LOW-COMPLEXITY AUTOMATIC SPEAKER RECOGNITION IN THE COMPRESSED GSM AMR DOMAIN

M. Petracca, A. Servetti, J.C. De Martin[1]

Dipartimento di Automatica e Informatica / IEIIT-CNR[1]
Politecnico di Torino
Corso Duca degli Abruzzi, 24 — I-10129 Torino, Italy
E-mail: [matteo.petracca|servetti|demartin]@polito.it

ABSTRACT

This paper presents an experimental implementation of a low-complexity speaker recognition algorithm working in the compressed speech domain. The goal is to perform speaker modeling and identication without decoding the speech bitstream to extract speaker dependent features, thus saving important system resources, for instance, in mobile devices. The compressed bitstream values of the widely used GSM AMR speech coding standard are studied to identify statistics enabling fair recognition after a few seconds of speech. Using euclidean distance measures on elementary statistical values such as coefficient of variation and skewness of nine standard GSM AMR parameters delivers recognition accuracies close to 100% after about 20 seconds of active speech for a database of 14 speakers recorded in a normal room environment.

1. INTRODUCTION

Automatic speaker recognition (ASR) has been a research topic for many years, during which various types of approaches, with increasing level of complexity and performance, have been studied. The general method to ASR consists of three steps: speech data acquisition, feature extraction, and pattern matching. Feature extraction maps speech intervals to a multidimensional feature space. Speaker identification is, then, performed comparing this sequence of feature vectors to available speaker models by pattern matching. State of the art speaker recognition systems typically use mel-frequency cepstral coefficients (MFCC) and Gaussian mixture models (GMM) for speaker modeling [1].

In recent years, due to the widespread use of digital speech communication systems, there has been increasing interest in the performance of recognition systems from coded speech. The effect of speech coding on speaker and language recognition tasks has been investigated for several coders and a wide range of bit rates [2]. The typical approach consists of extraction of the speech features from the decoded speech signal. This paper, instead, explores the possibility of working directly in the speech domain so that no decoding is needed, thus lowering the processing and memory requirements with respect to the standard approach.

We investigate the recognition accuracy achievable using medium-term statistical analysis of the coded bitstream to produce a feature set and a speaker model useful for speaker recognition. We focus on limiting the complexity of our model to the second and third order statistic of a few parameters, thus requiring just a fraction of the memory storage and processing power needed by systems based on GMMs. The proposed system is therefore targeted at applications that are allowed to identify speakers after a few seconds of active speech.

The structure of this paper is as follows. In Section 2 we investigate the statistical properties of coded speech parameters as speaker-dependent features. The recognition experiments and performance results are presented in Section 3 for a speech corpora with fourteen male and female speakers. Conclusions follow in Section 4.

2. SPEAKER-DEPENDENT INFORMATION IN THE GSM BITSTREAM

When a person speaks, he or she produces a set of signal features that characterize both the identity of the utterance as well as that of the speaker. In the literature there have been several studies on the choice of acoustic features in speaker recognition tasks [3]. Average fundamental frequency has been found to be a useful discriminating feature, as have gain measurements and long-term speech spectra. All these features are physically-based distinguishing characteristics related to the human speech production system.

In the approach under investigation, the feature space

The work was supported in part by the Motorola PCS Research Center in Torino, Italy.

0-7803-9332-5/05/$20.00 ©2005 IEEE

is instead derived from bitstream values of coded speech parameters. In this particular case our study regards the bitstream generated by the widely used GSM AMR speech coder. Table 1 shows, the bit allocation for each quantized speech parameter at 12.2 kb/s (20 ms speech frame) [4]. Although these values are non-linearly related to the more physically meaningful features, we intend to investigate if they can provide valuable speaker-dependent information. If proved true, then we can design a low complexity recognition system that saves time and power with respect to traditional systems based on parameters decoding or speech re-synthesis.

