16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

IMPROVING THE DETECTION EFFICIENCY OF THE VMR-WB VAD ALGORITHM ON MUSIC SIGNALS

Vladimír Malenovský, Milan Jelínek

Speech and audio research group, University of Sherbrooke
2500 Boul. Université, J1K 2R1, Sherbrooke, Québec, Canada
phone: + (1) 819-821-8000 (ext. 62134), fax: + (1) 819-821-7937, email: Vladimir.Malenovsky@USherbrooke.ca

ABSTRACT

Speech codecs are usually equipped with voice activity detection (VAD) algorithm to enable efficient coding of inactive frames and the discontinuous transmission mode (DTX). High VAD efficiency for speech in noisy environments is often traded off against its robustness for music. This is also the case of the VMR-WB codec recently standardized by 3GPP2. Its VAD fails to detect portions of some critical music samples. In this contribution we propose a method to improve the performance of the VMR-WB VAD on music signals. The idea is to measure the stability of tones in the spectral domain by means of per-tone correlation analysis. By using this approach, the music detection accuracy is increased to ~99% and the problem of misclassification is significantly reduced. The proposed method has been implemented in the G.718 codec being currently standardized by the ITU-T.

1. INTRODUCTION

Voice activity detection (VAD) has been typically used to save bandwidth and computational demand in speech communications by detecting inactive periods. During inactive periods, the background noise is only roughly coded and transmitted at very low bitrates, usually using discontinuous transmission (DTX). At the decoder side, the background noise is regenerated by means of a technique called comfort noise generation (CNG). An efficient VAD is of particular importance in CDMA systems and packet-based communications. In CDMA systems (e.g. cdmaOne and cdma2000) the DTX is not used, and the inactive speech is encoded using the lowest available bit rate. An efficient VAD is needed to maximize the system capacity and improve the overall performance. In packet-based communications, the packets have usually long headers with respect to the frame size of low bit rate speech coders. The DTX then allows for significant bandwidth savings by reducing the number of packets to be transmitted.

In recent years a new trend emerged in developing complex speech and audio coding systems with high transmission efficiency at low bitrates, and high perceptual rendering at high bitrates for both speech and audio inputs. This trend can be also observed in the standardization efforts within ITU-T. Many high-rate extensions to existing low-rate standards have been recently recommended or are being standardized, such as G.711.1, G.722.1 full-band or G.729.1. For these systems, the traditional VAD algorithms usually do not provide sufficient accuracy in the detection of audio signals and they are being replaced by more generic SAD (signal activity detection) algorithms. For example, Appendix III of the G.729B codec, addresses the problem of classifying portions of very long and high-level tonal signals as "inactive speech". In the AMR-WB codec, a tonal detector is used to detect information tones, vowel sounds and other periodic signals. A music detector is also included in the SMV vocoder to classify an input signal as music or non-music, which is then used in the rate-selection mechanism.

In this contribution we propose a method for improving the VMR-WB VAD algorithm so that it can be used as a SAD in the G.718 codec. The VMR-WB VAD algorithm is highly efficient in coding speech signals with various background noises, as was shown in 3GPP2 standardization tests. However, VMR-WB has been developed mainly for speech coding and it fails to detect certain critical music samples, e.g. low-pace piano or passages dominated by percussions.

The paper is organized as follows. In section 2 we summarize the VMR-WB VAD algorithm. In section 3 we describe our method to improve the VAD performance for music signals by means of tonal stability analysis and modified non-stationarity measure. Finally, in section 4, we provide some experimental results using the proposed method and compare it with some alternatives.

