TWO-STAGE SPEECH/MUSIC CLASSIFIER WITH DECISION SMOOTHING AND SHARPENING IN THE EVS CODEC

Vladimir Malenovsky[*], *Tommy Vaillancourt*[*], *Wang Zhe*[†], *Kihyun Choo*[‡], *Venkatraman Atti*[§]

*VoiceAge Corp., †Huawei Technologies, ‡Samsung Electronics Co., Ltd., [§]Qualcomm Inc.

ABSTRACT

In most internationally recognized standardized multi-mode codecs, signal classification is performed in a single step by either linear discrimination or SNR-based metrics. The speech/music classifier of the EVS codec achieves greater discrimination than these single-step models by combining Gaussian mixture modelling (GMM) with a series of context-based improvement layers. Additionally, unlike traditional GMM classifiers the EVS model adopts a short hangover period, allowing it to track transitions between music and speech. Misclassifications are mitigated by applying a novel decision smoothing and sharpening technique. The results in relatively static environments demonstrate that the new two-stage approach with selective hangover leads to classification accuracies comparable to speech/music classifiers with longer hangovers. They also show that the new approach leads to faster and more accurate switching of coding modes than conventional classifiers for more complex audio environments such as advertisements, jingles and speech superimposed on music.

Index Terms— EVS, GMM, speech/music classification, smoothing, sharpening

1. INTRODUCTION

The Enhanced Voice Services (EVS) codec comprises numerous coding modes, each of which is tailored for a specific class of input signals over a given range of bitrates [1]. For example, the ACELP mode is most efficient when applied on speech signals at low and mid bitrates. The MDCT mode is suitable for pure music or mixed content at low, mid and high bitrates. The Generic signal coder (GSC) technology provides good quality on generic audio content at low bitrates [1]. In order to apply the most appropriate coding mode for any type of input signal at any moment, the EVS codec uses a novel robust technique of speech/music classification.

Song et al. have proposed an improvement to the speech/music classifier of the 3GPP2 SMV codec [2] based on the GMM. In their proposed method, GMM features are calculated as running averages of parameters, including the following: LSF, signal energy, reflection coefficients and a periodicity counter. The initial decision of the EVS speech/music classifier is also based on GMM [3] but its features are calculated either instantaneously (in the current frame) or as a moving average between those in the current and the previous frames. Thus, the effective "memory" of the EVS GMM is just one or two frames. This shorter memory results in a quicker response time by the classifier to abrupt transitions from music to speech; a situation that happens frequently in many scenarios including professionally recorded radio transmissions, or informally when user talks in the presence of background music at discotheques or in shopping malls, cafés and pubs. Song et al. [2] report an almost 60% improvement in the detection of music by their GMM method when compared to the baseline SMV codec. This improvement is achieved on any signal containing some trace of music, incl. speech superimposed on music. Whilst this may be appropriate for the SMV codec, for the EVS codec this would lead to the selection of the MDCT coding mode resulting in distortion of speech utterances. In addition, there are frequent misclassifications of speech onsets resulting from long hysteresis.

The EVS speech/music classifier operates on 20ms frames previously declared as "active" by the VAD [1]. The internal sample rate of the classifier is 12800. The classifier has been trained and optimized on a signal database sampled at 16000 Hz and down-sampled to 12800 Hz. The algorithm has low computational complexity and relatively modest memory footprint. The block diagram of the classifier is shown in Fig. 1.

2. FEATURE SELECTION

The complexity of the classifier is minimized by reusing parameters that have been calculated in earlier stages of the codec pre-processing, i.e. during LP analysis, spectral analysis and open-loop pitch analysis. The initial selection of the feature set for the GMM was performed by analyzing the correlation matrix of the complete set of features whilst running the codec on a training database. This technique was reported by Karnebäck in [4] where the behavior of the 4Hz modulation feature was analyzed using 20 critical bands. We have extended Karnebäck's method by incorporating feature averages, their derivatives and logarithms computed between the current and the previous frames. In this way a set of 68 candidate features, with minimal mutual correlation, was derived and analyzed. The list was then pruned by examining histograms of individual features and calculating their discrimination potential according to the following metric,

