A ROBUST SPEECH/MUSIC DISCRIMINATOR FOR SWITCHED AUDIO CODING

Guillaume Fuchs

Fraunhofer Institut für Integrierte Schaltungen (IIS), Germany
guillaume.fuchs@iis.fraunhofer.de

ABSTRACT

Switching between speech coding and generic audio coding schemes was recently proven to be very efficient for coding a large range of audio materials at low bit-rates. However, it strongly relies on a robust classification of the input signal. The aim of the paper is to design a reliable speech and music discriminator (SMD) for such an application. Main attention was laid on getting a good tradeoff between accuracy, reactivity and stability of the decision while keeping the delay and complexity reasonably low. To this end, short-term and long-term features are dissociated before being conveyed to two different classifiers. The two classifier outputs are combined in a final decision using a hysteresis. Objective measures show that a more reliable switching decision is achievable. The SMD was successfully implemented in MPEG Unified Speech and Audio Coding (USAC). It allows the codec to show unprecedented audio quality.

Index Terms— Speech and Music Discrimination, Speech Coding, Audio Coding

1. INTRODUCTION

Switched audio coding is a new efficient approach for unifying speech and music coding at low bit-rates. Historically speech coding and music coding were developed separately for different applications. On the one hand, speech codecs are based on a parametric source model, which is very efficient for representing speech signal at low bit rates but too restrictive for coding other audio materials. On the other hand, generic audio codecs exploit the human auditory and show good perceived quality for music for a wide range of bit-rates. However, they do not perform well at low bit rates on speech signals.

For the sake of universality different works attempted to combine both technologies in a single coding scheme [1, 2]. The latest effort from MPEG to unify speech and music coding came to a switched audio coding approach [3], where a speech coder is used when speech is detected and a generic audio coder is applied to non-speech signals. More recently, the same principle was adopted for the low-delay conversational codec standardized by 3GPP for Enhanced Voice Services (EVS) over LTE [4].

Conventional Speech/Music Discriminators (SMDs) do not entirely fulfil the requirements for switched audio coding. They are generally too complex and introduce a far too high delay. Different works already underlined the problems and proposed real-time and low delay SMDs [5, 6]. However, SMDs are usually designed in view to maximize uniquely the accuracy rate of the classification without studying their behavior in real conditions. Speech over music is a typical scenario where classical SMDs may fail. For such a case, the switched coder should rather stay in one of the two coding schemes and not toggle from one paradigm to the other. Switching from one coding scheme to another is costly in term of bits and erodes the overall performances. Moreover, the juxtaposition of two different coding signatures in the same music section or speech utterance can engender a perceptual quality degradation. The reactivity of the classifier is another important characteristic, since coding artefacts occurring in onsets are perceptual very noticeable.

For the abovementioned reasons, we propose a new SMD, which takes into account the specificities of switched audio coding. As a first step, the features are selected and separated in two categories: the short-term and the long-term features. The short-term features are selected for their great reactivity in discriminating a signal while the long-term features are expected to bring their greater reliability and stability. The two sets of features are used in a short-term and a long-term classifier which are then appropriately combined with a hysteresis. The new SMD is evaluated with newly introduced performance measurements which aim to be more relevant for switched audio coding.

The paper is organized as follows. It begins in section 2 with a system overview. Section 3 describes the extracted features in detail. In section 4, the combination of two classifiers is presented before presenting the performance of the combined system in section 5. The paper is concluded by a final discussion regarding the applications of the proposed SMD.

2. SYSTEM OVERVIEW

The proposed SMD is depicted in Fig. 1. First the features are extracted. Two types of features are differentiated: the short-term (ST) and the long-term (LT) features. ST features are conveyed to both the short-term classifier (STC) and the

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

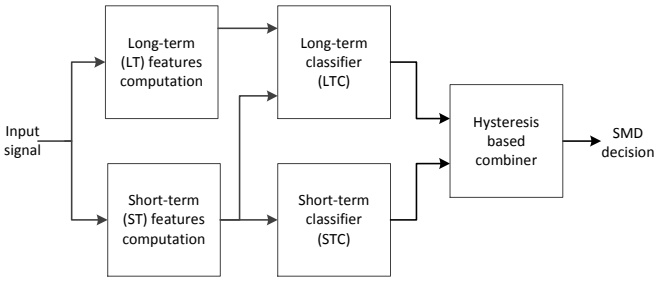


Fig. 1. Block diagram of the proposed SMD. Two classifiers are used and combined by a hysteresis.

