AN OBJECTIVE MEASURE FOR THE MUSICAL NOISE ASSESSMENT IN NOISE REDUCTION SYSTEMS

Nima Derakhshan, Mohsen Rahmani, Ahmad Akbari, Ahmad Ayatollahi

Iran University of Science and Technology, Narmak, 16844, Tehran, Iran.
Tel: +98-912-5478076, Fax: +98-21-77491128, nima_derakhshan@ee.iust.ac.ir, {m_rahmani, akbari, ayatollahi}@iust.ac.ir

ABSTRACT

In this paper, a new objective measure is proposed to assess the amount of the wide-band musical noise in an acoustic signal. The measure uses time and frequency characteristics of the musical noise to determine the musicalness of the residual noise. The proposed measure uses noise masking thresholds to measure how much audible the musical noise components are. Furthermore, a harmonicity measure is used to discriminate between the musical noise components and the speech components. At the end, the correlation of the proposed measure with the subjective listening test results is calculated and its performance is verified in experiments.

Index Terms— Speech enhancement, Acoustic noise

1. INTRODUCTION

Today, noise reduction is an important part of many speech processing systems. In the mobile telephony it is necessary to reduce the background noise before speech is coded and transmitted. The background noise is unpleasant and becomes even more so when coded. Noise reduction is always accompanied by noise distortion, which can be very annoying for a listener. The quality of speech and background noise should always be examined in the stage of system development to optimize the system performance. Consequently, a reliable measure, which can automatically predict the perceived quality of signals, becomes very valuable.

It is necessary to assess the performance of a noise reduction system separately in two aspects, the foreground speech quality and the background noise disturbance. Although much advancement is made in predicting the quality of the coded speech signals, there are few objective measures developed for assessing the performance of the noise reduction systems. The work carried out in this area mostly consists of combining the objective measures, originally developed for assessing codecs, to assess noise reduction systems [1]. ITU recommendation P.835 [2] provides a good methodology in the subjective evaluation of noisy speech signals by asking listeners to rate the speech and noise quality separately. Using this recommendation listeners rate the quality in three aspects, the speech distortion, the background noise disturbance, and the overall quality.

Musical noise is one of the most significant distortions introduced into the background noise when a noisy signal is processed by a noise reduction system. It is necessary to develop a method to measure the amount of the musical noise after noise reduction. The major work carried out in this area uses the tonality measure of the residual noise [3-5]. In [4], musical noise is detected by finding isolated tones in short-time spectra of signals. In [4], noise masking thresholds [6] are used to find audible components and the noisy speech signal is used as a reference signal to discriminate between musical noise components and speech signal components. In [5], this idea is extended to critical bands and the musical noise is detected by using the tonality coefficient in the critical bands. However, these works mostly focus on the problem of musical noise detection.

In this paper, we propose a non-intrusive objective measure for detecting and evaluating wide-band musical noise in an acoustic signal. In this paper, we use noise masking thresholds to characterize the isolated tones in acoustic signals. Furthermore, we use a harmonicity measure to discriminate between musical noise components and speech components.

This paper is organized as follows. At first, we review the origin and properties of the musical noise. Afterwards, three principal measures are defined to discriminate the musical noise from harmonic signals and white noise. Subsequently, an objective measure is proposed to assess the musicalness of the residual noise. At the end, the performance of the proposed measure is evaluated using subjective listening tests.

2. THE ORIGIN OF THE MUSICAL NOISE

Musical noise is one of the most unpleasant distortions introduced into the background noise when noisy signals are processed by a spectrum modification-based noise reduction system. The musical noise is named so because of transient random single-tones in its short-time spectrum, which produce a distorted music-like sound. The residual musical noise, for an uncorrelated noise, consists of random components in time and frequency with almost a constant rate of generation and degeneration [7].

The musical noise has two major properties; it consists of separate components, randomly distributed over frequencies; and a varying shape of spectrum over time. Using these two properties we propose a method for musical noise detection in three steps,

1. Detecting distinct peaks in the short-time spectrum
2. To discriminate the musical noise from harmonics of a voiced speech signal, it is tested to see if the detected peaks are not harmonics of a fundamental frequency.
3. Inspecting the similarity of consecutive frames to detect the transient spectral components.

In this paper, these steps are implemented by defining three principal measures. These measures are then used to detect the musical noise in the acoustic signals. Finally, an objective measure is proposed to assess the amount of the detected musical noise in the residual noise.