$$U_{ftr} = \frac{1}{2}\sum_{j=0}^{M}\left| m_{ftr}^{(sp)}(j) - m_{ftr}^{(mus)}(j)\right|, \qquad (1)$$

where $m_{ftr}^{(sp)}$ and $m_{ftr}^{(mus)}$ are the histograms of the feature *ftr* calculated on the speech and the music training database, respectively, and $M=256$ is the total number of bins covering the values of the feature *ftr*, normalized in the range [0;1]. Note that each histogram was normalized with the total number of frames in the database. The discrimination potential U ranges from 0 (no discrimination) to 1.0 (maximum discrimination). This is illustrated in Fig. 2. This procedure leads to the following set of 12 features (with their respective discrimination potentials) finally

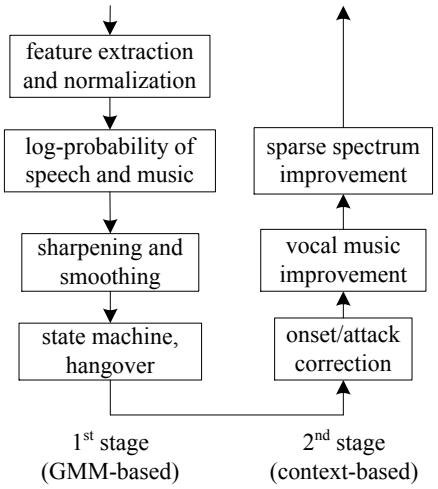

Fig. 1: Schematic diagram of the EVS speech/music classifier

being selected from the initial 68 candidates; open-loop pitch (0.36), normalized correlation (0.37), 5 LSF parameters (0.1-0.28), tonality (0.56), non-stationarity (0.54), residual LP error energy (0.36), spectral difference (0.38) and spectral stationarity (0.35). A detailed description of these features and the exact method of their calculation are provided in [1].

3. THE STATISTICAL MODEL

The GMM uses 6 components (mixtures). It has been trained by the Expectation Maximization (EM) algorithm [5] on a large database containing 2 hours of clean speech, 2 hours of noisy speech and 3 hours of music (classical, modern, rock and pop, jazz, etc.). The clean speech database contained both male and female talkers in 7 different languages. The noise database contained car, street, office and babble noise at various SNR levels. The music database was selected from a collection of various genres of classical and modern music, mainly instrumental. The level of all input signals was normalized to -26 dBov prior to training.

The GMM is a weighted sum of K-component Gaussian densities (K=6) given by the equation

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (2)$$

where $\mathbf{x}$ is a N-dimensional feature vector (N=12) representing the current frame, w_k, k=1,...,K are the component weights and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are the component Gaussian densities. The feature vector is normalized prior to the probablity calculation. Fig. 2 (dotted lines) illustrates the marginal Gaussian densities of the GMM after training. The GMM yields two raw probabilities, p_s and p_m, for the speech and the music models, respectively.

By comparing the values of p_s and p_m in each frame it is possible to obtain a raw discrimination measure between speech and music (SM-RAW). This is achieved by calculating the difference of the log-probability as

$$f_{SM} = \log(p_m) - \log(p_s). \quad (3)$$

The results of the classification based on the SM-RAW are shown later in this paper in Table III. The percentage of correct decisions appears relatively low due to the fact that only 1-2 frames are taken into account during the feature extraction phase.

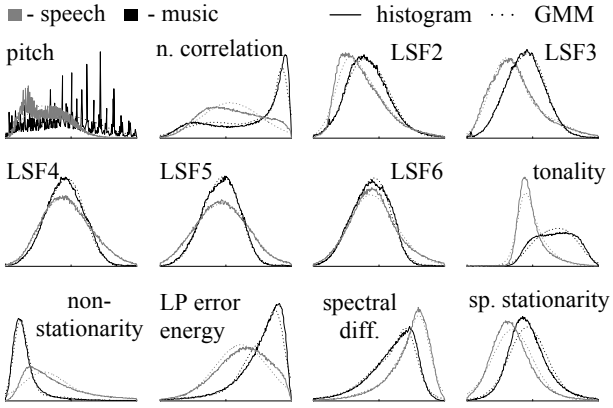

Fig. 2: Normalized histograms of features and marginal densities of the GMM (x-axis scaled to [0;1])