2.1. Speech features in the compressed domain

The essential characteristics of the speech process change relatively slowly; speech signals are, therefore, usually parametrized over relatively long time periods of 10 to 25 ms called frames. Feature vectors produced by individual speakers are assumed to be samples from a continuous density probability distribution. The distributions of different speakers overlap and share the speaker space, but are ideally distinguishable from each other so that speaker identification can be achieved. For instance, if we consider the fundamental frequency, although it is a controllable attribute of stress and intonation which may vary widely, each person appears to have a mean fundamental frequency value which is relatively constant over a reasonable time span independently of linguistic content [3].

Our assumption is that, over a sufficiently long interval of speech, also the distributions of bitstream values conserve the desirable property of being distinguishable among different speakers. Statistical analysis on GSM AMR coded parameters shows, indeed, that some of them can really be considered part of the individual speaker's characteristics. The *adaptive codebook index*, for example, is an almost linear quantization of the fundamental frequency, and it presents nearly the same distribution.

Moreover, from our preliminary analysis, inter-speaker variability is also evident in the distribution of the *adaptive codebook gain*, the *fixed codebook gain*, and the *LP coefficients* of the prediction filter that models the vocal tract.

Parameter	**Subframe**			
	1	**2**	**3**	**4**
Line Spectral Frequencies (LSFQ 1-5)	7 + 8 + 9 + 8 + 6			
Adaptive Codebook index	9	6	9	6
Adaptive Codebook gain	4	4	4	4
Fixed Codebook index	35	35	35	35
Fixed Codebook gain	5	5	5	5

Table 1. Bit allocation of the GSM AMR 12.2 kb/s speech coding standard.

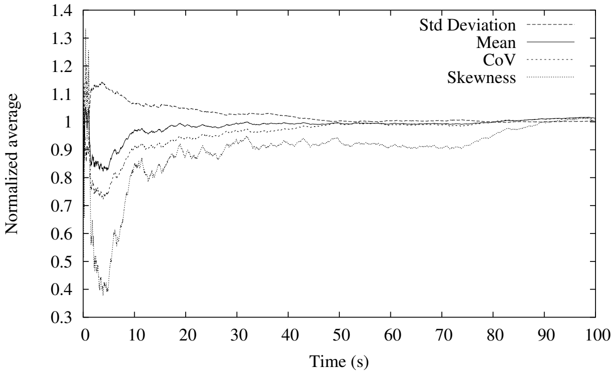

Fig. 1. Long-term average convergence of the adaptive codebook index mean, standard deviation, CoV, and skewness as a function of time.

This last result is particularly interesting: the split vector quantized LSF indexes, in fact, were not expected to preserve a great deal of the physical meaning attached to the corresponding LP coefficients. No discrimination, instead, arose from the fixed codebook excitation, that, as expected, was almost uniformly distributed and statistically identical between different speakers.

The group of nine input parameters which appeared to give the best recognition performance was selected. For each frame, a vector, consisting of two adaptive codebook indexes, the adaptive and fixed codebook gains, and the five vector-quantized LSFs, was derived.

2.2. Speaker characterization

In order to characterize the probability density distribution of speaker discriminant features – in this case the raw bitstream values of the speech coder parameters – we need to chose significant figures of merit. Mean, standard deviation, coefficient of variation [1] (CoV) and skewness are some of the possible choices limiting the search to elementary statistical functions. For example, the standard deviation of a speaker's fundamental frequency could be expected to be relatively small for a "monotone" speaker as compared to an "expressive" speaker.

To effectively employ such measures for speaker identification we need to know if early estimation of these values over a short period of time (a few seconds) can be accurate. Therefore we need to study the relation between these measures over an initial part of a speaker's sentence and the complete talk. For this purpose, in Fig. 1 we represent the convergence of *adaptive codebook index* mean, standard deviation, CoV, and skewness to their long-term average (one hour of active speech).

[1] Coefficient of variation is here defined as standard deviation over mean.

