OPTIMAL SELECTION OF BITSTREAM FEATURES FOR COMPRESSED-DOMAIN AUTOMATIC SPEAKER RECOGNITION

Matteo Petracca, Antonio Servetti, Juan Carlos 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

Low-complexity compressed-domain automatic speaker recognition algorithms are directly applied to the coded speech bitstream to avoid the computational burden of decoding the parameters and resynthesizing the speech waveform. The objective of this paper is to further reduce the complexity of this approach by determining the smallest set of bitstream features that has the maximum effectiveness on recognition accuracy. For this purpose, recognition accuracy is evaluated with various sets of medium-term statistical features extracted from GSM AMR compressed speech coded at 12.2 kb/s. Over a database of 14 speakers the results show that, using 20 seconds of active speech, a recognition ratio of 100% can be achieved with only nine of the 18 statistical features under analysis. This is a complexity reduction by a factor of two with respect to previous works. Moreover, the robustness of the proposed system has been assessed using test samples of different length and varying levels of frame losses, and proved to be the same of previous approaches.

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[1]. 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 network 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. 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 packet streams of compressed voice packets. This method has been previously presented as *compressed-domain automatic speaker recognition* (CD-ASR) in [2] [3] where voice feature vectors are made up of compressed bitstream values from coded speech frames.

In [2] a tentative implementation limited to the GSM AMR Adaptive Multi-Rate (AMR) standard at 12.2 kb/s showed that, in some circumstances, speaker recognition in the compressed domain is possible after the analysis of about 20 seconds of active speech. The objective of this paper is to investigate if the complexity of that recognizer can be further reduced and with what impact on the accuracy of the results. For this purpose, for each compressed speech feature used in [2], we analyze its discriminant power as a single classifier, as well as its contribution to the overall recognition accuracy when used in conjunction with other features. Then, using a database of 14 speakers to test the feasibility of this approach, we order the features by their effectiveness and we identify the smallest set that achieves perfect recognition in this simple context. A recognizer with only nine out of the 18 features under analysis is proved to improve the recognition accuracy over previous approaches and to have the same robustness to packet losses.

The rest of this paper is organized as follows. An overview of automatic speaker recognition approaches is presented in Section 2. Besides traditional systems that use clean voice waveforms as input, we describe other approaches that work, at different levels, with coded speech. Compressed-domain automatic speaker recognition is then discussed in

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

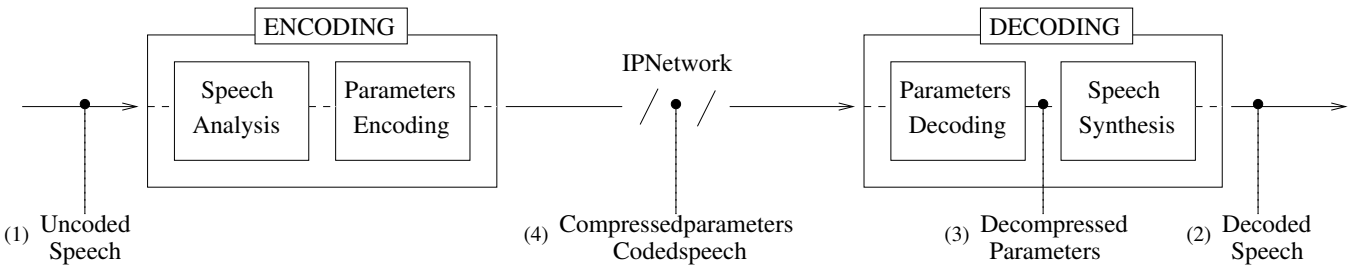

Fig. 1. In VoIP communications, the sender applies encoding standards to reduce the amount of information that is sent through the IP network. Hence, speech data traverses the network in a coded format and it has to be decoded and resynthesized at the receiver to obtain a voice signal similar to the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

Section 3 . In Section 4 we investigate the discriminant power of GSM AMR coded speech parameters for speaker recognition. The feature subset with the best tradeoff between complexity and recognition rate is then compared to previous results to evaluate its robustness to packet losses. Conclusions follow in Section 5.

2. OVERVIEW OF AUTOMATIC SPEAKER RECOGNITION APPROACHES

Figure 1 illustrates the encoding, transmission and decoding chain for VoIP communications. Within this context, the four 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 [4].