2. VMR-WB VAD ALGORITHM

The VMR-WB VAD [1] proceeds in two stages. In the 1st stage, the VAD algorithm (Figure 1) makes its decision about speech activity by comparing an average SNR per frame to a certain threshold, which is a function of long-term SNR. The average SNR per frame is calculated using energies in critical bands [8]. It is defined as

$$SNR = 10\log\left(\sum_{j=b_{\min}}^{b_{\max}} \frac{E(j)}{N(j)}\right), \qquad (1)$$

where $b_{\min}$ and $b_{\max}$ are the indices of the minimum and the maximum critical band of the useful bandwidth, respectively. $E(j)$ is the average active signal energy in a critical band j and $N(j)$ is the estimated noise energy in the same critical band. As a result of the 1st stage, called "decision" in Figure 1, a flag v_f is set to one if signal activity is detected.

The 2nd stage of the VAD algorithm, called "noise estimation" in Figure 1, serves to decide when to update the

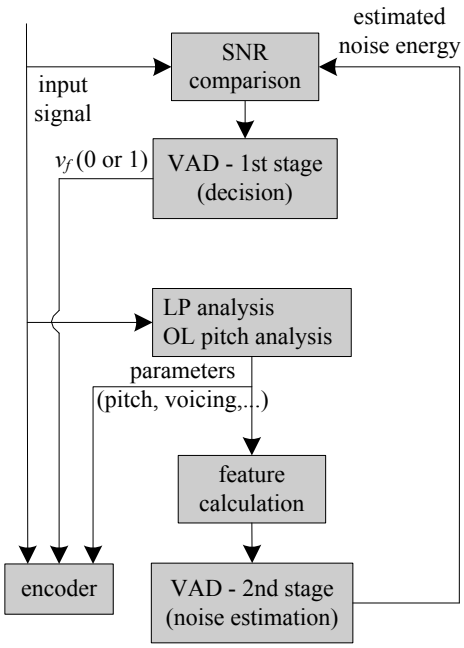

Figure 1 - Schematic description of the two-stage VAD algorithm in the VMR-WB and G.718 codecs

estimated noise energy. It is based on a set of parameters, independent of the average SNR per frame, and relatively insensitive to noise-level variations. Given that the characteristics of background noise evolve slowly, the 2nd stage of the VAD is allowed to make more incorrect decisions on inactive signal, but it is not allowed to make any incorrect decision on active signal.

By dividing the VAD in two stages, the thresholds of activity decision in the 1st stage can be adjusted to a desired sensitivity without affecting the robustness of the decisions in stage 2. Further, it solves the problem of locking the VAD decision in case of a sudden increase of the background noise level.

The following set of four features is used in the 2nd stage of the VAD algorithm:

- signal non-stationarity
- pitch stability
- voicing measure (max. normalized correlation)
- ratio between 2nd order and 16th order linear prediction (LP) residual error energies.

A detailed description of these features is provided in [2]. They are calculated on a frame-by-frame basis and, if none of them exceeds its threshold, the noise energy per critical band $N(j)$ is updated.

Of particular interest is the signal non-stationarity parameter, which is defined as:

$$p_{nonsta} = \prod_{i=b_{\min}}^{b_{\max}} \frac{\max\left\{E(i), \bar{E}(i)\right\}}{\min\left\{E(i), \bar{E}(i)\right\}}, \qquad (2)$$

where $\bar{E}(i)$ is the long-term average of the active signal energy. The signal non-stationarity feature triggers most active speech decisions. This is shown in Table 1 where the percentage of active frames, successfully detected by each of the parameters, is presented.

Table 1 - Percentage of active decisions triggered by individual features in the 2nd stage of the VMR-WB VAD algorithm.

content	signal non-stationarity	pitch stability	voicing	energy ratio
clean speech	99.51	50.24	46.45	75.67
rock & pop - instrum.	86.53	47.14	38.44	31.20
rock & pop - vocal	95.05	57.56	32.13	47.87
classical - instrumental	85.06	70.27	31.43	41.46
classical - vocal	89.22	67.43	27.81	72.54

Table 2 - Percentage of correct decisions in the 1st and 2nd stage of the VMR-WB VAD algorithm.