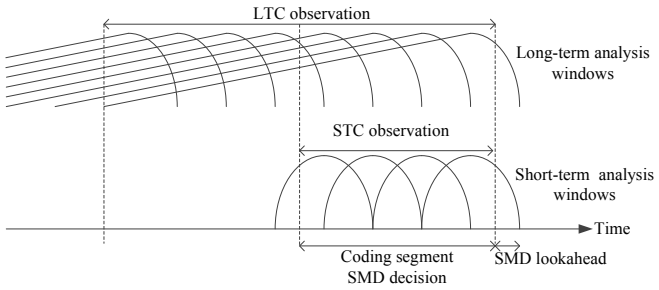


Fig. 2. Analysis windows of the SMD and its framing. The window shapes are only illustrative.

long-term classifier (LTC), while LT features are only considered by LTC. As a consequence the two classifier outputs have a different latency. STC is expected to have an almost instantaneous decision while LTC provides a more stable but also delayed decision.

The paper considers mainly the switched codec MPEG-D USAC, which can switch between the two coding schemes with a granularity of 1024 samples, i.e. 64ms at 16kHz. Since speech is considered to be quasi-stationary on 10-20 ms duration, the coding segments are further divided within the classifiers into 4 frames of 256 samples, i.e. 16 ms at 16 kHz. For each frame a set of features is computed using overlapping analysis windows. Fig. 2 illustrates the analysis window shapes and the time division of the two classifiers. The ST features use analysis windows with limited overlapping region, while LT features use asymmetric windows overlapping with several past frames. Moreover, STC observes the statistics of its features only on the current coding segment, while LTC considers additional past segments. The LTC decision shows then a latency, which will be further taken into account when combining the classifier outputs. As it can be observed the lookahead needed by the whole SMD is low, and corresponds to half of a frame, i.e. 8 ms. For a lower delay switched coder, the present method can be adapted to a granularity as low as 1 frame. In this case a decision is computed every 16 ms for an overall delay as low as 24 ms.

3. FEATURE EXTRACTION

ST features have the aim to capture instantaneous information about the nature of the input signal. They are related to short-term attributes of the signal which can rapidly and at any time change. In consequence ST features are expected to have a very low latency.

LT features result from longer observations of the signal and therefore permit to achieve more reliable classification. In many works [7], variances of static features have been observed to be more discriminating than features themselves. As a rough general rule, music can be considered more stationary and exhibits usually lower variance. On the contrary, speech can be easily distinguished by its remarkable 4-Hz energy modulation as the signal periodically changes between voiced and unvoiced segments. Moreover the succession of different phonemes makes the speech features less constant. In the present work, we consider two long-term features, one based on a variance computation and another based on a priori knowledge of the pitch contour of the speech.

3.1. Short-term features

The present work uses mainly the Perceptual Linear Perception Cepstral Coefficients (PLPCCs) as ST feature. PLPCCs are widely used in speech and speaker recognition [8]. PLPCCs are also used in contrast to Mel-Frequency Cepstral Coefficients (MFCCs) because they share a great part of the functionalities with the Linear Prediction (LP) analysis of the speech coder. PLPCCs can extract the formant structure of the speech as LP does, but by taking into account perceptual considerations PLPCCs are more speaker independent and thus more relevant regarding the linguistic information. An order of 16 is used on the 16 kHz sampled input signal.

Besides PLPCCs, a voicing strength is computed as a short-term feature. The voicing strength permits to distinguish voiced and unvoiced phonemes of the speech and is expected to create at least two clusters in the feature space. The voicing strength uses the speech classification parameters from the VMR-WB speech codec [9]. Zero crossing rate (zc), spectral tilt ($tilt$), pitch stability (ps), and normalized correlation of the pitch (nc) are derived after a Linear Prediction (LP) and a Long-Term Prediction (LTP) analyses of the signal. Both predictions are already present in the speech coding scheme. All 4 parameters are normalized between 0 and 1 such as 0 corresponds to a completely unvoiced signal and 1 to a completely voiced signal. The voicing strength $v(k)$ is then defined for the frame index k as:

$$v(k) = \frac{1}{5}(2 \cdot nc(k) + ps(k) + tilt(k) + zc(k)) \qquad (1)$$

3.2. Moving variance of PLPCCs

The moving variance of PLPCCs consists of computing the variance of the PLPCCs over an overlapping analysis window

