Voiced/unvoiced classification for short time frames and its application to frequency-domain cryptanalysis

Angelo M. C. R. Borzino,[†] Dirceu G. da Silva,[†‡] and José A. Apolinário Jr.,[†]

[†]**IME – SE/3**
Praça General Tibúrcio, 80 – Urca
22290-270 Rio de Janeiro, RJ – Brazil
aborzino@uol.com.br, apolin@ieee.org

[‡]**PUC-Rio – CETUC**
Rua Marquês de São Vicente, 225 – Gávea
22453-900 Rio de Janeiro, RJ – Brazil
dirceu@ime.eb.br

***Abstract*— This paper introduces a novel voiced/unvoiced classification of speech signal for short time frames, sica 8ms, which do not encompass more than one pitch period. The classification scheme takes into account not only usual features related to the speech frame like its energy and its normalized power spectrum density but also statistical features like the median and the difference between the maximum and the minimum values of the latter. With only the three last mentioned features, as will be seen, this classification can be used in the cryptanalysis of frequency-domain ciphered speech. The voiced/unvoiced decision in this case is needed to improve classical cryptanalysis results by adding the concept of two codebooks: one for voiced ciphered speech and another one for unvoiced ciphered speech.**

***Index Terms*— Speech processing, speech scrambler, cryptanalysis of ciphered speech.**

I. Introduction

In many speech analysis systems, there is a need to decide whether a given segment of speech should be classified as voiced or unvoiced. In the technical literature, it can be found a number of methods used to make this decision [1]. Most of them use frames with a duration encompassing more than one pitch period and, therefore, relying on this feature for the classification. Few articles have addressed short duration speech frames (about 10ms or less). In this work, we tackle the problem of voiced/unvoiced classification of speech frames as short as 8ms.

Another goal of this research is to find features which are robust to frequency scramblers, i.e., they do not change after a speech signal is ciphered. Most features addressed in the literature do not apply for this kind of scrambler; this is so because they change when we frequency sub-bands are permuted (which is the basic procedure used by frequency-domain scramblers, as seen in Subsection IV-A).

In [2], a voiced/unvoiced/silence classification technique for short time frames (10ms) is introduced. However, four out of the five features used in this method do not apply when speech signals are ciphered in the frequency-domain. This happens because zero-crossing rate, autocorrelation coefficient, and prediction error change when a speech signal is scrambled.

The authors thank FAPERJ and CNPq for partial funding of this paper.

This article introduces a voiced/unvoiced classification for short duration speech frames, around 8ms, which roughly corresponds to the average pitch period for male speakers. Yet, with only 3 features, it is possible to apply this same classification in the problem of cryptanalysis of frequency-domain ciphered speech.

The classifier used here is the Gaussian Mixture Model (GMM) [3], which was applied successfully in speaker recognition [4]. GMM can be seen as a hybrid between two effective models: a unimodal Gaussian classifier and a vector quantization (VQ) [5] codebook. This scheme combines the robustness and smoothing properties of the parametric Gaussian model with the arbitrary modeling capability of a non-parametric VQ. The GMM performs the spatial separation of voiced/unvoiced classification and its main difference comparing to VQ concerns the fact that distances are not used to separate the classification but probabilities from a set of Gaussian probability density functions previously estimated. The GMM can also be understood as a single state HMM (Hidden Markov Model) [6], having as observations mixtures of Gaussian PDFs (probability density functions). These components may model the two classifications: voiced or unvoiced. This fact justifies its use in the decision whether a frame is voiced or not.

This paper is organized as follows. Section II describes the proposed voiced/unvoiced classification technique while Section III shows its simulations results. Section IV details the application: cryptanalysis of frequency-domain ciphered speech. Finally, Section V concludes this work.

II. Voiced/unvoiced classification of short frames

In this section, the proposed voiced/unvoiced technique is described. Subsection II-A details the features and Subsection II-B explains the classifier.

A. Selected Features

The choice of the features must be carried out such that they vary consistently from one class to another (voiced to unvoiced). We will show that the following features, whether

considered together or only a part, are able to classify well 8ms frames as voiced or unvoiced:

- Log-energy (LE).
- 23 normalized power spectral density ($NPSD$) coefficients.
- $Median$ of the 23 $NPSD$ coefficients.
- Difference between the maximum and minimum values of the 23 $NPSD$ coefficients (Dif).