We can clearly see that the values converge to one after 50 seconds except for the skewness that takes about 90 seconds. This is important, because it means that after 90-100 seconds further analysis is not necessary to identify the "true" speaker characteristics (needed for building identification models), and that after the first 20-40 seconds possibly enough information can be gathered to discriminate between different speakers.

2.3. Speaker discrimination

An estimate of the usefulness of a parameter for speaker discrimination is commonly derived from the ratio of its between-speaker to within-speaker variance, also known as *F-ratio*. In an ASR system, promising parameters are, in fact, those that show an high degree of variability between various speakers, but that concentrate their values in a short range for the same speaker over different time spans. The F-ratio of a feature is computed as the ratio between the variance of its inter-speaker means and the average of its intra-speaker variance.

Table 2 lists CoV (that includes in its definition mean and standard deviation) and skewness F-ratios of the coded parameters under investigation. The ordering of the parameters is of some interest. These values can, in fact, be used as an indicator of the relative effectiveness of each feature. For example, we see that CoV of LSFQ3 provides very little discrimination among speakers, whereas skewness of adaptive codebook indexes appears to provide the maximum discrimination among speakers over all parameters.

A second way of quantitatively showing the parameters effectiveness in speaker identification is to inspect two-dimensional scatter plots of selected pairs of input parameters for various lengths of the test samples. Figure 2 presents a scatter plot of LSFQ1 CoV versus LSFQ2 CoV for speakers A and B, also shown are two-sigma ellipses for each speaker distribution. A significant decrease in the dispersion of the data is seen as the sample length increases. The two speakers are partially overlapped for limited amounts of sample duration (i.e., five and ten seconds), but they are clearly separable for longer sample durations.

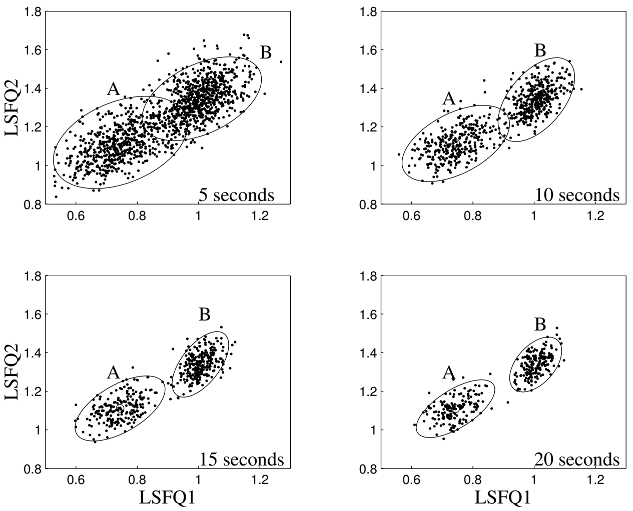

Fig. 2. Scatter plots for speakers A and B along two parameter dimensions (LSFQ1, LSFQ2) for different lengths of test samples (5, 10, 15, 20 seconds).

	10 seconds		20 seconds	
	CoV	Skewness	CoV	Skewness
Adaptive cbk idx 1-3	1.24	35.21	1.48	45.39
Adaptive cbk idx 2-4	1.38	22.65	1.73	32.92
Adaptive cbk gain	1.32	3.25	1.55	4.74
Fixed cbk gain	2.47	17.86	3.67	21.70
LSFQ 1	2.62	7.73	3.34	10.24
LSFQ 2	1.14	0.50	1.50	0.66
LSFQ 3	0.06	3.82	0.08	5.98
LSFQ 4	2.11	4.14	2.70	5.12
LSFQ 5	1.23	1.54	1.44	2.11

Table 2. F-ratios of long term average CoV and skewness for 10 and 20 seconds of active speech.

Thus these results suggest that a speaker recognition application can be implemented using simple statistics on coded speech parameters together with a basic Euclidean distance metric to discriminate between different speakers. The next section will cover in detail our experiments and the corresponding preliminary results.