covering several frames. In order to emphasize the last frame and to limit the introduced latency, the analysis window is asymmetric with a limited lookahead as illustrated in Fig 2. In a first step, the moving average of the PLPCCs is computed over the last N frames as follows:

$$ma_m(k) = \sum_{i=0}^{N-1} PLPC_m(k-i) \cdot w(i), \qquad (2)$$

where $PLPC_m(k)$ is the mth cepstral coefficient of the M-dimensional feature of the kth frame. The moving variance is then defined as:

$$mv_m(k) = \sum_{i=0}^{N-1} (PLPC_m(k-i) - ma_m(k))^2 \cdot w(i) \qquad (3)$$

where w is a window of length N which is in the present work a ramp slope defined as follows:

$$w(i) = (N-i)/(N.(N+1)/2) \qquad (4)$$

The moving variance is finally averaged over the cepstral dimension:

$$mv(k) = \frac{1}{16} \sum_{m=1}^{16} mv_m(k) \qquad (5)$$

The long-term features are computed with $N = 25$, i.e. considering 400 ms of history of the signal.

3.3. Pitch-based speech merit

The pitch is an important attribute of speech and its time variation over time called the pitch contour follows a particular pattern. Indeed pitch contours in speech fluctuate smoothly during the voiced segments but is seldom constant. During unvoiced segments, the pitch is irrelevant since there is no periodicity. On the contrary, music exhibits very often constant pitch during the whole duration of a note and abrupt deviations during transients. The proposed feature tries to encompass these characteristics by observing the pitch contour on a long time segment. A pitch contour parameter is defined as,

$$pc(k) = \begin{cases} 0 & \text{if } |p(k) - p(k-1)| < 1 \\ 0.5 & \text{if } 1 \leq |p(k) - p(k-1)| < 2 \\ 1 & \text{if } 2 \leq |p(k) - p(k-1)| < 20 \\ 0.5 & \text{if } 20 \leq |p(k) - p(k-1)| < 25 \\ 0 & \text{otherwise} \end{cases} \qquad (6)$$

where $p(k)$ is the pitch lag (inverse of the fundamental frequency $F0$) computed at the frame index k. From the pitch contour parameter, a speech merit, $sm(k)$, is derived based on the observation that speech displays a smooth pitch contour in voiced frames and a strong spectral tilt toward high frequencies in unvoiced frames:

$$sm(k) = \begin{cases} nc(k) \cdot pc(k) & \text{if } v(k) \geq 0.5 \\ (1 - nc(k)) \cdot (1 - tilt(k)) & \text{otherwise} \end{cases} \qquad (7)$$

where $nc(k)$, $tilt(k)$, and $v(k)$ are defined as in (1). The speech merit is then weighted by the window $w(i)$ defined in (4) and integrated over the last N frames.

$$asm(k) = \sum_{i=0}^{N} sm(k-i) \cdot w(i) \qquad (8)$$

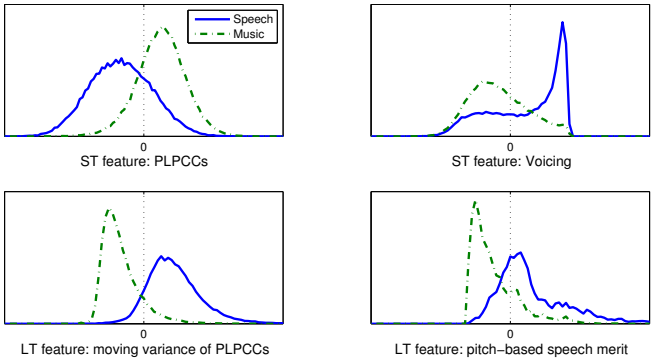


Fig. 3. Histograms of the different features compiled over the training sets.

3.4. Discrimination ability of the features

The discrimination ability of each feature can be observed in the histograms reported in Fig. 3. The statistics were computed on the training set which is composed of 20 min of active speech from different languages and 20 min of music of diverse genres. For illustrative purposes, a Linear Discriminant Analysis (LDA) is applied to the PLPCCs in order to reduce their dimension to 1 and to be able to plot a 1-dimensional histogram. All features are normalized across the speech and music training sets for having a mean of 0 and a variance of 1.