In order to explain how to obtain these features, let $\boldsymbol{x}_i$ be the vector with the M samples of the $i-th$ speech frame (note that $M = 64$, when the sampling frequency is $8000Hz$, corresponds to an 8ms frame) and $\boldsymbol{X}_i$ be a vector with its 64 points DFT coefficients $x_{ij} = [\boldsymbol{X}_i]_j, j = 1, 2, \cdots, 64$. Due to the particular application at hand and the fact that radio and telephone channels usually destroy information below 300Hz and above 3200Hz, these side bands were eliminated (the corresponding DFT coefficients were not considered) in our investigation. With this observation and the fact that it is necessary to keep the symmetry of the DFT, so that the signal remains real, it is sufficient to work with the coefficients 4 to 26 (total of 23).

The features are obtained as follows:

(1) Log-energy (LE)

$$LE = 10log(\sum_{i=4}^{26} |\boldsymbol{X_i}|) \quad (1)$$

(2) 23 normalized power spectral density ($NPSD$) coefficients

$$NPSD_j = 10log(|[\boldsymbol{X_i}]_j|^2/\|\boldsymbol{X_i}\|),\ j = 4,\ 5,\ ...\ 26. \quad (2)$$

(3) Median of the 23 $NPSD$ coefficients ($Median$) Sorting these coefficients in an ascending order such that $\boldsymbol{z}_1$ is the lowest and $\boldsymbol{z}_{23}$ is the highest, then

$$Median = \boldsymbol{z}_{12} \quad (3)$$

(4) Difference between the maximum and minimum values of the 23 $NPSD$ coefficients, i.e., for the $4-th$ to $26-th$ components (Dif)

$$Dif = max(NPSD) - min(NPSD) \quad (4)$$

Fig. 1 depicts the typical behavior of the normalized power spectral density (in dB) of short frames for a voiced speech frame and for an unvoiced speech frame. From this figure, we can observe that:

- The difference between the maximum and the minimum values is expected to be larger for voiced frames.
- The median is expected to be lower for voiced frames.
- Voiced frames usually have larger energy in lower frequency bands.

From these observations, it is possible to develop a GMM-based voiced/unvoiced classification, as will be seen in the next subsection. The motivation for this approach comes from the results shown in Fig. 1, Fig. 2, and Fig. 3, where the classification capabilities of the before mentioned features can be observed.

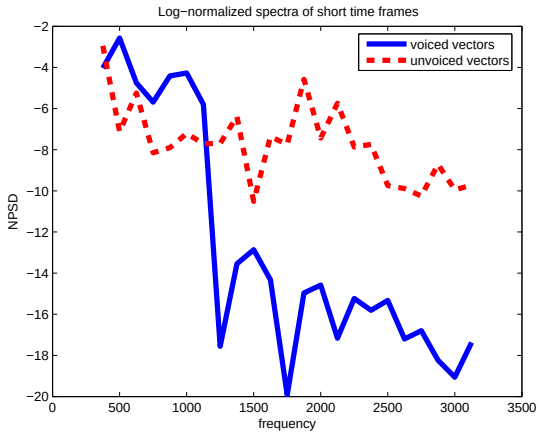

Fig. 1. Log-normalized spectra of short time frames.

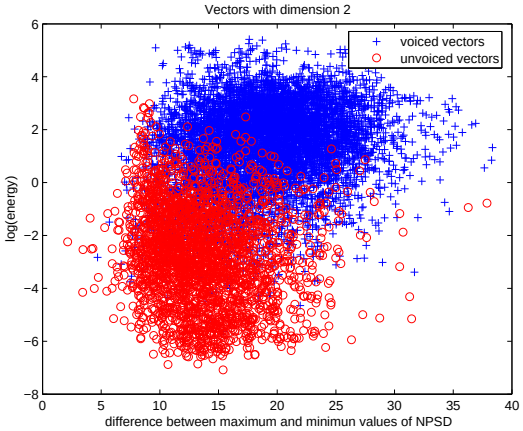

Fig. 2. Difference between maximum and minimum values of $NPSD$ versus log-energy (LE).