3. SPEAKER RECOGNITION RESULTS

Speaker recognition tests have been conducted to assess the feasibility of such a system based on the raw bitstream values of the GSM AMR parameters. Speaker recognition performance was verified using a speech corpora recorded under normal room noise conditions. We collected recordings from fourteen speakers, seven men and seven women. Two separate data sets were made from the recordings. The first set, consisting of ninety seconds of active speech, was used to train the system. As argued in Sec. 2.2, in fact, after this period the time average of most features already corresponds to the "true" long-term average. The second data set contained unknown samples of different lengths to be used to test recognition. In this set there were a total of 150 seconds for each speaker, successively divided in thirty, fifteen, ten, and seven sequences of five, ten, fifteen, and twenty seconds, respectively.

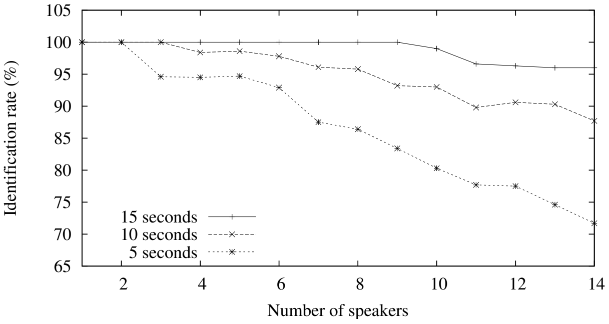

Fig. 3. Speaker identification rate as a function of the number of speakers for different lengths of the test samples.

3.1. Experiments

In the following experiments, recognition was based on the CoV and skewness of the nine bitstream values in Table 2. A reference value for each speaker was firstly estimated from the ninety second reference set. Then the same measures were performed on the test sets. The squared Euclidean distances from the test samples to each reference value were computed, and the test sample was assigned to the reference which yielded the smallest distance. For evaluating the speaker identification rate, the number of correct choices over the total number of test sequences was recorded.

After trying several distance measures, good recognition rates were achieved with the following linear combination of CoV (δ) and skewness (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 vector to be classified, Y_i is the vector for speaker i, and α is an experimentally derived optimal weighting parameter ($\alpha = 0.48$). Identification success rates for different lengths of the test sequences are given in Table 3. This metric achieves perfect recognition with at least twenty seconds of active speech and it does not degrade too much even with only five seconds of test material.

Finally, in Fig. 3, we provide a plot of the recognition rate as a function of the number of speakers included in the data set. Scores for different lengths of the test sequences are presented. We do not represent the case of 20-second samples that was already shown achieving zero percent error rate. Clearly the percentage of correctly identified speakers is directly proportional to the sample length and inversely proportional to the speaker population. However longer sample lengths show a reduced recognition degradation with an increasing number of speakers.

Length (s)	Identification rate (%)
5	71.7%
10	87.7%
15	96.0%
20	100%

Table 3. Speaker identification rate using Equation (1) for different lengths of the test samples, for a database of 14 speakers.

Because of the relatively small size of the recorded database we could not test our recognition system with a larger set of speakers, but, in this limited population size, statistics on the coded speech parameters demonstrated a fairly good performance. Future work will assess different environmental conditions, more representative speech corpora, and other speech coders.

4. CONCLUSIONS

An experimental implementation of a low complexity speaker recognition system working in the compressed speech domain has been investigated. Bitstream values of the GSM AMR speech coder parameters are studied to identify sufficient statistic that enables fair recognition after few seconds of speech. Even if coded, linear prediction coefficients, gains, and pitch delay still show some inter-speaker variability. Euclidean distance measures on elementary statistic variables such as coefficient of variation and skewness obtain recognition accuracy of 100% after 20 seconds of active speech for a database of 14 speakers recorded in a normal room environment.

5. REFERENCES

[1] J. P. Campbell, “Speaker recognition: A tutorial,” *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, September 1997.

[2] 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.

[3] J. D. Markel and S. B. Davis, “Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 1, pp. 74–82, February 1979.

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