COMBINED SPEECH AND AUDIO CODING BY DISCRIMINATION

Ludovic Tancerel, Stéphane Ragot, Vesa T. Ruoppila, and Roch Lefebvre

University of Sherbrooke, Department of Electrical Engineering
Sherbrooke, Québec, J1K 2R1, Canada

ABSTRACT

We propose in this paper a general solution for combined speech and audio coding. Particularly, we describe a speech/music discrimination procedure for multi-mode wideband coding. The speech/music decision is updated only when a low-energy frame is detected, and kept unchanged otherwise. The signal is classified using second-order statistics of discriminant parameters. An experimental CELP/transform coder operating at 16 kbit/s is demonstrated. Results show improved performance when compared to single-mode encoding.

1. INTRODUCTION

Current speech and audio coders fail to achieve good quality on a variety of wideband audio signals at low bit rates. Recent advances have considered hybrid approaches such as multi-mode coding based on open-loop [1] or closed-loop [2] mode decision, or backward-forward linear predictive coding [3, 4]. In this paper we propose to employ high-level open-loop signal classification consisting of a speech/music discrimination (SMD). Figure 1 illustrates the encoding principle in a simplified case, where we consider an automatic switch between two mode coders only: a speech and an audio coder. The problem could be further generalized to select the best coder among a collection of available speech and audio coders, given some channel information or some user defaults.

Several constraints can be set to render the discrimination tractable and to ease the system design. Firstly, we aim at minimizing changes inside coders, so one requirement is to have a reset command used to reinitialize a coder before operation. Secondly, multi-mode coders usually cause audible artifacts when switching from one mode to another. Instead of designing any transient or intermediate mode, we prefer to allow switches in low-energy frames or silences only. Finally, the default coder is set to the audio coder, since it yields better performance on more general audio material than the speech coder. For all these reasons, the speech/music discrimination problem is formulated as a combination of signal activity detection together with a speech/non-speech discrimination. Note that the proposed source-controlled encoding solution is independent from the real nature of mode coders. We consider here a CELP coder for speech and a transform coder for music.

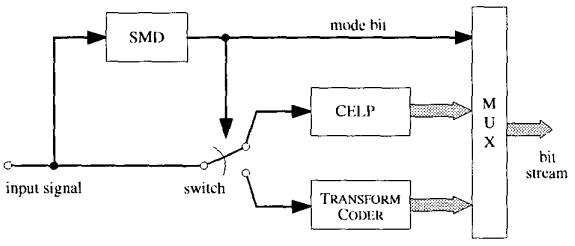

Figure 1: Encoder architecture.

Speech/music discrimination is described in Section 2. Section 3 will present the discrimination performance and subjective audio quality obtained from an experimental wideband ACELP/G.722.1 coder operating at 16 kbit/s.

2. SPEECH/MUSIC DISCRIMINATION

The speech/music discrimination is essentially based on the classical pattern recognition framework consisting of feature extraction followed by classification. To maximize the performance of the classifier we use a long lookahead. In this paper we restrict its value to 500 ms. Relaxing the delay constraint helps to achieve good discrimination performance using long-term parameter trajectories.

Figure 2 shows a block diagram of the proposed signal classifier. In the following we describe each sub-block separately. The signal/noise discriminator is derived from a voice activity dectector (VAD). Its detailed

This work was financed by VoiceAge Corp. and the NSERC.
E-mail: {tancerel,ragot,ruoppila,lefebvre}@gel.usherb.ca

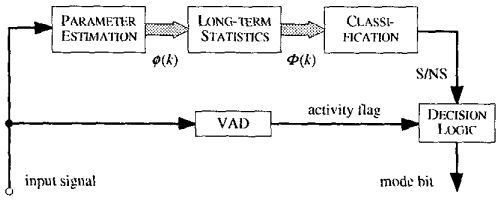

Figure 2: High-level description of the speech/music discriminator.

description can be found in [5].

2.1. Parameter estimation

Because of the variety of music signals, speech/music discrimination is essentially based on speech characteristics. We have used a small set of time and frequency parameters extracted framewise. The parameter vector can be written $\phi(k) = (e, v, g_p, r_2)$ at each frame index k. We compute the short-term energy e of the input signal every 5 ms on frames of length 15 ms using a Hamming window. This parameter is related to the syllabic articulation of speech. The pitch is another important speech characteristic. Its variation is slow but seldom null. For music, these variations are either null or rapid. We use a cross-correlation based algorithm to find a rough estimate for the pitch delay every 5 ms together with the pitch gain g_p. The pitch value is refined using coherence with the neighbouring frames, a voicing measure v is then defined by tracking the pitch evolution relatively to some predefined thresholds. The parameter r_2 represents the evolution of the formants in the signal. We compute the LSF vector $\omega(k)$ from a linear predictive analysis of order 16 and check the correlation r_2 with a preceding frame, where

$$r_n = \frac{\omega(k)^T \omega(k-n)}{\|\omega(k)\| \cdot \|\omega(k-n)\|}.$$

2.2. Long-term statistics and normalization

Trajectory patterns are discriminated based on the second-order statistics of the parameter vector ϕ. Some more elaborate techniques could also be used, such as hidden Markov models. The discrimination efficiency increases with the size of the lookahead used to compute the statistics. We found that a length of 500 ms is a good compromise between global discrimination efficiency and local switching accuracy. The statistical features are the mean μ_e and the variance σ_e^2 of the time envelope, the variance $\sigma_{g_p}^2$ of the pitch gain, and the variance $\sigma_{r_2}^2$ of the inter-frame LSF correlation. They were normalized in the same range to prevent parameters from having more importance from one to another in the classification step. Some non-linear scaling functions was also applied to obtain marginal pdf's close to a Gaussian shape (see Figure 3). The feature vector $\Phi(k)$ can eventually be written as $(\mu_e, \sigma_e^2, v, \sigma_{g_p}^2, \sigma_{r_2}^2)$.