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) [5]. 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 [6]. 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 recently, a fourth approach, compressed-domain ASR, 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 [3] or medium-term statistical analysis [2]. One of the benefit from 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 mean 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, 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 builds on the results in [2]

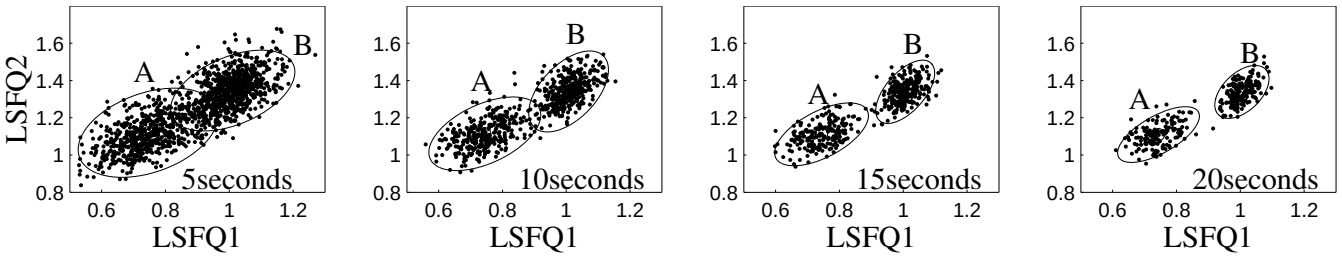

Fig. 2. Scatter plots for speakers A and B along two parameter dimensions (LSFQ1, LSFQ2) for different lengths of test samples (5, 10, 15, 20 seconds).

and regards the bitstream generated by the widely used GSM Adaptive Multi-Rate (AMR) speech coder at 12.2 kb/s, the default speech coder for GSM 2+ and WCDMA third generation wireless systems [7]. Although compressed speech parameters are non-linearly related to the more physically meaningful features, each compressed voice packet implicitly 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 [2], parameters that appeared to give the best recognition performance were selected. For each frame, a vector, consisting of the adaptive codebook indexes of the even (ACe) and odd (ACo) subframes, the adaptive (ACG) and fixed (FCG) codebook gains, and the five vector-quantized LSFs (LSFQ1-5), was derived. The recognition algorithm was then based on the computation of the coefficient of variation (CoV) and skewness (SKEW) on sequences of those parameters. For each sequence a feature vector of 18 elements was derived. Firstly, a reference feature vector (i.e., reference model) for each speaker was estimated from a ninety second reference voice stream. Then the same measures were performed on the test streams. The squared Euclidean distances from the test streams to each speaker reference model were used as the identification criterion. Results that appeared 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 happened to score a recognition ratio of 100% in initial tests with a small speech corpora of 14 speakers recorded in normal room noise conditions.

4. ANALYSIS OF RECOGNITION ACCURACY IN THE COMPRESSED DOMAIN

In the previous work we proposed to use a linear combination of skewness and coefficient of variation of the bitstream parameters as a tentative low-complexity approach for speaker recognition in the compressed domain. For the particular case of the GSM AMR speech coder at 12.2 kb/s and for a database of 14 speakers, this technique achieved good recognition accuracy after processing about 20 seconds of active speech. Experimental results showed that medium-term statistical features of compressed voice from different speakers appear to be separable. In Fig. 2, in fact, scatter plots of selected pairs of discriminant features for various lengths of the test samples illustrate a significant decrease in the dispersion of data as the sample length increases.

In this section, our objective is to analyze the recognition accuracy of each discriminant feature used in [2] in order to build a CD-ASR system of lower complexity from a subset of those features. Robustness of this new classifier is then assessed using test samples of different duration and simulating frame losses.

4.1. Single feature F-ratio

Intuitively, a good parameter for speaker recognition is one for which the individual speaker distributions are as narrow and as widely separated as possible. A statistic which has been found useful in quantifying this desired property is the F-ratio. The statistic is proportional to the ratio of the variance of the means of each speaker's feature distribution to the average value of the variance of each distribution. Given a total of K speakers, these expression can be mathematically defined by

$$F = \Big(\frac{1}{K}\sum_{j=1}^{K}(\mu_j - \mu)^2\Big) \Big/ \Big(\frac{1}{K}\sum_{j=1}^{K}\sigma_j\Big) \qquad (2)$$

where μ_j and σ_j are the mean and variance of the j^{th} speaker's feature distribution and μ is the overall mean of the feature distributions. Table 1 lists coefficient of variation (CoV) and skewness (SKEW) F-ratios of the coded parameters under investigation for 20-second long samples of active speech. The farther apart the individual distributions are with respect to their average spread, the higher the F-ratio. Although the F-ratio has been used as an indication of a feature's effectiveness, it is not optimal because a feature with a high ratio does not necessarily contribute more to the performance of a recognition system than a feature with a lower ratio.