ST and LT features are further evaluated by using Gaussian Mixture Models (GMMs) as a classifier. The discrimination ability is measured as the percentage of true positives in a speech test set and in a music test set. Both test sets are different from the training sets but have the same duration. Three classifiers are compared in Table 1; one using the 2 ST features, another the 2 LT features, and the last one considering all 4 ST and LT features. The effect of the number of mixtures from 1 to 10 is also studied. A decision is computed every segment of 64 ms, i.e. 4 successive frames, as it is done in MPEG-D USAC. Although the LT features have a much lower dimensionality (dimension of 2), it has a better discrimination ability than the ST features (dimension of 17). The best performance comes from the combination of

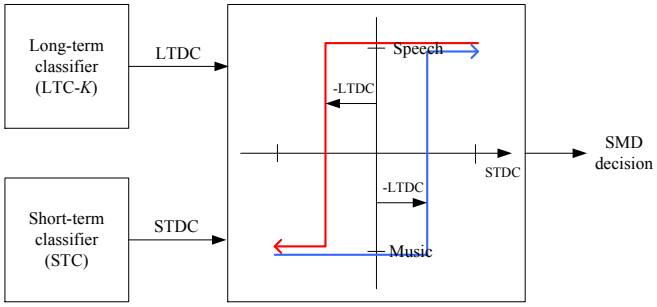


Fig. 4. Hysteresis-based combined classifier system.

the two types of feature. A mixture of 5 Gaussians shows a good compromise between performance and complexity.

4. COMBINED CLASSIFIER SYSTEM

ST and LT features are usually mixed in a single classifier when designing SMDs. However, using the two types of features in a single classifier tends to give more importance to the LT features due to their higher discriminative power. Thus, great other properties of the ST features are not efficiently exploited [6]. In particular, the reactivity of the classification is not explicitly optimized during the training phase, although it has an important role for most of the applications.

The proposed SMD exploits the results of two classifiers: a short-term classifier (STC) considering ST features and a long-term classifier (LTC-K) considering both ST and LT features. STC computes its likelihoods of having speech and music by using a 5-GMMs on the features collected in the current coding segment, i.e. the last 4 frames in USAC. In contrast, LTC-K runs its 5-GMMs over the K last segments, i.e. the $4 \times K$ last frames. A short-term discriminating cue (STDC) and a long-term discriminating cue (STDC) are derived as the logarithmic ratio of the speech model likelihood over the music model likelihood given by the STC and the LTC-K respectively. STDC and LTDC are bounded between -1 and 1, where -1 is the highest likelihood to have music and 1 the highest likelihood to have speech.

As it is depicted in Fig. 4, STDC and LTDC are combined in a hysteresis. STDC is the instant input of the hysteresis, while LTDC controls the width of the hysteresis thresholds. If LTDC and STDC have the same sign, the system will output the common decision. Otherwise, LTDC will prevent or ease the change of decision state by increasing or reducing the hysteresis effect.

5. PERFORMANCE

In order to evaluate the proposed system, and its suitability for switched audio coding, three performance measures are computed. The first one corresponds to the classification accuracy, called here speech against music (SvM) performance. It is evaluated over a set of 20 min of music and a set of 20 min of speech. A second performance measure is done on a 10 min audio item alternating between speech and music every 3 seconds. The classification accuracy is then called speech after/before music (SabM) performance and reflects mainly the reactivity of the system. Finally, the stability of the decision is evaluated by classifying a set of 10 min of speech over music items with different levels of mixture. The speech over music (SoM) performance is derived from the number of state changes of the decision happening in a item and is defined as follows:

$$SoM = \frac{(L-1) - \sum_{l=1}^{L-1} d(l) \neq d(l-1)}{L-1} \qquad (9)$$

where $d(l)$ is the switching decision at the segment l and L is total number of segments to classify.

The long term classifier LTC-K and the short-term classifier STC are used as references for evaluating conventional single classifier approaches. Their performances are plotted in Fig. 5 (a). The STC shows a good SabM performance while having the lowest stability and overall discrimination ability. On the other hand, the LTC-K, and especially by increasing K, can reach a better stability and discrimination ability but only by compromising the reactivity of the decision.

The performance of the combined classifier system, STC/LTC-K, is plotted in Fig. 5 (b) which shows that the combined system is able to reach very interesting tradeoffs between SvM, SabM and SabM performances. Such tradeoffs are difficult, or even impossible, to reach with a conventional single classifier approach.