content	1st stage		2nd stage	
	active	inactive	active	inactive
clean speech	98.97	99.98	100.00	61.26
speech + office noise	98.03	99.24	99.33	41.26
speech + car noise	99.46	99.85	99.87	73.21
speech + street noise	99.80	99.47	99.96	71.97
rock & pop - instrumental	90.67	81.09	98.82	80.07
rock & pop - vocal	97.02	100.00	99.28	78.23
classical - instrumental	92.03	99.75	98.18	87.36
classical - vocal	95.27	100.00	98.96	90.88

The results in Table 1 were calculated using several music samples at 16 kHz, each having a length of 20s. They were selected from various music genres and categorized as *classical* (incl. jazz, blues, etc.) and *rock&pop* (incl. dance, disco, hip-hop, etc.) and then sub-divided as *vocal* and *instrumental*. For each of the four categories, one long file of approximately 7 min. has been created by concatenating several samples together. There was a pause of about 2-3s between each two consecutive samples. All recordings were manually labelled with active and inactive signal marks. From Table 1 we see that signal non-stationarity successfully triggered 85-95% active music decisions, whereas the other features only 27-73%. In addition, all features except non-stationarity are dependent on the music genre.

Table 2 shows the percentage of correct decisions of the VAD algorithm in each stage when all features are combined together. The first and the second column correspond to the 1st stage and the other two columns to the 2nd stage. It can be seen that, for noisy speech, the VAD's detection accuracy in the 1st stage is more than 98%. For classical instrumental music it is only 92% and for rock&pop instrumental music even 91%. This is caused by an incorrect detection of active frames in the 2nd stage of the VAD (third column). Incorrectly detected active frames are called "Type II" errors. From Table 2 we see that the percentage of Type II errors (the complement to 100%) in the 2nd stage is approximately the same for speech and for music. While for noisy speech, the Type II errors occur in low-energy regions where speech is generally buried in noise and the error has not a big impact, for music the Type II errors occur often in perceptually important segments with significant energy. In music, a Type II error basically means that an active signal has been erroneously declared as inactive, and that background noise energy has been updated with active signal energy. Subsequently, the first-stage VAD often stops detecting active signal and hangs in a wrong state for several frames until the noise energy is correctly re-updated. In systems with DTX operation, the signal

is encoded using a CNG during this period which provides a very poor quality for music signals.

In the second and fourth column of Table 2 we see the percentage of inactive frames that were correctly detected. The incorrectly detected inactive frames are called Type I errors. They are not critical to the quality of synthesized signal, but should be kept as low as possible as they generally increase the average data rate of the codec. The Type I errors in the 2nd stage reduce the update rate of the background noise energy which affects the accuracy of the decision in the 1st stage. The Type I errors in the 1st stage force the codec to use more bits for encoding which translates to higher bit rate.

3. THE PROPOSED FEATURES

Our goal was to enhance the performance of VMR-WB by eliminating Type II errors on music signals without significantly affecting the VAD decision efficiency for noisy speech. When analyzing the Type II errors of the VMR-WB VAD algorithm on music signals, we observed that the failures in the second-stage mostly happened:

(a) after certain attacks of signal energy (beats of drums, cymbals or piano);
(b) in a high-frequency energetic signal, such as handclapping or castanets;
(c) during low-pace piano concerts between two consecutive key strokes.

It is not possible to solve these failures by only adjusting the parameters and the thresholds of the VMR-WB VAD. This would lead to a dramatic increase of Type I errors and, consequently, to much lower efficiency of the first-stage VAD. Instead, we propose to solve problems (a) and (b) by modifying the non-stationarity feature of the VMR-WB VAD, and problem (c) by adding a new feature called "tonal stability".

