SPEECH/MUSIC DISCRIMINATION FOR UNIVERSAL AUDIO CODING

Ludovic Tancerel, Stéphane Ragot, and Roch Lefebvre

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

ABSTRACT

Current low bit rate audio compression technology fails to achieve a good quality for a wide class of audio signals. For instance at 16 kbit/s and a 16 kHz sampling rate, LP-based speech coders perform poorly on music signals, while subband and transform coders cause audible artifacts on speech signals. Speech/music discrimination is therefore an attractive approach to achieve better quality by switching between two coding models – typically a linear predictive speech coder and a transform coder. In this paper we propose a multimode encoding strategy based on pattern recognition techniques. We describe a set of discriminative features and compare several signal classifiers. These are a K-nearest neighbor classifier, a Gaussian mixture model and a multilayer perceptron.

1. INTRODUCTION

The objective of this paper is to describe the design of a speech/music discriminator for audio compression. We consider only audio signals sampled at 16 kHz and restricted to the 50 Hz – 7 kHz bandwidth. This is the typical bandwidth of videotelephony, videoconferencing and low bit rate Internet audio broadcasting. Future wireless standards will also support this bandwidth, yielding a face-to-face quality for speech communication. The application foreseen in this paper is not real-time duplex communications but specifically broadcasting. Therefore problems due to algorithmic or buffering delay are circumvented. Relaxing the delay constraint should also help in achieving a good discrimination performance.

1.1. Motivations and Objectives

This work is mainly motivated by the fact that there is currently no mature technology available for universal audio coding at low bit rates (16 kbit/s, one bit per sample). The control of audio quality often requires the user interaction to select the best coder for a given communication context. There are roughly two classes of low-bit-rate high-quality audio coding systems. On one hand, speech coders use analysis-by-synthesis and take advantage of vector quantization, simple masking techniques, and linear prediction which provides a loose model of speech production. Most of them are currently derived from the Code-Excited Linear Prediction (CELP) model [10]. On the other hand, audio coders are built on a common basic framework. They employ frequency decomposition (a transform or a filterbank), a perceptual model to adapt the stepsizes of scalar quantization, and entropy coding. Typical instances of these coders comprise the MPEG family including MPEG-AAC, Dolby AC-3 and the ITU-T G.722.1 standard. At low bit rates none of these two different approaches can guarantee good quality on both speech and music signals.

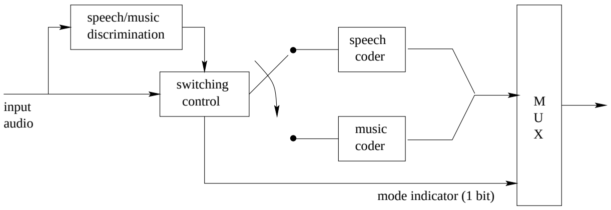

Figure 1: Multimode audio coding by signal classification.

We propose here to use a multimode encoding strategy. In this paper we restrict the mode selection to a speech/music discrimination as described in Figure 1. The two coders operate at the rate of 16 kbit/s and the decision is made on a frame-by-frame basis. The design of the system is focused here on finding a minimal set of parameters and a classifier in order to minimize a classification error probability. This criterion is however loose since even a human being will not always be able to discriminate transients or mixtures of speech and music signals. The real design objective is actually to provide a decision which improves decoded audio quality by switching between two coding methods, compared to using a single coding model.

1.2. Related Work

The speech/music discrimination problem is related to other classification problems like gender and speaker identification, which are based on feature extraction and statistical

† This work was funded by VoiceAge Corp. and the NSERC.

pattern recognition. Therefore, classical feature extraction and classification techniques can be applied to our problem.

In [8], a speech/music discriminator is used for audio statistics. In [9, 11], a discriminator was designed for an automatic speech recognition system for general audio data. Parameters, such as amplitude and pitch features, mel-frequency cepstral coefficients and zero-crossing rates, proved to be discriminant [8, 11, 9, 6]. In [4], the classification is based on line spectral frequencies and zero-crossing rates.

1.3. Outline of the Paper

The paper is structured as follows. The feature extraction is first described in detail before presenting decision and switching procedures. The discrimination results are discussed afterwards.

2. FEATURE EXTRACTION

Feature extraction is a classical front-end function of classifiers. It is intended to reduce the dimension of the classification problem, and also to ease the decision by illuminating the regularities and variable patterns in input signals. Our work is driven by the following basic principles:

- The trajectory of a parameter is often more discriminant than its instantaneous value.
- Information should be extracted in both time and frequency.