3. DETECTION OF THE ISOLATED TONES

The first step to detect musical noise is detecting isolated tones in the short-time spectrum. Let $a(m,k)$ be the short-time

spectrum magnitude in the mth frame and the kth frequency bin. The ensemble $\mathbf{K}_m$ is defined as

$$\mathbf{K}_m = \left\{ k \middle| a(m,k) > a_{th}(m), a(m,k) > a(m,k-\ell), 0 < |\ell| \le L_a \right\}, \quad (1)$$

whose members are the local maxima of the spectrum in the frame m. In (1) $a_{th}(m)$ are peak detection thresholds which are calculated using means and variances of the frames' spectral magnitudes, and L_a is the maximum search length. The method of determining $a_{th}(m)$ and L_a is described later on. Let the number of the $\mathbf{K}_m$ members be denoted by $\|\mathbf{K}_m\|$. The musical noise components are those spectral peaks with significantly higher magnitudes than the other components. For the pth detected peak of the spectrum in the mth frame a measure of distinctiveness is defined as

$$v_m(p) = \frac{a(m,k_p)}{\max\left\{ \left(\prod_{i=k_p-L_a}^{k_p+L_a} a(m,i) \right)^{\frac{1}{2L_a+1}}, \varphi(m,k_p) \right\}}, \quad (2)$$

in which $k_p \in \mathbf{K}_m$ is the index of the pth ($p = 1,...,\|\mathbf{K}_m\|$) local spectral maximum and $\varphi(m,k_p)$ is the magnitude of the noise masking threshold [6] in the frame m and the frequency bin k_p. The $v_m(p)$ measure shows how much audible the pth isolated spectral peak is. The averaged value of $v_m(p)$ in the frame m shows the average distinctiveness of the frame peaks:

$$f_d(m) = \frac{1}{\|\mathbf{K}_m\|} \sum_{p=1}^{\|\mathbf{K}_m\|} \min\left(v_m(p), v_{\max}\right), \quad (3)$$

in which $v_{\max}$=10 is the maximum allowable value of v_m.

In this paper, all the measures are calculated at the sampling frequency of 16 kHz by using Short-Time Fourier Transform (STFT). The STFT window size is 512 samples with no overlap between consecutive frames. The Discrete Fourier Transform (DFT) length is increased to 1024 samples by using zero-padding.

$a_{th}(m)$ and L_a were chosen such that f_d measure makes maximum discrimination between white Gaussian noise and its musical residual noise. Different amounts of musical noise were generated using spectral subtraction and the f_d measure was calculated. The most discriminations were made using L_a=41 and

$$a_{th}(m) = \underset{k}{mean}\{a(m,k)\} + 0.25 \times \underset{k}{std}\{a(m,k)\}.$$

4. HARMONIC PEAKS DETECTION

The detected isolated peaks should be checked to find any harmonic relationship. At first, a new spectral magnitude is constructed such that, it equals $a(m,k)$ at the locations of the detected isolated peaks and their neighborhoods, and equals zero at the other places. Let this spectral magnitude be denoted by $\hat{a}(m,k)$ in the mth frame and the kth frequency bin. The unbiased autocorrelation sequence of $\hat{a}(m,k)$ in positive frequencies is calculated as

$$R_X(m,l) = \frac{1}{K/2-l+1} \sum_{i=0}^{K/2-l} \hat{a}(m,i+l) \times \hat{a}(m,i), l = 0,..., K/2, \quad (4)$$

where K is the DFT length and l are the autocorrelation lags. The ensemble $\mathbf{\Lambda}_m$, which contains the indices of the local maxima of $R_X(m,l)$, is defined as

$$\mathbf{\Lambda}_m = \left\{ \lambda \middle| R_X(m,\lambda) > R_{th}, R_X(m,\lambda) > R_X(m,\lambda-\ell), |\ell| \le L_R \right\}, \quad (5)$$

where L_R and R_{th} are the maximum search length and the peak detection threshold, which are set to 11 and 0.01, respectively. The number of the $\mathbf{\Lambda}_m$ members is denoted by $\|\mathbf{\Lambda}_m\|$. It is assumed that λ are sorted increasingly in $\mathbf{\Lambda}_m$, such that the following condition for the pth member of $\mathbf{\Lambda}_m$ is satisfied:

$$\forall p > p_0, \qquad \lambda_p > \lambda_{p_0}, \quad p_0 = 1,...,\|\mathbf{\Lambda}_m\|. \quad (6)$$