3.1. Modified non-stationarity

Most of the occurrences of problems (a) and (b), described above, coincide with non-stationarity failures. By analysing the behaviour of this feature, it was found that it failed mostly when a sharp energy attack in a signal was followed by a slow energy decrease. From Equation (1) we see that non-stationarity depends on a long-term average of the active signal energy. The long-term average is given by

$$\bar{E}(i) = \alpha \bar{E}(i) + (1-\alpha)E(i)\ , \text{ for } i=b_{min},..,b_{max}. \qquad (3)$$

The forgetting factor $\alpha = 0.024E_t - 0.235$ is dependent on a total frame energy, E_t, and is limited by $0.5<\alpha<0.99$.

To overcome the failures, a new long-term average $\bar{E}'(i)$ is established and calculated in the same way as in Equation 2. Unlike (2), its updating is reset during energy attacks, by setting $\alpha = 0$. The energy attacks are detected in the following way. First, for critical bands $10,..,b_{max}$, which corresponds to the range from 1270 Hz to the half of the sampling frequency, ratios between energies are calculated as

$$r(i) = \frac{\max\{E(i), E_{-1}(i)\}}{\min\{E(i), E_{-1}(i)\}}\ , \text{ for } i=10,..,b_{max}. \qquad (4)$$

Thus, only energy bins exceeding the 10th critical band are considered in order to increase the discrimination capabilities of energy attacks. The index E_{-1} refers to the energies of the previous frame. A weighted sum of the ratios is then calculated as follows

$$E_{att} = \frac{\sum_{i=10}^{b_{\max}} \max\{E(i), E_{-1}(i)\}.r(i)}{\sum_{i=10}^{b_{\max}} \max\{E(i), E_{-1}(i)\}}, \qquad (5)$$

which is subject to a threshold $t_{att} = 5.0$. The threshold has been found experimentally on a large database of speech and music signals. If the threshold is exceeded, an energy attack is detected, and the updating of $\bar{E}'(i)$ is reset. The modified non-stationarity is then evaluated in the same way as the original non-stationarity (see Equation (2)), but using the long-term average of Equation (3). Since the forgetting factor α is reset in every energy attack, the modified non-stationarity triggers active decisions in a few frames following the attack. This replaces exactly the failures of the original non-stationarity and the problem is eliminated.

3.2. Tonal stability

The tonal stability capitalizes on the harmonic stationary structure of music signals and stems from ideas about signal tonality in [3-7]. In [3], spectral flux (SF) parameter is used to measure a spectral difference between the current and the previous frame. This idea is used also in the proposed method and further expanded to include multiple past frames. Hawley [4] proposes a music detector based on harmonic entropy measures and in [5], spectral flatness measure (SFM) is used, which is an estimation of the tone-like quality of a spectrum. The proposed method analyzes peaks of the spectrum to create an image about signal tonality. Then, an entropy measure is applied to quantify the variance of the image. This concept is also seen in [6], where spectral centroid (SC) is introduced to measure the spectral shape and "brightness" of the spectrum. Finally, in a recent work of Hosseinzadeh and Sridhar [7], a new set of features is developed for speaker recognition. These features are used in conjunction with the classical features characterizing the vocal tract. Among them, there is spectral crest factor (SCF), which provides a measure for quantifying the tonality of the signal. A detailed description of the proposed method is given below.

In the spectrum of an audio signal there are typically several peaks. For harmonic music signal these peaks represent tones. Musical tones usually remain stable (position and amplitude) for several frames whereas for noise signals they tend to diminish quite rapidly. To take an advantage of this phenomenon we first need to detect the tones in the spectrum. Let the spectrum of the current frame be denoted as $S(k)$, where $k=0,1,..,N-1$ and the vector of indices corresponding to its spectral minima as m_l, $l=0,1,..,L-1$. We calculate a spectral floor, which is a piece-wise linear function connecting all spectral minima. Each piece between two consecutive minima m_l and m_{l+1} is defined as