4. DECISION SMOOTHING AND SHARPENING

It was observed that the dynamic range of SM-RAW is relatively low and fluctuates around zero, especially in the presence of mixed signals. This is illustrated in Fig. 3. On the other hand, it is clear that SM-RAW reacts very quickly to transitions from music to speech and vice versa. To fully exploit the discrimination potential of SM-RAW, f_{SM} is smoothed and sharpened by the following adaptive auto-regressive (AR) filter

$$\bar{f}_{SM} = \gamma_c f_{SM} + (1-\gamma_c)\bar{f}_{SM}^{[-1]}, \quad (4)$$

where γ_c is a filtering factor in the range [0;1] and the superscript [–1] denotes a value from the previous frame. The filtered decision, $\bar{f}_{SM}$, is denoted SM-SS. The filtering factor is a combination of smoothing and sharpening effects. It is calculated in each frame according to the following formula

$$\gamma_c = \gamma_E r_{SM}, \quad \gamma_c > 0.01, \quad (5)$$

where γ_E is the scaled relative frame energy and r_{SM} is the slope of SM-RAW. The relative frame energy E_r is computed as the ratio between the current frame energy and the estimated background noise energy [1]. It is scaled in the speech/music classifier as

$$\gamma_E = 1 + E_r/15, \quad 0.01 < \gamma_E < 1. \quad (6)$$

The scaled relative frame energy has values close to 1 in energetically significant segments and values close to 0.01 in the background noise. Therefore, if the signal energy is high more emphasis is put on f_{SM} to follow the raw decision more closely. On the other hand, if the SNR is low, it is naturally more difficult for the classifier to make a correct short-term decision and more reliance is put on past data; the decision is therefore smoothed. The gradient of SM-RAW is also used to sharpen the decision during transitions from music to speech (potential speech). This situation potentially occurs when $f_{SM} < 0$ and $f_{SM} < \bar{f}_{SM}^{[-1]}$. In this case

$$r_{SM} = r_{SM}^{[-1]} + \frac{\bar{f}_{SM}^{[-1]} - f_{SM}}{20}, \quad 0.1 < r_{SM} < 1, \quad (7)$$

where $r_{SM}^{[-1]}$ is initialized (reset) to the value of $-f_{SM}$ each time when $f_{SM} < 0 < \bar{f}_{SM}^{[-1]}$. Thus, the gradient of SM-RAW is positive only during frames when f_{SM} is falling below zero and it represents a quantitative measure of the decrease. This creates a sharpening effect which is illustrated in Fig. 3 (see last but one trace for f_{SM}).

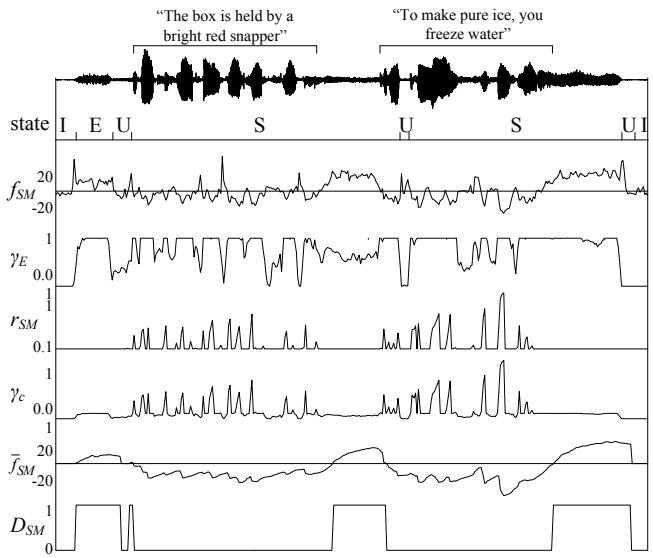

Fig. 3: Decision smoothing and sharpening on a speech sample with background music

5. SIGNAL PARTITIONING AND HANGOVER

The efficiency of the speech/music classiffier is further improved by combining the raw decisions from 0-7 previous frames. This is referred to as the hangover addition or decision smoothing. Applying hangover in every frame would lead to misclassifications for speech onsets where past information has little relevance. In turn, it would lead to the selection of an inapropriate coding mode. To avoid this, the signal is first partitioned and a taylored variable-length hangover is individually applied to each section.