	COV			SKEW		
Parameter	F-ratio	Recognition Accuracy	Effectiveness Order	F-ratio	Recognition Accuracy	Effectiveness Order
ACo	10.95	50.63%	#1	12.49	48.75%	#2
ACe	4.32	35.00%	#15	0.10	6.88%	#16
ACG	0.68	23.75%	#14	1.54	33.13%	#17
FCG	3.26	30.00%	#13	1.19	16.88%	#7
LSFQ1	0.37	32.50%	#12	1.40	30.62%	#4
LSFQ2	0.41	27.50%	#18	1.68	21.25%	#6
LSFQ3	0.65	18.75%	#5	0.10	12.50%	#10
LSFQ4	0.39	29.37%	#9	3.52	30.00%	#3
LSFQ5	2.81	29.37%	#11	5.15	28.13 %	#8

Table 1. F-ratio, recognition accuracy and effectiveness order of medium-term average COV and SKEW for GSM AMR codec parameters. Values refer to 20-second long test samples.

4.2. Single feature recognition performance

A better criterion for feature selection is based upon the obvious fact that the goal of a speaker recognition system is to classify an unknown speaker correctly. This goal implies that the relative merit of a feature should be based upon its contribution to the performance of recognition. In practical terms, if a feature, G, yields a smaller rate of error than another feature, then G may be a better feature for recognizing speakers. Given that the ultimate utility of a feature really depends upon the nature of the classification system that follows it, we evaluate the relative merit of a feature with respect to the previous defined distance measure, as in Eq. (1). Table 1 reports the recognition accuracy of COV and SKEW for each compressed speech parameter over a set of 160 twenty-second long test speech samples. A match occurs if a test vector is labeled to the right speaker, i.e., the intra-speaker distance is smaller than all the inter-speaker ones. Accuracy is then obtained by evaluating the percentage of matches. We note that the recognition ratio is generally higher for the coefficient of variation than for the skewness even if this last one showed an overall higher F-ratio. This result is clearly related to the fact that the skewness takes about 90 seconds to converge to its long-term average, as shown in [2], so the accuracy that can achieve highly depends on the length of the test speech samples (in this case only 20 seconds).

Given the recognition accuracy of each single feature we can try to incrementally build a recognition system with a slightly increased complexity and accuracy at each step. For example, joining the first best feature with the second one (i.e., COV ACo with SKEW ACo) we improve the recognition rate from 50.63% to 75.63%. Clearly, successive steps will present a reduced gain due to the relative dependency between the features in the ability to discriminate among speakers. A good tradeoff, but not necessarily the best, is the result achieved with the top nine features (in the following referred as TOP-9). This system requires half the complexity of the one using all the 18 features (REF-18) and enables a recognition accuracy of 96,25% instead of 100%.

Length (s)	**REF-18**	**TOP-9**	**BEST-9**
5	71.74%	80.54%	83.68%
10	87.72%	92.12%	92.72%
15	96.05%	95.39%	98.16%
20	100%	96.25%	100%

Table 2. Comparison of speaker identification rate for three recognition systems varying the length of test samples.

4.3. Feature-subset recognition performance

However, the approaches so far do not guarantee the optimality of the recognizer for the given data set. In fact, since the features are not statistically independent, there could be other combination of features that work better, i.e. two features that are not the best individually can give the best performance in combination since they carry complementary information. Hence, we employed an experimental technique for ordering the effectiveness of each feature when used in conjunction with other features under the assumption that the *relative effectiveness* of a set of features may be defined as inversely proportional to the error performance of a classifier that uses that feature set. Starting from a total number of features that is equal to N, the method begins by evaluating the recognition ratio of each of the N feature subsets with $N-1$ members. The most effective feature subset is then determined, and the feature not included in this subset is defined as the least important feature. This feature is then eliminated from further consideration. The procedure continues until all the features are eliminated from consideration. The ordered effectiveness of the features is then given by the inverse sequence of the eliminated features.

