seconds of active speech. Table 2 shows a comparison of the results obtained by the speaker recognition algorithm for different speech formats. A database with a total of 14 speakers has been used to

Speech Coder	Length (s)			N
	10	20	30	
GSM AMR 12.2 kb/s	95.8%	100%	100%	9
GSM AMR 7.40 kb/s	85.2%	91.9%	95.3%	6
GSM AMR 6.70 kb/s	85.5%	93.8%	97.1%	6
GSM AMR 4.75 kb/s	84.9%	92.5%	96.2%	6
G.729 8.00 kb/s	77.0%	83.1%	87.7%	7
G.723 6.3 kb/s	76.4%	85.0%	90.6%	6
G.723 5.3 kb/s	75.5%	86.1%	87.6%	6
MELP 2.4 kb/s	86.1%	95.6%	97.2%	9
iLBC 15.20 kb/s	75.8%	82.5%	92.5%	10
iLBC 13.33 kb/s	77.9%	89.4%	95.3%	13

Table 2. Comparison of speaker identification rate with different length of the test samples for various encoders. The last column (N) reports the number of compressed parameters used for recognition in each speech format.

test speech segments of increasing length.

GSM AMR at 12.2 kb/s, that has been the subject of our studies so far, achieves a speaker recognition rate of 100% just after 20 seconds. That is because the system was previously optimized for this particular case. However, performance remains above 90% for all the GSM AMR rates we tested. The decrease in the recognition accuracy is mainly due to the decreased number of parameters available for the collection of speaker statistics. If we consider the complete set of compressed speech formats, we note a degradation of the recognition rate of nearly 18% in the worst case. In spite of that serious performance reduction we can however notice that the proposed technique remains somewhat valid for any coder. Results can be improved both with a wise choice of the best compressed features (and distance metrics) and with an increase of the length of the tested sequences just above 20 seconds.

An important result is that the recognition accuracy does not strictly depend on the coding rate used by the different encoders. As shown in Fig. 2, MELP coded speech at 2.4 kb/s presents an high recognition percentage of 97.2% after 30 seconds of active speech. This is almost the same result of GSM AMR at 6.7 kb/s and it is above the result of the iLBC at 13.33 and 15.20 kb/s. On the contrary, we should note that performance of speaker recognition using speech re-synthesized from GSM at 12.2 kb/s, G.729 at 8 kb/s and G.723.1 at 5.3 kb/s was shown to generally degrade with coder bit rate [5]. In a CD-ASR context performance of the system depends instead on the number of parameters used and on their size in bits. The proposed distance measure relies, in fact, on statistical features like CoV and skewness that are based on the probability density function of the parameters. As a consequence the discrimination power of Eq. (1) is clearly affected by the quantization resolution of the compressed speech features.

5. CONCLUSIONS

In this paper we presented an initial study on compressed-domain automatic speaker recognition that tries to extend previous results on GSM AMR at 12.2 kb/s to a selection of widely used speech coders: GSM AMR, G.729, G.723, MELP and iLBC. Although we applied this approach to dissimilar speech coding formats, promising results show that the proposed technique may achieve homogeneous results with different speech compression algorithms. With a database of 14 speakers, speaker recognition experiments show an accuracy close to 100% after the analysis of 30 seconds of active speech for most of the considered speech coders and rates. Of particular importance is the result that the recognition accuracy does not strictly depend on the quality of the speech encoders. In fact, the recognition performance does not always decrease with the coder bit rate: MELP coder at 2.4 kb/s and GSM at 4.75 kb/s achieve recognition rates comparable to the ones of toll quality coders.

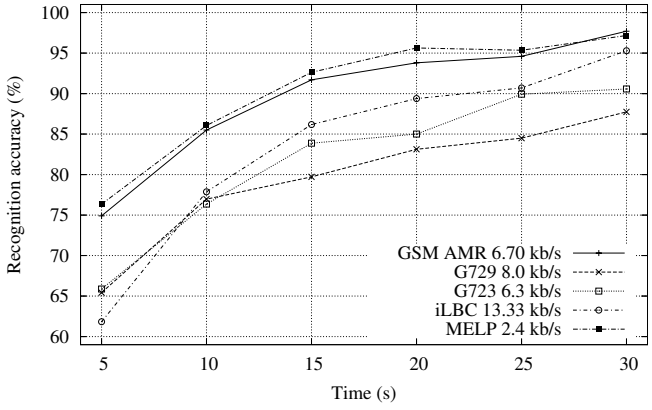

Fig. 2. Speaker recognition accuracy for various speech codecs as a function of the length of the test samples.

6. REFERENCES

[1] M. Petracca, A. Servetti, and J.C. De Martin, "Low-complexity automatic speaker recognition in the compressed GSM-AMR domain," in *Proc. IEEE Int. Conf. on Multimedia & Expo*, Amsterdam, The Netherlands, July 2005, pp. 662–665.

[2] C. Aggarwal, D. Olshefski, D. Saha, Z.-Y. Shae, and P. Yu, "CSR: Speaker recognition from compressed VoIP packet stream," in *Proc. IEEE Int. Conf. on Multimedia & Expo*, Amsterdam, The Netherlands, July 2005, pp. 970–973.

[3] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, January 1995.

[4] R.B. Dunn, T.F. Quatieri, D.A. Reynolds, and J.P. Campbell, "Speaker recognition from coded speech in matched and mismatched conditions," in *A Speaker Odyssey. The Speaker Recognition Workshop*, Crete, Greece, June 2001, pp. 72–83.

[5] T.F. Quatieri, R.B. Dunn, D.A. Reynolds, J.P. Campbell, and E. Singer, "Speaker recognition using G.729 speech codec parameters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Istanbul, Turkey, June 2000, vol. 2, pp. 1089–1092.

[6] ETSI EN 301 704 V7.2.0, "Digital cellular telecommunications system (phase 2+); adaptive multi-rate(AMR) speech transcoding," 1999.

[7] ITU-T Recommendation G729, "Coding of speech at 8 kbit/s using conjugate-structure algbraic-code-excited linear prediction (CS-ACELP)," *ITU-R*, 1996.

[8] ITU-T Recommendation G723.1, "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s," *ITU-R*, 1996.

[9] S. Andersen, A. Duric, H. Astrom, R. Hagen, W. Kleijn, and J. Linden, "Internet low bit rate codec (iLBC)," *RFC 3951*, December 2004.

[10] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, "MELP: the new federal standard at 2400 bps," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Munich, Germany, April 1997, vol. 2, pp. 1591–1594.