The input signal is partitioned using a simple state machine as shown in Fig. 4. The INACTIVE state is selected as the initial state. It is switched to ENTRY state when the VAD goes to unity. The ENTRY state marks the first onset after a longer period of silence. After 8 frames in the ENTRY state the classifier enters into the ACTIVE state which marks a stable signal with sufficient energy. If the energy suddenly drops closer to the level of the background noise the classifier's state is changed to UNSTABLE where it may stay for up to12 frames. After this period it reverts back to the INACTIVE state. If the energy suddenly increases while the classifier is in the UNSTABLE state, the classifier enters the ACTIVE state, bypassing the ENTRY state. This ensures continuity of classification during short pauses.

With the signal partitioned by the state machine hangover is applied according to the conditions defined in Table II. The final decision of the speech/music classifier (SM-HO) is binary. With the state machine in the INACTIVE state, the classifier output is always zero. In the UNSTABLE state, the classifier decision of the previous frame is maintained. In the ENTRY state, the classifier output is based on a weighted sum of k_{ENT} previous values of SM-RAW where k_{ENT} is the frame counter of the ENTRY state. The weighting factors α_k are given in Table III.

6. CONTEXT-BASED IMPROVEMENT LAYERS

The final decision of the speech-music classifier (SM-HO) is reviewed and potentially corrected in the second stage of the classifier. The concept of decision correction in the second stage has been reported by Chou et al. in [8]. Here, the corrections are applied only during specific signal contexts and are usually based on long-term statistics.

As an example, unaccompanied background vocal music may contain strong tonal characteristics. Typically, the first stage of the classifier declares these frames as "speech" and suggests using the ACELP paradigm. In reality, the MDCT or the GSC mode are better suited to encode such content. Depending on the signal context, a three-way classification is performed at certain bitrates to select among ACELP, GSC or MDCT for improved efficiency while avoiding any serious artifacts. To achieve this, two further independent state machines are deployed gathering statistics over eight SM-HO decisions. In addition, tonal features and some other features from the first stage are re-analyzed to determine potential errors in the classification. In the case that the tonal feature analysis indicates the presence of strong vocal music and if the previous decisions were mainly indicating "speech", then the decision is adjusted to "music" [1].

The context-based switching mechanism also employs the following two new features; spectral sparseness and LP efficiency. Spectral sparseness is calculated from the log-energy spectrum obtained from the spectral analysis and sorted in descending order

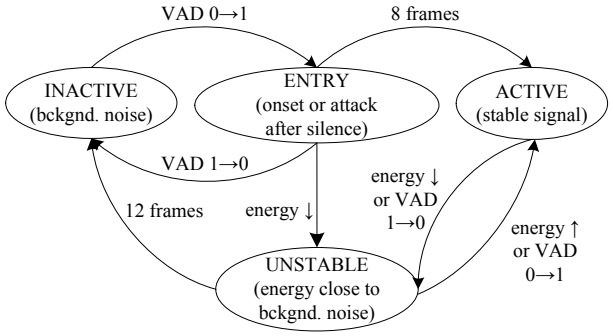

Fig. 4: State machine for signal partitioning

TABLE I: Final decision with hangover

state	condition	SM-HO (D_{SM})
INACTIVE	–	0
UNSTABLE	–	$D_{SM}^{[-1]}$
ENTRY	–	$\alpha_0 f_{SM} + \sum_{k=1}^{k_{ENT}} \alpha_k f_{SM}^{[-k]} > 2$
STABLE	$\bar{f}_{SM} > 0, D_{SM}^{[-1,..,-3]} = 1$	1
	$\bar{f}_{SM} < 0, D_{SM}^{[-1]} = 0$	0
	otherwise	$D_{SM}^{[-1]}$