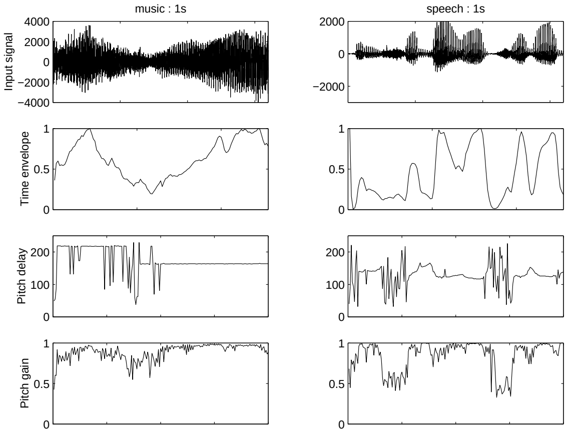

Figure 2: Input signal, time envelope, pitch delay and gain for music (left) and speech (right).

In total 5 features are computed by long-term statistics over 500 ms. These are

- the mean μ_e and variance σ_e^2 of the time envelope e,
- a voicing measure v computed from the pitch delay T and gain g_p,
- the variance $\sigma_{g_p}^2$ of the pitch gain g_p, and
- the variance $\sigma_{E_T}^2$ of the energy in pitch harmonics.

The decision between speech and music is made on frame-by-frame fashion every 20 ms frame with a lookahead of 480 ms, i.e. 24 frames of 20 ms. As a consequence the decision is delayed by a 500 ms in total. A typical trajectory for the envelope and pitch delay and gain is presented in Figure 2. Note that each feature has to be normalized (in the range [0,1] here) prior to classification.

2.1. Time Envelope

Speech signals alternate between high-energy voiced segments and low-energy unvoiced segments. Music signals on the contrary are relatively stationary; their time envelope evolves slowly except for strong beats. A simple way to represent this observation is to compute the time-domain envelope of the input signal. We compute the short-term energy $e[k]$ of the signal $x[n]$ every 5 ms on frames of length 15 ms:

$$e[k] = \sum_{n=0}^{N-1} w_h[n] |x[n - kN]| \qquad (1)$$

where w_h is a Hamming window of 15 ms length, n the time index, N the frame length in samples and k the frame index.

2.2. Pitch Delay and Gain

Pitch delay and gain have proved to be efficient features for discriminating speech and music [6]. For speech signals, pitch tracking is reliable in voiced regions and ranges typically from 60 Hz to 120 Hz for male speakers, and from 120 Hz to 200 Hz for female speakers. The variation of the pitch delay in voiced segments is quite smooth, but seldom null. For music signals, these variations are either null or rapid. Voiced segments usually have a normalized pitch gain near unity whereas unvoiced segments have a very low gain. Music signals usually have a more uniform normalized gain.

We use a cross-correlation based algorithm to find a rough estimate T of the pitch delay every 5 ms, together with the pitch gain g_p. The value of T is then refined by using two frames on each side to improve the results. After that, a voicing level measure v is defined over a 500 ms frame. It tracks the slow evolution of the pitch by thresholding the variation of the pitch delay over the decision frame.

2.3. Energy Ratio in Pitch Harmonics

Music is more stationary than speech. This property was exploited in [9] with frequency-based parameters. However, this method achieves better results for frames larger than 1 second, and so does not apply well in our context. To reduce the dimension of features, we used a single frequency feature: the energy ratio E_T in pitch harmonics. This is motivated by the fact that for voiced speech, the energy is located in the harmonics of the fundamental frequency. In music, the harmonic structure usually arises from different fundamental frequencies. To extract this information, we compute the Fast Fourier transform (FFT) every 20 ms, and use the pitch estimate to quantify the amount of energy E_T in the frequency peaks corresponding to pitch harmonics.

3. DECISION PROCEDURES

After the parameter estimation, a decision has to be made on the nature of the signal to control the switching between the two coding models. This mode decision is a pattern recognition problem. We present here several techniques for this purpose. Three pattern recognition techniques are tested: the K-nearest-neighbor classifier (K-NN), a classifier based on Gaussian mixture models (GMM) and a multilayer perceptron network. A brief description of these classifiers is presented below.

3.1. K-Nearest Neighbor Classifier

K-nearest neighbors classification is a non-parametric technique of statistical pattern recognition. It gives a local estimate for the density function of different classes around an unclassified point [3]. A search is done into a set of previously classified points to find their K-nearest neighbors. The Euclidean distance is commonly used as the neighborhood metric. The most represented class among these K points is assigned to the unclassified point.

The disadvantages of this technique are the need to store a large number of vectors to have an accurate estimate for the density functions, and the number of distances to compute. A remedy to these problems is to condense the number of stored vectors by leaving only a set of representative vectors. This can be made by learning vector quantization.