B. Voiced/unvoiced classifier

As mentioned before, the classifier used in this work is the GMM which is explained in detail as follows.

A mixture of Gaussian probability densities is a weighted sum of M densities, and is given by

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (5)$$

where $\vec{x}$ is a random vector of dimension D, $b_i(\vec{x})$, $i = 1, ..., M$, are the density components, and p_i, $i = 1, ..., M$, are the mixtures weights. Each component density is a D variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{e^{\left(-\frac{1}{2}(\vec{x}-\vec{\mu})' K_i^{-1} (\vec{x}-\vec{\mu})\right)}}{(2\pi)^{\frac{D}{2}} \sqrt{|\boldsymbol{K}_i|}} \quad (6)$$

with mean vector $\vec{\mu}_i$ and covariance matrix $\boldsymbol{K}_i$.

Note that the weighting of the mixtures satisfies $\Sigma_{i=1}^{M} p_i = 1$. The complete Gaussian mixture density is parameterized by a vector of means, covariance matrix, and a weighted mixture

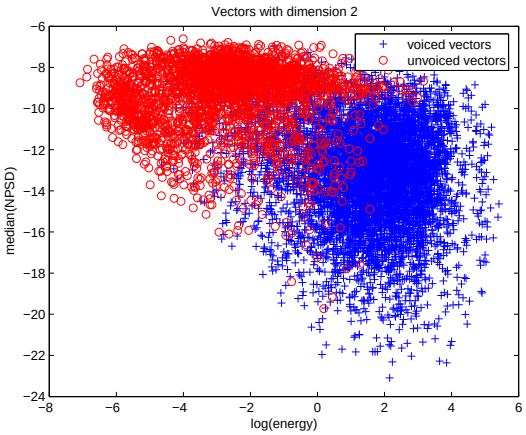

Fig. 3. Log-energy (LE) versus $Median$.

of all component densities (λ model). These parameters are jointly represented by the following notation:

$$\lambda = \{p_i, \vec{\mu}_i, K_i\} \qquad i = 1, ..., M. \tag{7}$$

The GMM can have different forms depending on the choice of the covariance matrix. The model can have a covariance matrix per Gaussian component as indicated in (7) (nodal covariance), a covariance matrix for all Gaussian components for a given model (grand covariance), or only one covariance matrix shared by all models (global covariance). A covariance matrix can also be complete or diagonal [3].

For a set of training data, the estimation of the maximum likelihood is necessary. In other words, this estimation tries to find the model parameters that maximize the likelihood of the GMM and may be obtained recurrently, using the *Expectation Maximization* (EM) algorithm.

The voiced/unvoiced classification scheme comprises the following steps:

- Manual classification (voiced/unvoiced) of a number of 8ms phonetically balanced speech frames, for training.
- Feature extraction of each frame to be used by a Gaussian Mixture Model (GMM) in order to produce voiced and unvoiced models: in our case, we have used 23 $NPSD$ coefficients, their log-energy (LE) without normalization, the median of the $NPSD$ ($Median$) and the difference between their maximum and minimum values (Dif), i.e, a vector with a total of 26 elements.
- Model generation (GMM with different number of Gaussians) for voice and unvoiced frames.
- Feature extraction of test data (in our experiment, 8ms speech frames from 10 phonetically balanced phrases were used) forming vectors of 26 elements.
- Validation of the test data by comparing the vectors from the last step with those models trained with the GMM, as illustrated in Fig. 4.

Overall, we have used 6528 voiced vectors with 2566 unvoiced vectors, for training, and 904 voiced vectors with 252 unvoiced vectors for testing.

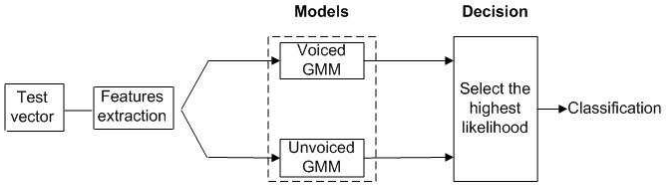

Fig. 4. Voiced/unvoiced identification using GMM.