TABLE II: Weighting factors for SM-HO in ENTRY state

k_{ENT}	α_0	α_1	α_2	α_3	α_4	α_5	α_6	α_7
0	1							
1	0.6	0.4						
2	0.47	0.33	0.2					
3	0.4	0.3	0.2	0.1				
4	0.3	0.25	0.2	0.15	0.1			
5	0.233	0.207	0.18	0.153	0.127	0.1		
6	0.235	0.205	0.174	0.143	0.112	0.081	0.05	
7	0.2	0.179	0.157	0.136	0.114	0.093	0.071	0.05

of magnitude. It represents the minimum spectral bandwidth that covers 75% of the total signal energy, i.e.

$$\sum_{f=0}^{f_{ss}} S(f) = 0.75E_t \,, \tag{8}$$

where f_{ss} is the bandwidth representing spectral sparseness, E_t is the total signal energy and $S(f)$ is the sorted per-bin log-energy spectrum [1]. The LP efficiency is based on a ratio of residual energies of the LP analysis in the logarithmic domain, i.e.

$$\varepsilon_p = \log\left(\frac{E_{err}(13)}{E_{err}(1)}\right), \tag{9}$$

where $E_{err}(13)$ and $E_{err}(1)$ are the residual energies of 13th order and 1st order LP analysis, respectively. To reduce frequent switching the LP efficiency is smoothed by summing ε_p over a period of eight consecutive frames, i.e

$$\bar{\varepsilon}_p = \sum_{i=-7}^{0} \varepsilon_p^{[-i]} \,. \tag{10}$$

It has been found experimentally that signals with high spectral sparseness are better coded with the MDCT mode. At the same time, signals with high prediction gain are more naturally encoded with the GSC technology. The MDCT coding mode is selected by default. The GSC coding mode is selected only for signals with high LP efficiency and non-sparse spectrum. Some hysteresis is applied to the mode selection if the spectral sparseness is reasonably stable. A detailed description of all of the context-based improvement technologies is provided in [1].

7. EVALUATION AND TESTING

The EVS speech/music classifier has been evaluated using several test signals. The clean speech samples (~3-5s) were taken from a speech corpus of Recommendation ITU-T P.800-compliant sentence pairs [9]. The music samples (~15-20s) were obtained from a proprietary database of music extracts. Clean and noisy speech, instrumental and vocal music and mixed signals from real radio recordings have been evaluated. The samples in each test were concatenated by inserting approximately 3s of silence. Background signal (noise or music) was added over the clean signal at appropriate SNR levels. The manual labeling of the real radio recordings was achieved by subjectively listening to the synthesized signal obtained by forcing the codec to either ACELP or MDCT coding mode and picking the optimum quality decision. Recommendation ITU-T G.191 library software tools were used for signal manipulation and level adjustment [6].

Comparison of the classifier described here with the GMM method of Song et al. [2] and with the RMS and zero crossings method of Panagiotakis et al. [7] is summarized in Table III. The method by Song et al. is denoted "SMV" and the method by Panagiotakis et al. is denoted "RMS0X". The table shows the percentage of correct "speech" and "music" decisions on the following testing signals. The "clean" signal contains only clean speech and silence. The "15 dB car"signal contains noisy speech signal with 15 dB car noise in the background. In case of "20 dB music", the speech signal is combined with classical background music. The "radio mix" contains real recordings of some Canadian radio stations in which speech is overlapping with multi-genre music at various SNR levels (ranging from 30 dB up to 0 dB). The "rock and pop" and "classical" testing signals contain only instrumental, single-genre music whereas "opera" contains only vocal music.