3.2. Gaussian Mixture Model

A Gaussian mixture model is a statistical parametric technique which models each class of data by a linear combination of several multivariate Gaussians in the feature space. We have to find the maximum likelihood model $L_i(x)$ of each class i:

$$L_i(x) = \sum_{i=1}^{n} \pi_i G_{\mu_i,\Sigma_i}(x), \qquad (2)$$

$$G_{\mu_i,\Sigma_i}(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}, \qquad (3)$$

where n is the number of Gaussians, d is the dimension of the data, and provided that

$$\sum_{i=1}^{n} \pi_i = 1. \qquad (4)$$

The mean vectors μ_i, the covariance matrices Σ_i and the weights π_i of the Gaussians can be iteratively derived by the expectation-maximization algorithm [7].

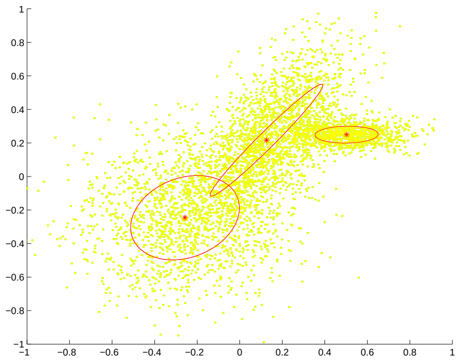

Figure 3: Gaussian mixture modeling of a two-dimensional distribution.

Figure 3 illustrates a Gaussian mixture model for a two-dimensional distribution. The ellipses represent the covariance matrix of each Gaussian, centered on their mean. In practice, the number of Gaussians has to be determined experimentally.

3.3. Multilayer Perceptron

Multilayer perceptrons [5] are a powerful pattern recognition approach, but they are difficult to design. The design of the network architecture is usually based on heuristic methods. In this paper, we consider only multilayer perceptron networks with one hidden layer of nodes. Only one output node is required because there are only two classes to be discriminated. We employ the back-propagation algorithm for optimizing the network parameters. It defines a non-linear separation for multi-dimensional space based on the training database.

4. SWITCHING PROCEDURES

It is very important to avoid classification errors for several reasons. Firstly, the human ear is sensitive to the inter-frame evolution of the signal. Secondly, most of speech and audio coders operate with a memory (e.g an adaptive codebook in CELP coding, or the previous MLT coefficients needed for overlap-add synthesis in the G.722.1 standard). Switching from one mode to another requires to reset the internal coder state. Finally, Coding distortion at the current frame depends on the mode (speech or music) selected at the previous frame. And different coding models may introduce perceptually different coding noise.

We propose here to use a hysteresis based on the previous decision or state latency decision to stabilize the mode selection. Such high-level logic can force efficiently the coherence of the decision.

Another important aspect of the discrimination is also how to handle switching when a transition or short-term silence occurs. It is essential to tune the time response of the discriminator. Yet, the faster it reacts, the more it may switch. Furthermore the discriminative capacity of parameters proved to be very dependent on the length of the decision frame. We achieved satisfying trade-off by varying the lookahead of decision frame.

5. RESULTS

The composition of the training database is a critical part of the discriminator design, since it must represent all possible realizations of the signal. The speech training database is multilingual and composed of phonetically balanced sentences. The music training database is composed of different kinds of music, like rock'n roll, rap, country, techno and classical music. The training database comprises in total 117876 frames of 20 ms.

The test database must also be composed of various segments to test the classification robustness and stability. Obviously the classification performance is strongly dependent on test conditions, such as the noise level, the number of transitions between speech and music and the superposition of speech on music or vice versa. We used two different ways to evaluate the discrimination performance. We first tested the system with a database comprising 136863 frames of speech and music. It gave a good overview of the statistical average performance. The discriminator was also tested in harsh real conditions by using another test database consiting of 10941 manually classified frames with multiple transitions, high noise level and mixtures of speech and music. Such conditions are typical in the context of audio coding.

5.1. Discrimination Statistics

The classification performance is presented in Table 1. It was significantly improved by using an hysteresis. Note that this technique was not used for the K-nearest neighbor classifier.

The Gaussian mixture models and K-nearest neighbor classifier outperformed the perceptron. The Maximum Likelihood (ML) is a subcase of GMM, where each distribution is modeled by only one Gaussian. We observed that increasing the number of Gaussians does not bring any improvement beyond a certain number. The K nearest neighbor classifier is accurate, but computationally very demanding. Clustering like learning vector quantization or tree-search vector quantization could lower its complexity. The number of hidden nodes in the multilayer perceptron was selected after several attempts.