$$S_{fl}(k) = a.(k-m_l)+b\ , \text{ for } k = m_l,..,m_{l+1}-1 \qquad (6)$$

where a is the slope of the line and $b=S(m_l)$. The slope is calculated as

$$a = \frac{S(m_{l+1}) - S(m_l)}{m_{l+1} - m_l}\ . \qquad (7)$$

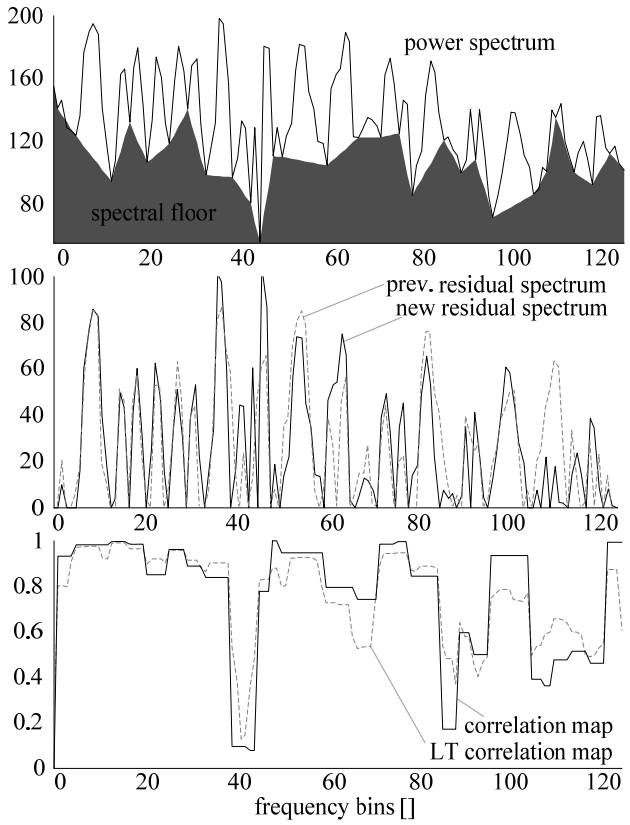

Figure 2 - Tonal stability calculation

Before the first minimum and after the last minimum, the spectral floor is identical to the original spectrum, i.e.

$$S_{fl}(k) = S(k), \quad \text{for } k = 0,..,m_0 - 1$$
$$S_{fl}(k) = S(k), \quad \text{for } k = m_{L-1},..,N-1. \tag{8}$$

The spectral floor is then subtracted from the orig. spectrum

$$S_{res}(k) = S(k) - S_{fl}(k), \text{ for } k = 0,..,N-1, \tag{9}$$

and the residual spectrum is correlated piece-by-piece with the residual spectrum of the previous frame, i.e.

$$C_l = \frac{\sum_k S_{res}(k) S_{p,res}(k)}{\sum_k S_{res}^2(k) \sum_k S_{p,res}^2(k)}, \quad k = m_l,..,m_{l+1}, \tag{10}$$

where l is the piece number, S_{res} (k) is the residual spectrum of the current frame and $S_{p,res}$ (k) is the residual spectrum of the previous frame. A "correlation map" is then created by concatenating C_l of all pieces, i.e.

$$C(k) = C_l, \text{ for } l = 0,..,L-1 \text{ and } k \in < m_l : m_{l+1}). \tag{11}$$

To take a decision about tonal stability, a long-term (LT) correlation map is calculated as

$$C_{LT}(k) = \alpha C_{p,LT}(k) + (1-\alpha) C(k), \text{ for } k = 0,..,N-1, \tag{12}$$

where $C_{p,LT}$ (k) is the LT correlation map calculated in the previous frame and α=0.9 is the filtering factor. The algorithm is started with $C_{p,LT}$ (k)=0, k=0,..,N-1. Note that the LT correlation map is updated in every frame on a bin-by-bin basis. Finally, the whole LT correlation map is summed together to obtain a quantitative measure of tonal stability, i.e.

$$p_{tonal} = \sum_{k=0}^{N-1} C_{LT}(k). \tag{13}$$

The result is then compared to an adaptive threshold, which is limited by 49 and 60, and initialized to 56. These values have been found empirically on a large database of speech and music signals. If tonal stability is detected (the threshold is exceeded) the threshold is decreased by 0.2, otherwise it is increased by 0.2. By decreasing the threshold, the probability of declaring a frame as active is increased and vice-versa. Such behaviour is useful at the end of active signal periods since it introduces the effect of hangover. The main steps of the tonal stability calculation are illustrated in Figure 2. The spectrum in this figure was calculated on active segment of string music, using a 256-point FFT analysis.