2.3. Speech/music classification

The reader is referred to [6] for a comparison of several classifiers applied to speech/music discrimination. We retain here the Gaussian mixture models (GMM) only. It is a statistical parametric technique which models each class of data by a linear combination of several multivariate Gaussians in the feature space. The mean vectors and covariance matrices of the Gaussians can be derived iteratively by the expectation-maximization algorithm [7].

After classification, the S/NS decision is post-processed by an hysteresis based on the previous decision.

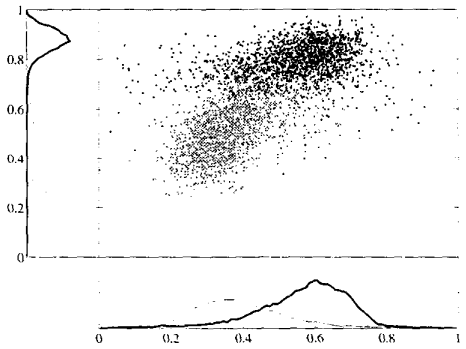

Figure 3: Distribution of features $(\sigma_{g_p}^2, \sigma_{r_2}^2)$ and related marginal probability densities.

2.4. Final mode decision

The switching procedure consists of resetting the coder chosen by the classification. Signal distortion is limited by constraining the switching decision in low-energy frames according to the VAD decision, as explained in Figure 2. The S/NS discrimination is effective only in these parts of the signal. The preference is given to the audio coder, because of its generality. The CELP coder is used only when clean speech is detected.

3. RESULTS

Table 1 shows the estimated average error probabilities of the S/NS classification for different numbers

of Gaussians and with the high-level hyteresis logic. These results were obtained with a training database of 1 h, and a test database consisting of over 40 mn of audio material. The audio files were composed of multilingual clean speech and different types of music. The maximum likelihood (ML) classifier corresponds to multivariate Gaussian classification using one Gaussian only. It is sufficient when data have exact Gaussian shape, but in practice, the performance usually increases with the number of Gaussians.

Table 1: Average error rates for speech/non-speech decision (S/NS).

Classifier	Without Hysteresis (%)	With Hysteresis (%)
ML	5.93	4.44
5-GMM	3.28	2.30
10-GMM	3.17	2.17
20-GMM	3.03	1.98

These classification statistics are strongly database-dependent, since transients or mixtures of speech/music may easily deteriorate the results.

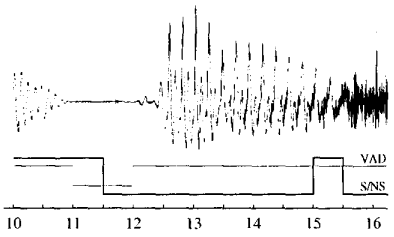

Figure 4: Output example of signal detection (VAD) and speech/non-speech classification (S/NS).

Note that by using a VAD decision to allow switching during signal inactivity only, erroneous S/NS decisions are not considered, as described in Figure 4.

3.1. Coding performance

The speech/music discriminator has been inserted into a two-mode audio coder operating at 16 kbit/s. One bit per frame was needed to describe the selected mode, as shown in Figure 1. The speech coder was an ACELP coder – similar to that of [2]. The audio coder was an implementation of the G.722.1 coder based on the description given in [8]. Its bit rate was reduced to 16 kbit/s (minus one bit per frame). The overall subjective quality was improved compared to single coders.

4. CONCLUSION

We presented a general coding solution based on speech/music discrimination. A large lookahead was used to garantee good classification performance. For this reason, it is suitable only for broadcasting or storage.

ACKNOWLEDGEMENTS

The authors wish to thank Milan Jelinek and Bruno Bessette for providing the voice activity detector and ACELP coder, respectively.

REFERENCES

[1] S. Ramprashad, "A multimode transform predictive coder (MPTC) for speech and audio", IEEE Workshop on Speech Coding, Porvoo (Finland), June 1999.

[2] B. Bessette, R. Salami, C. Laflamme, and R. Lefebvre, "A wideband speech and audio codec at 16/24/32 kbit/s using hybrid ACELP/TCX techniques", IEEE Workshop on Speech Coding, Porvoo (Finland), June 1999.

[3] ITU-T SG16 Contribution (COM16-66), "Draft description of Annex E to Recommendation G.729 - 11.8 kbit/s CS-ACELP speech coding algorithm", Sept. 1998.

[4] J. Schnitzler, J. Eggers, C. Erdmann, and P. Vary, "Wideband speech coding using forward/backward adaptive prediction with mixed time/frequency excitation", IEEE Workshop on Speech Coding, Porvoo (Finland), June 1999.

[5] M. Jelinek and F. Labonté, "Robust signal/noise discrimination for wideband speech and audio coding", submitted to IEEE Workshop on Speech Coding, Delevan, Wisconsin, U.S.A., Sept. 2000.

[6] L. Tancerel, S. Ragot, and R. Lefebvre, "Speech/music discrimination for universal audio coding", 20th Biennial Symposium on Communications, Kingston (Ontario), Canada, May 2000.

[7] T.K. Moon, "The expectation-maximization algorithm", *IEEE Sig. Proc. Mag.*, pp. 47–60, Nov. 1996.

[8] ITU-T SG16 Contribution (COM16-93), "ITU-T G.722.1 Proposed for decision: 7 kHz audio - Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss", Sept. 1999.