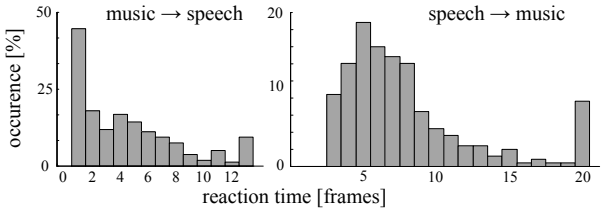

Fig. 5: Reaction time to content switching (SM-HO)

TABLE III: Comparison of speech/music classification accuracy

		RMS0X	SMV	SM-RAW	SM-SS	SM-HO
speech	clean	-	1.000	0.947	0.999	0.999
	15 dB car	0.736	-	0.932	0.994	0.997
	20 dB street	0.711	-	0.840	0.973	0.984
	20 dB music	0.651	0.380	0.681	0.837	0.839
	radio mix	0.729	0.170	0.638	0.773	0.775
music	rock and pop	0.679	0.950	0.879	0.936	0.934
	classical	0.739	0.710	0.988	0.994	0.995
	opera	0.532	-	0.910	0.941	0.939
	radio mix	0.477	0.830	0.823	0.818	0.817

From the results it can be seen that the EVS speech/music classifier clearly outperforms the method by Panagiotakis et al. on all tested signals. The method proposed by Song et al. works well for certain music signals but it completely fails to detect speech in the radio mix passages. The raw decision of the EVS speech/music classifier (SM-RAW) is useful in clean and noisy speech and in pure music, but for mixed content, the reliability falls to 60-70%. The smoothing and sharpening technique (SM-SS) provides the best results on all tested signals. The accuracy is improved by up to 13% without sacrificing the performance in rapidly changing content (radio mix). The detection of music in radio mix signal is only 0.5% worse in SM-SS compared to SM-RAW. Finally, the addition of hangover to the decision (SM-HO) slightly improves the detection of speech at the expense of some loss in the detection of music. This is expected as the objective of SM-HO is to reduce accidental switching of SM-SS during stable sections while maintaining the fast detection of speech onsets. Fig. 5 shows the measured reaction time of the classifier for transitions from music to speech and vice-versa. The classifier can be seen to react to sudden transitions from music to speech within 20 ms (1 frame). The reaction time in the opposite direction is between 100 – 160 ms (5–8 frames). By comparison, the reaction time of the classifier without the smoothing and sharpening method is approximately 160 – 400 ms in both directions, i.e. more than eight times slower.

8. CONCLUSION

In this paper we have presented the EVS speech/music classifier. Unlike traditional GMM-based techniques, this classifier has a very short hangover period. Frequent misclassification is prevented by means of a novel technique based on decision smoothing and sharpening. The method described here uses adaptive hangover logic which achieves very short reaction times for speech onsets. The final decision is also adjusted by a series of context-based improvement techniques. The overall accuracy on pure speech and music signals is more than 92% which is comparable to state-of-the-art classifiers and results show that the EVS speech/music classifier outperforms the methods of Panagiotakis et al. and Song et al on all tested signals. The algorithm forms part of the 3GPP EVS codec published as TS 26.445 [1] .

9. REFERENCES

[1] 3GPP Spec., Codec for Enhanced Voice Services (EVS); Detailed Algorithmic Description, TS 26.445, v.12.0.0, Sep. 2014.

[2] Song, Ji-Hyun, et al. "Analysis and improvement of speech/music classification for 3GPP2 SMV based on GMM." *IEEE Signal Processing Letters,* 2008, pp. 103–106.

[3] J. Pinquier, C. Senac, R. Andre-Obrecht, Speech and music classification in audio documents, in *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, Vol. 4, pp. 4164–4166.

[4] S. Karnebäck. Discrimination between speech and music based on a low frequency modulation feature. *European Conf. on Speech Comm. and Technology*, pages 1891–1894, 2001.

[5] A. P. Dempster, N. M. Laird and D.B. Rubin "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Statiscal Soc.*, vol. 39, no. 1, pp.1–38, 1977

[6] ITU-T G.191, Software tools for speech and audio coding standardization. International Telecommunication Union (ITU), Series G., 2001

[7] C. Panagiotakis and G. Tziritas, A speech/music discriminator based on RMS and zero-crossings, *IEEE Transactions on Multimedia*, Vol. 7, No. 1, Feb. 2005.

[8] W. Chou and L. Gu, „Robust singing detection in speech/music discriminator design“, in *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2001, Vol. 2, pp. 865–868.

[9] ITU-T P.800, Methods for Subjective Determination of Transmission Quality. International Telecommunication Union (ITU), Series P., August 1996.