4. EXPERIMENTAL RESULTS

The proposed features were tested on the VMR-WB VAD algorithm and their performance compared with several alternatives. They were integrated in the 2nd stage of the VMR-VAD and their thresholds optimized. Each threshold was set to a value which results in the elimination (if possible) of Type II errors in the 1st stage of the VAD and, the lowest rate of Type II errors in the 2nd stage of the VAD.

The same testing signals as in section 2 were used in the performance evaluation. For all tested features, performance on noisy speech was recorded only for office noise as it is one the most difficult noises with respect to VAD decision. The results are summarized in Figure 3.

In Figure 3 above, we see a percentage of correctly detected active frames in both, the 1st stage and the 2nd stage of the "improved" VMR-WB VAD. Note that only one feature at time was tested, in order to obtain comparable results. From the graph it is seen that the best detection accuracy is achieved with the tonal stability both, in stage I and stage II. The other features are less efficient, especially for instrumental music. The tonal stability improves the percentage of correctly detected music frames with respect to VMR-WB VAD in the following way:

rock&pop instrumental	from 90.67 to 97.30,
rock&pop vocal	from 97.02 to 98.24,
classical instrumental	from 92.03 to 97.97,
classical vocal	from 95.27 to 97.67.

The reference values are taken from Table 2 (1st stage).

In Figure 3 below wee see the results for inactive signal detection. In the 1st stage, the percentage of correctly detected inactive frames is close to 100% for the tonal stability, the modified non-stationarity and the spectral flux. Further, for these features, the detection accuracy does not depend on music genre. For the other features, the accuracy is low mainly for rock&pop instrumental music. In the 2nd stage of the VAD, the detection accuracy is more or less the same for all features, with the exception of spectral centroid, which scores low for classical instrumental music. Also, the tonal stability has a slightly lower performance for classical music. However, an important result is that, for speech signal corrupted by office noise, the tonal stability does not increase the rate of Type I errors in the 2nd stage by more than 5% when compared to the original VAD. This is also verified and confirmed for the other types of noise, used in Table 2.