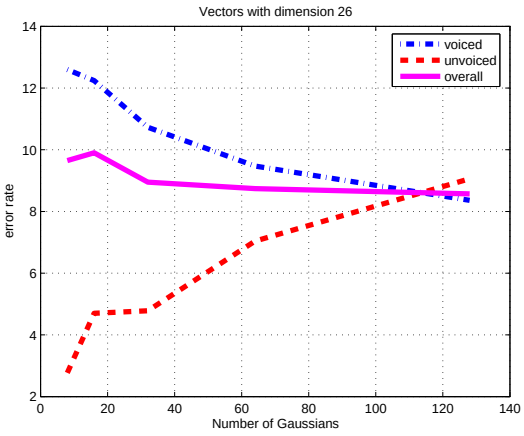

Fig. 5. Error rate (in %) of voiced/unvoiced decision for feature vectors with dimension 26.

III. Simulation Results

Fig. 5 depicts the error rates of the validation test when vectors of dimension 26 are used, for different numbers of Gaussians.

From Fig. 5, it can be observed that it is possible to decrease the overall error rate if we use different numbers of Gaussians to model voiced and unvoiced vectors; the figure suggests that 8 Gaussians are enough to model unvoiced frames while 128 (or even more) are necessary to model voiced frames. With this simple approach, the overall error rate would drop according to the voiced/unvoiced frame rates found in speech.

Fig. 6 shows the error rates, when vectors of dimension 3 are used—the log-energy (LE), the $Median$ (of the $NPSD$), and the difference between its maximum and minimum values (Dif).

Comparing Figures 5 and 6, we note that with only the 3 features used in Fig. 6 we are are able to classify a speech frame with nearly the same performance as of the 26 features used in Fig. 5. This fact can be explored in frequency domain cryptanalysis since these three features do not change when a speech signal is ciphered by a frequency-domain scrambler.

The voiced/unvoiced decision can be used to implement two codebooks instead of the single codebook approach used by [8] in an attempt to improve the performance of the frequency-domain cryptanalysis.

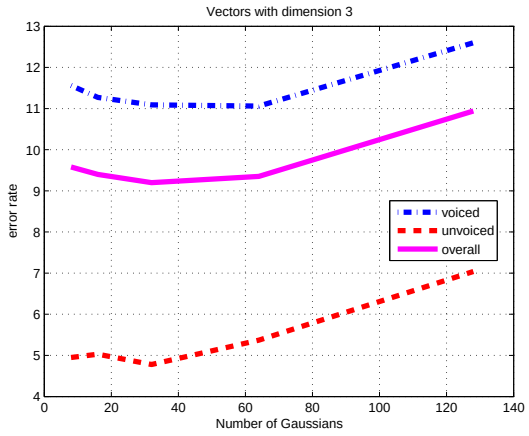

Fig. 6. Error rates (in %) for feature vectors with dimension 3.

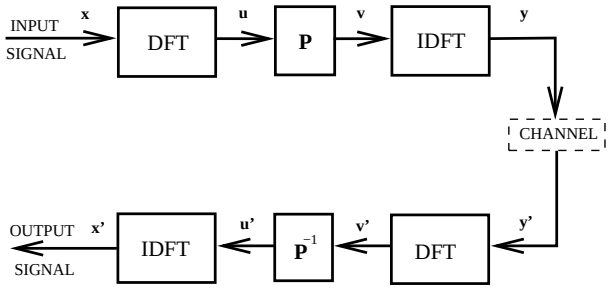

Fig. 7. *Scrambler* DFT.

IV. Cryptanalysis of Frequency-domain Ciphered Speech

This section presents an application where the classifier proposed in Section II was successfully used: the frequency-domain cryptanalysis. In Subsection IV-A, the frequency scrambler is briefly explained and, in Subsection IV-B, the cryptanalysis scheme is detailed.