It is important to keep in mind that the ordering is established in accordance with the measurements of a given, small, set of recordings and that the order may slightly vary for a different set. Nevertheless, the ranking shown in Table 1 affords a general idea of what GSM AMR compressed features are important in recognizing an unknown speaker These important features include: a) the adaptive codebook index, only

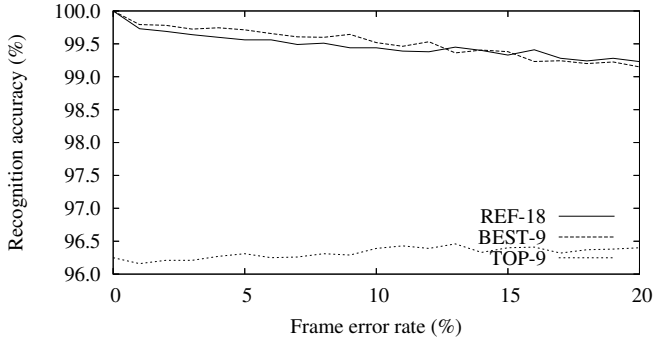

Fig. 3. Speaker recognition accuracy as a function of the frame loss rate for the three recognition systems under analysis. Speech samples of 20 seconds are considered.

the value of the odd subframes appears to be significant because the value of the even subframes is differentially coded with respect to the preceding odd subframe; b) the formant frequencies, that retain their discriminant power even if vector quantized.

On the other side, marginal or null contribution is provided by the gain related features. Only FCG is of some relevance because it is part of the nine best features.

With a database of 14 speakers and test speech samples of 20 seconds, our experiments show that, of the 18 features under analysis, only nine are sufficient to obtain a recognition ration of 100%. This new system, BEST-9, is based on the coefficient of variation of LSFQ3, LSFQ4, ACo, and on the skewness of LSFQ1, LSFQ2, LSFQ4, LSFQ5, ACo, FCG. Table 2 shows its accuracy compared to the other classifiers for different lengths of the test samples. We note that a more accurate choice of the compressed speech features used in the recognition system can also improve the accuracy with short test samples.

4.4. Robustness experiments

This section considers the adverse effects on speaker recognition accuracy caused by packet losses present in fixed and mobile IP communications. Some degree of packet loss is inherent in VoIP communications where lost packets might be caused by congestion in Internet routers or errors in the communication channel. Packet dropping has a great impact on the decoder ability to reconstruct the voice signal because compression of a speech frame is strongly based on its correlation to the preceding and successive frames. When these data are not available, decoding results in a poor voice waveform, not of sufficient quality for accurate voice recognition analysis in the decoded signal domain. Thus, we expect that CD-ASR is particularly robust against packet losses because, with respect to other techniques that use the speech waveform, it does not need to decode the speech signal.

The effect of the degradation caused by unreliable packet transmission is assessed on the three recognition systems: REF-18, BEST-9, TOP-9. To simulate packet losses, a varying percentage of speech frames are discarded from each test sample with uniform probability. No attempt is made to recover the corresponding lost features. In the results, as presented in Fig. 3, the accuracy curves confirm that the system is highly resilient with minimal decrease in the recognition accuracy over the range of possible frame loss rates. Moreover, the choice of the BEST-9 recognizer, while reducing the complexity of the REF-18 algorithm by one half, does not reduce the robustness to transmission errors.

5. CONCLUSIONS

In this paper we presented a study on the identification of the most effective features from compressed speech for low-complexity automatic speaker recognition. For each feature, extracted from the GSM AMR bitstream at 12.2 kb/s, we analyzed the discriminant power as a single classifier, as well as the contribution to the overall recognition accuracy when combined with other features. The selection of the most effective features allowed the construction of a recognizer with only half of the features used in previous works and with increased accuracy. This system has also been shown robust to packet losses in IP networks with a degradation in the recognition rate of less than 1% for a maximum frame error rate of 20%. Further investigations are in progress to validate this approach for different speech coders and a large number of speakers.

6. REFERENCES

[1] B. Goode, "Voice over internet protocol (VoIP)," *Proceedings of the IEEE*, vol. 90, no. 9, pp. 1495–1517, September 2002.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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