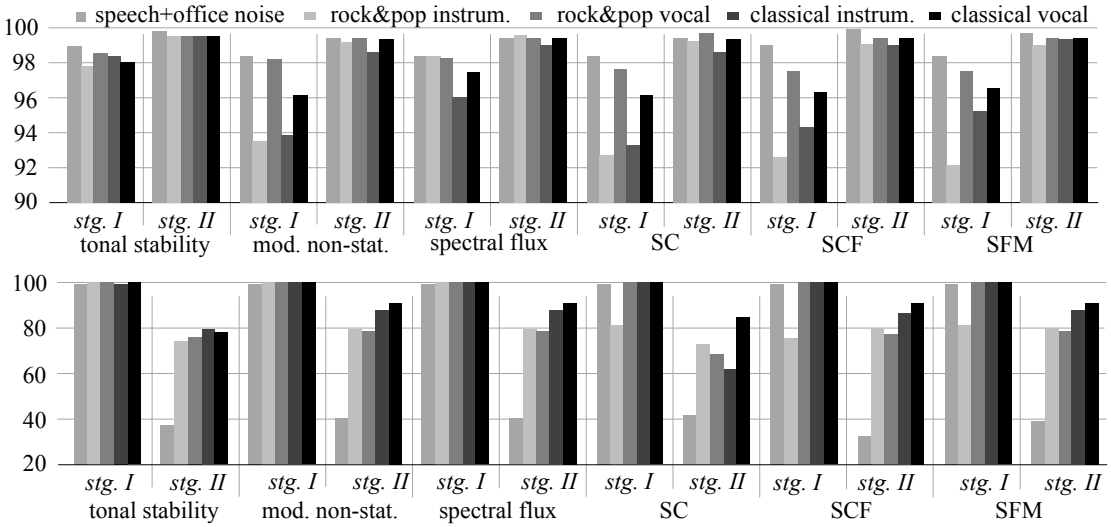

Figure 3 – Comparison of different features implemented in the VMR-WB VAD algorithm: active decisions (above), inactive decisions (below).

The performance of the VAD is further improved by combining the tonal stability with other feature(s). The same framework and the same signals as in the first experiment are used. The combination of the tonal stability with the spectral flux increases the detection accuracy of active music signals in the 1st stage by 0.1-0.4%. On the other side, a conjunction of the tonal stability with the modified non-stationarity increases the detection accuracy of rock&pop instrumental signal by 1.72% and the classical instrumental signal by 0.99%. This is a significant progress which is explained by the fact that for these types of music, the two features are complementary. The tonal stability improves the detection of harmonic parts of the signal whereas the modified non-stationarity improves the detection of percussion-dominated passages. An incorporation of other features does not increase substantially the detection accuracy. At the same time the percentage of correctly detected inactive frames is basically unchanged when the tonal stability is combined with the modified non-stationarity. It was also verified that a combination of three features does not lead to another improvement of VAD performance and would only result in a wasting of computational resources.

5. CONCLUSION

In this paper we consider an improvement to the VMR-WB VAD algorithm for the detection of music signals. Two features are proposed for the 2nd stage of the VAD algorithm, the tonal stability and the modified non-stationarity. The tonal stability capitalizes on a harmonic structure of some critical music samples and measures its long-term invariance. On the other side, the modified non-stationarity exploits information about energy attacks of certain percussion-dominated signals to complement the original non-stationarity feature.

The proposed features were compared with several alternatives and tested on four types of music and noisy speech. It was shown that, with the proposed method, the percentage of active music frames, successfully detected by the VAD algorithm in the 1st stage, can be significantly increased. For vocal music, it is by ~1-2% and, for instrumental music, by ~6-7%. Compared with the original VMR-WB VAD, the false detection of active music is basically eliminated, whereas the detection accuracy of inactive signal is decreased only by 5% in the 2nd stage of the VAD.

The proposed method was implemented in a SAD algorithm of the G.718 codec.

REFERENCES

[1] M. Jelínek and R. Salami, "Wideband speech coding advances in VMR-WB standard," in *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. 15, pp. 1167-1179, 2007.

[2] M. Jelínek and R. Salami, "Noise reduction method for wideband speech coding," in *Proc. EUSIPCO*, Vienna, Austria, pp. 1959–1962, Sep. 2004.

[3] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. IEEE Conf. on Acoustic Speech and Signal Processing*, Munich, Germany, pp. 1331-1334, Apr. 1997.

[4] M. Hawley. *Structure out of sound*. PhD thesis, MIT Media Laboratory, 1993.

[5] P. Cano, et al., "A review of audio fingerprinting". *Proc. IEEE Workshop on MMSP*, vol. 41, no. 3, pp. 271-284, 2005.

[6] A. Ramalingam and S. Krishnan, "Gaussian mixture modelling using short time Fourier transform features for audio fingerprinting", *IEEE Trans. on Forensics and Security*, vol. 1, issue 4, pp. 457-463, Dec. 2006.

[7] D. Hosseinzadeh and S. Krishnan, "Combining vocal source and MFCC features for enhanced speaker recognition performance using GMMs", *Proc. IEEE Workshop on MMSP*, Oct. 2007.

[8] J.D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *Proc. IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314-323, Feb. 1988.