A. A simple example of a frequency-domain scrambler

Fig. 7 depicts an example of a frequency-domain scrambler based on the Discrete Fourier Transform DFT). Note that the upper half of this figure comprises the transmitter where a block of M samples of the original speech is processed with or without overlap. The DFT of each frame (or block of samples) in then computed and M coefficients in the transformed domain are obtained. Following, these coefficients are permuted by permutation matrix $\boldsymbol{P}$ and the inverse transform is applied to return the signal to the time domain, allowing its transmission over a channel to the receiver.

The lower part of Fig. 7 depicts the reception process. The *de-ciphering* is carried out in a similar way but using the key or the inverse permutation ($\boldsymbol{P}^{-1}$) to recover the clear signal. Signal y (or y') in Fig. 7 would be used in the cryptanalysis.

B. Frequency-domain cryptanalysis

In order to perform the cryptanalysis, we implement a vector quantization (VQ) in the ciphered domain. Therefore, a codebook is to be designed to accomplish this task. Moreover, the speech signal used for training must have similar distortions and environmental noise as those of the signal we wish to intercept and recover intelligibility through cryptanalysis.

Fig. 6 has shown that the voiced/unvoiced error rate is relatively low even when considering only 3 features (log-energy, the median of the NPSD, and the difference between its maximum and the minimum values). These features are not altered when the signal is ciphered with a frequency-domain scrambler. Therefore, it is possible to classify each block using two distinct codebooks: one for voiced frames and a second one for unvoiced frames.

In this work, we have assumed that the number of samples M of each ciphered frame is known (we have used $M = 64$ in our experiments), that the signal is synchronized, and also that frequencies components below 300 Hz and above 3200 Hz are heavily attenuated.

Knowing that the DFT must always keep its symmetry, we work only with coefficients 4 to 26 (a total of 23) when considering 64 points DFTs, the assumed frequency selective channel, and a sampling frequency of 8000Hz.

Assuming that *NQ* is the number of ciphered frames and *NQc* the number of vectors belonging to the codebook, the cryptanalysis procedure is:

1) Compute the 64 DFT of each ciphered frame, keeping the results in vectors $\mathbf{X}_i$, $i = 1, 2, ..., NQ$, which components are x_{ij} (j-th component of the i-th vector), $j = 1, 2, ..., 64$.
2) Compute the absolute values $\mathbf{X}_i$, storing them in vectors $\mathbf{V}_i$, $i = 1, 2, ..., NQ$, which components are v_{ij}. as explained before, we only work with $j = 4, 5, 6, ..., 26$.
3) We classify $\mathbf{V}_i$ as voiced or unvoiced according to Section II and assign one particular codebook for each case. Therefore, in the following steps, when mentioning the word "codebook", it means the codebook assigned for the type of frame classified in this step.
4) Considering the $i-th$ ciphered signal frame: components v_{ij}, $j = 4, 5, 6, ..., 26$ of vector $\mathbf{V}_i$ are sorted in a descending order.
5) We assume that the $k - th$ vector of the codebook, $\mathbf{U}_k$, was formed from the same frame that formed vector $\mathbf{V}_i$ (under analysis). The components u_{kj}, $j = 4, 5, 6, ..., 26$ of vector $\mathbf{U}_k$ are also sorted in a descending order.
6) We name p_1 the highest component of the vector corresponding to the ciphered frame under analysis ($\mathbf{V}_i$), p_2 the second highest, and so on till p_{23}. Also, we name q_1 the highest component of vector $\mathbf{U}_k$, q_2 the second highest, and so on till q_{23}. We then store the pairs (p_1, q_1), (p_2, q_2), ..., (p_{23}, q_{23}).
7) We form a 64×64 permutation matrix, $\hat{\boldsymbol{P}}$, where all elements are zeros except:
   a) the elements located on row p_a with columns q_a, in which the paires (p_a, q_a), $a = 1, 2, ..., 23$ are the same of item item 6;

b) the elements located on row z with column z, in which z is the order of all DFT coefficients not permutable, i.e., 1 to 3, 27 to 39, 63 to 64;
c) the elements located on row $(66 - p_a)$ with column $(66 - q_a)$, such that the signal is kept real (due to the need of keeping the symmetry of the DFT).