Table 1: Estimated average error probabilities.

Classifier	Without Hysteresis (%)	With Hysteresis (%)
ML	6.70	4.02
2-GMM	6.72	4.50
5-GMM	6.15	2.64
10-GMM	7.96	5.20
20-GMM	8.03	4.55
1-NN	12.17	.
3-NN	7.84	.
5-NN	6.03	.
perceptron [25 hidden nodes]	9.02	7.73

Table 2: Error rates for hard context decision.

Classifier	Without Hysteresis (%)	With Hysteresis (%)
ML	17.46	12.25
2-GMM	14.65	11.39
5-GMM	15.36	11.44
10-GMM	14.79	10.36
20-GMM	14.54	10.28
1-NN	19.60	.
3-NN	16.65	.
5-NN	16.27	.
perceptron [25 hidden nodes]	17.45	11.95

The results for the hard decision context are presented in Table 2. We observed that most of the errors occured during speech and music mixtures, transient, or silence and noise segments. Again the introduction of an hysteresis was significant. It improves classification results because it takes previous decision into account. It prevents the encoding system from switching unexpectedly.

5.2. Subjective Audio Quality

The speech/music discriminator has been inserted into a two-mode audio coder operating at 16 kbit/s. In addition, one bit per frame is needed to describe the selected mode, as shown in Figure 1. The speech coder was an ACELP coder – similar to that of [1]. The music coder was an implementation of the G.722.1 coder based on the description given in [2]. Its bit rate was reduced to 16 kbit/s.

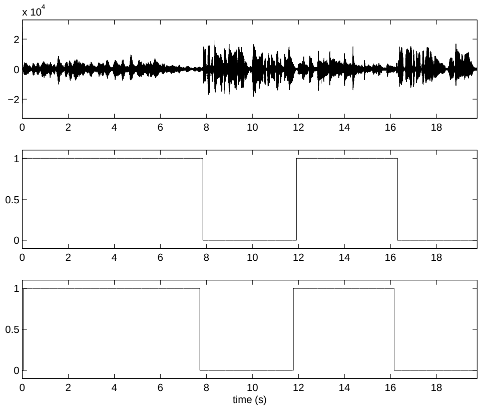

Figure 4: Discrimination result : construction by hand (top), discriminator output (bottom) for speech (0) and music (1).

Figure 4 presents a typical example of our speech/music discrimination versus manual classification. The coder met the requirements of achieving a better quality for the decoded audio. However a few open problems remained. The switching is very criticial, and it may cause a few artifacts. The main problem with the coders is due to coder memory. CELP coding is based on adaptative codebooks, and the G.722.1-based coder needs previous MLT coefficients for overlap-add synthesis.

6. CONCLUSION

This paper presented an application mixing both audio coding and pattern recognition techniques. A speech/music discriminator was designed to bridge the gap between speech and transform coders, and to give an open-loop decision for controlling an adaptive switch. The objective was to improve the quality of broadcasted audio by switching automatically between two coders. The overall algorithmic complexity is lower than in a closed-loop decision proposed in [1]. Furthermore the classification error rate was decreased by an ad hoc high-level switching procedure.

It is important to note that the problem coped in this paper and the proposed solution will be obsolete once a mature universal audio coding model is available. However in short-term the proposed solution is attractive.

ACKNOWLEDGEMENTS

The authors wish to thank Claude Laflamme for helpful suggestions and Milan Jelinek for details on robust pitch tracking. They are very grateful to Vesa T. Ruoppila for reviewing the paper thoroughly.

REFERENCES

[1] 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.

[2] 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.

[3] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory*, 13(1):21–27, 1967.

[4] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal. Speech/music discrimination for multimedia applications. *Proceedings of ICASSP*, to appear 2000.

[5] S.S. Haykin. *Neural Networks: A Comprehensive Foundation*. Maxwell Macmillan International, 1994.

[6] E.S. Parris M.J. Carey and H. Lloyd-Thomas. A comparison of features for speech, music discrimination. *Proceedings of ICASSP*, pages 149–152, 1999.

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

[8] J. Saunders. Real time discrimination of broadcast speech/music. *Proceedings of ICASSP*, pages 993–996, 1996.

[9] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. *Proceedings of ICASSP*, pages 1331–1334, 1997.

[10] M.S. Schroeder and B.S. Atal. Code-excited linear prediction (CELP) : high-quality speech at very low bit rates. *Proceedings of ICASSP*, pages 937–940, 1985.

[11] M.S. Spina and V.W. Zue. Automatic transcription of general audio data: preliminary analyses. *Proc. ICSLP*, pages 594–597, 1996.