6. CONCLUSION

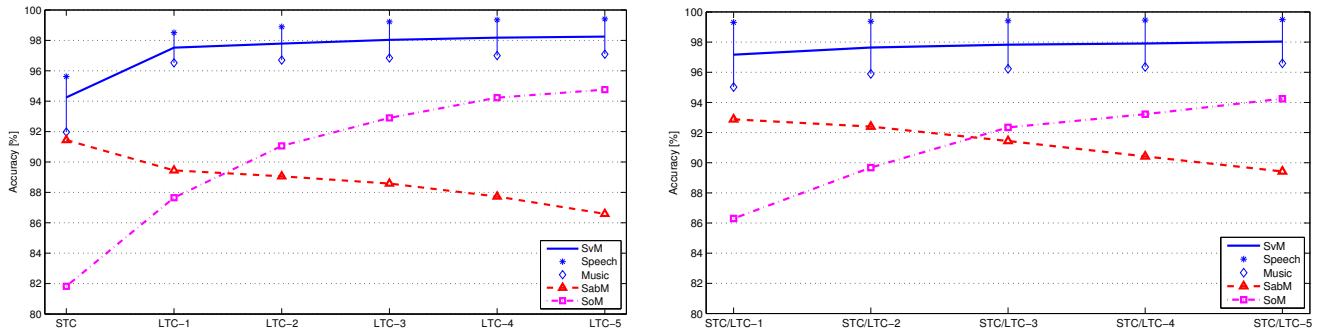
In the present paper, we propose a new SMD using two classifiers. The two classifiers compute a short and a long-term discriminating cues which are combined in a hysteresis-based final decision.

Further, we define new performance measures relevant for switched audio coding and use them to evaluate the new SMD. They show that the proposed combined classifier system has a suitable behavior for switched audio coding. The proposed SMD (namely STC/LTC-3) was used in the MPEG-D USAC performance characterization [10]. Test results confirm that the proposed SMD is well suited for such an application.

The present SMD can be also adapted for lower delay switched coders. Since the signal has to be classified on smaller segments, the performance of the system will be affected. For small segments of about 20 ms or less, a closed-loop approach or a low complex version of it [11] can afford to encode and decode the same segment by the two complete or simplified coding schemes before taking a decision. It

	ST features			LT features			ST and LT features		
	1-GMMs	5-GMMs	10-GMMs	1-GMMs	5-GMMs	10-GMMs	1-GMMs	5-GMMs	10-GMMs
Speech	95.3	95.6	96.3	95.3	97.5	98.1	98.3	98.5	99.2
Music	92.1	92.0	92.8	95.6	95.9	95.4	95.6	96.5	96.2
Average	93.7	94.2	94.6	95.5	96.7	96.7	96.9	97.5	97.7

Table 1. Classification accuracy in %, decisions taken every 64 ms.



(a) (b)

Fig. 5. Performances of different SMDs: conventional single classifiers (a), proposed combined classifier systems (b). The combined classifier systems can reach interesting tradeoffs between accuracy, reactivity and stability of the decision.

will be then interesting to compare the performance of the proposed SMD against the closed-loop decision.

REFERENCES

[1] L. Tancerel, S. Ragot, V.T. Ruoppila, and R. Lefebvre, "Combined speech and audio coding by discrimination," *Proc. 20th Biennal Symposium on Communications*, 2000.

[2] S.A. Ramprashad, "The multimode transform predictive coding paradigm," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 2, pp. 117–129, March 2003.

[3] M. Neuendorf et al., "Unified speech and audio coding scheme for high quality at low bitrates," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP*, 2009.

[4] M. Dietz et al., "Overview of the EVS architecture," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP*, 2015.

[5] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/Music Discrimination for Multimedia Applications," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP*, 2000.

[6] Jun Wang, Haojiang Deng, and Qin Yan, "Real-time speech/music classification with a hierarchical oblique decision tree," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP*, 2008.

[7] M.J. Carey, E.S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP*, 1999.

[8] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc.*, vol. 87, no. 4, April 1990.

[9] M. Jelinek and R. Salami, "Wideband Speech Coding Advances in VMR-WB Standard," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, May 2007.

[10] S. Quackenbush and R. Lefevbre, "Performance of MPEG Unified Speech and Audio Coding," in *Audio Engineering Society Convention 131*, 2011.

[11] E. Ravelli, C. Helmrich, G. Fuchs, and M. Multrus, "Low-complexity and robust coding mode decision in the EVS coder," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP*, 2015.