8) We multiply vector $\mathbf{V}_i$ by the inverse of matrix $\hat{\boldsymbol{P}}$ obtaining vector $\mathbf{V}_i^{'}$. For the 23 permutable elements of $\mathbf{V}_i$, we compute a permutation error named $e(k)$ defined as the squared norm of the error vector $\mathbf{E}_k = \mathbf{V}_i^{'} - \mathbf{U}_k$, i.e., $e(k) = \mathbf{E}_k^T \mathbf{E}_k$.
9) Repeat steps 5 to 8 for all vectors of the codebook obtaining an $e(k)$ for each $k = 1, 2, ..., NQc$. The index *k* corresponding to the lowest error (k_m) will be assumed related to closest codevector to the ciphered frame.
10) Since all NQc $\hat{\boldsymbol{P}}$ matrices are easily available, we use the one corresponding to k_m-th codevector and pre-multiply its inverse by vector $\mathbf{X}_i$ obtaining the cryptanalysed vector $\mathbf{X}_i^c$.
11) We apply the IDFT to vector $\mathbf{X}_i^c$, obtaining the $i - th$ frame of 64 samples in the time domain.
12) Steps 4 to 11 are repeated to all frames of the ciphered signal.

For the performance evaluation of the cryptanalysis, it was used the IME 2002 corpus which consists of 200 phonetically balanced phrases of the Portuguese language spoken in the city of Rio de Janeiro, recorded from 50 male speakers.

The phrases were the same introduced in [9]. For training the codebook, we have used 9 minutes of speech of the corpus, using 10 speakers and 180 sentences. For the test, it was used 1 minute of speech from the corpus, with 4 speakers and 8 sentences not present in the training.

The tests consisted in ciphering the phrases with the frequency-domain scrambler of Subsection IV-A, with blocks of 64 samples, permuting 23 of them (total of 23! possible permutations per block).

Since the goal of the cryptanalysis is not the perfect reconstruction of the signal but making the scrambled speech intelligible to trained ears, no regular objective measure usually used to access speech quality applies.

Instead, a subjective evaluation was carried out in order to check if the intelligibility of the analyzed signal was possible, i.e., if the meaning of the phrases could be recovered. The percentage of the words correctly understood by the listeners was 78.4%.

V. Conclusions

A new scheme for classifying voiced/unvoiced speech frames of short duration, around 8ms, was detailed. It was also figured out that only three features are able to provide a satisfactory classification, specially for the case of frequency-domain scrambled speech for which traditional methods fail. The reason for this attractive result comes from the fact that the statistical features used are robust to frequency scrambling techniques.

References

[1] L. R. Rabiner, et all, *A Comparative Performance Study of Several Pitch Detection Algorithms*. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-24, no. 5, October 1976.

[2] B. A. R. Al-Hashemy and S. M. R. Taha, *Voiced-Unvoiced-Silence Classification of Speech Signals Based on Statistical Approaches*. Applied Acoustics 25, pp. 169-179, 1988.

[3] D. A. Reynolds, *Speaker Identification and Verification using Gaussian Mixture Speaker Models*. Speech Communication, vol. 17, pp. 91-108, 1993.

[4] C. B. de Lima, A. Alcaim and J. A. Apolinário Jr. *GMM Versus AR-Vector Models for Text Independent Speaker Verification*. In Proc. of SBT/IEEE International Telecommunication Symposium (ITS 2002), Brazil, September 2002.

[5] R. M. Gray, *Vector Quantization*. IEEE ASSP Magazine, pp. 4-29, 1984.

[6] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[7] B. Goldburg, S. Sridharan and E. Dawson, *Cryptanalysis of frequency domain analogue speech scramblers*. IEEE Proceedings-I, vol. 140, no. 4, pp. 235-239, August 1993.

[8] B. Goldburg, S. Sridharan and E. Dawson, *On the use of a frequency domain vector codebook for the cryptanalysis of analog speech scramblers*. IEEE International Symposium on Circuits and Systems, vol. 1, pp. 328-331, June 1991.

[9] A. Alcaim, J. Solewicz, e J. Moraes, *Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no Rio de Janeiro*. Revista da Sociedade Brasileira de Telecomunicações, vol. 7, no. 1, p. 23-41, dezembro de 1992.