Using this assumption, the distances between $R_X(m,l)$ peaks are calculated as

$$\Delta_m(p) = \lambda_p - \lambda_{p-1}, \quad \lambda_p \in \mathbf{\Lambda}_m, \quad p = 2,...,\|\mathbf{\Lambda}_m\|. \quad (7)$$

If the frame under study contains harmonic peaks, $\Delta_m(p)$ becomes almost constant and its standard deviation becomes very close to zero. Thus, the frames with harmonic components can be distinguished from the frames which are mixtures of some non-harmonic components. To study $\Delta_m(p)$ its normalized standard deviation and mode percentage are defined as

$$\sigma_{\mathrm{P}}(m) = std\{\Delta_m(p)\} / mean\{\Delta_m(p)\}, \quad (8)$$

$$\theta_{mod}(m) = N_{mod}(m) / \|\mathbf{\Lambda}_m\|, \quad (9)$$

respectively, where N_{mod} is the number of the instances of the mode (most frequent value) of $\Delta_m(p)$. N_{mod} becomes close to $\|\mathbf{\Lambda}_m\|$ for harmonic signals. As a result $\theta_{mod}(m)$ becomes close to one for harmonic signals and it is less than one for non-harmonic signals. These two measures are combined to define a measure with low error in discriminating harmonic from non-harmonic signals:

$$f_h(m) = \frac{\min\left(\sigma_{\mathrm{P}}(m), thr_\sigma\right)}{thr_\sigma} \times \frac{1 - \max\left(\theta_{mod}(m), thr_\theta\right)}{1 - thr_\theta}, \quad (10)$$

where thr_σ and thr_θ are set to the mean values of σ_{P} and θ_{mod} for white Gaussian noise, i.e. 0.2 and 0.3, respectively. For harmonic signals (like a voiced speech) when isolated peaks are detected, the values of $f_h(m)$ become very close to zero. In contrast, the values of $f_h(m)$ are far from zero for non-harmonic signals. As a result, $f_h(m)$ is used to discriminate between the musical noise and speech.

5. CONSECUTIVE FRAMES SIMILARITY

As mentioned before, one of the major properties of musical noise is frame-to-frame variations in its short-time spectrum. The spectral similarity between the frames i and j is defined as

$$S_a(i,j) = \frac{\sum_{k=1}^{K/2} a(i,k) \times a(j,k)}{\sqrt{\sum_{k=1}^{K/2} a^2(i,k)} \sqrt{\sum_{k=1}^{K/2} a^2(j,k)}}. \quad (11)$$

And the spectral regularity in the frame m is defined as

$$f_s(m) = \max\left(S_a(m,m-1), S_a(m,m+1)\right). \quad (12)$$

The average value of f_s for musical noise is much less than the values for speech and white noise. Consequently, f_s is used in combination with f_d and f_h to discriminate the musical noise from the speech and white noise.

Table 1. C_{MN} Statistical Properties

	Frames	C_{MN}		
		Mean	Trimmed Mean	Median
Speech	622	0.023	0.009	0
White Noise	3375	0	0	0
Musical (Lo)	3375	0.923	0.945	1.000
Musical (Hi)	3375	0.938	0.956	1.000

6. MUSICAL NOISE DETECTION

f_d, f_h and f_s are combined to discriminate the musical noise from harmonic speech and white noise as

$$C_{MN}(m) = \hat{f}_h(m) \cdot \hat{f}_d(m) \cdot (1 - \hat{f}_s(m)), \quad (13)$$

where $\hat{f}_d$, $\hat{f}_h$ and $\hat{f}_s$ are the mapped values of f_d, f_h and f_s into the range between zero and one. C_{MN} statistics are shown in Table 1. The statistics for speech are calculated for twelve 3-second long utterances on the speech frames. The *Lo* and *Hi* tags in Table 1 correspond to the musical noise signals with the low and high levels of the musical noise respectively. The level of the musical noise is determined in the subjective listening tests described in section 8. It is concluded that the musical noise can be effectively discriminated from white noise and speech by comparing the mean value of C_{MN} with a constant threshold, e.g. 0.2.

7. MUSICALNESS MEASURE

Once the musical noise is detected using C_{MN}, its musicalness can be determined using the f_d and f_s measures. We define the musicalness measure as

$$F_M(m) = {f_d}^{\gamma(m)}(m) \cdot \left(\frac{\max(thr_{s,w} - f_s(m), 0)}{thr_{s,w}} \right)^{1-\gamma(m)}, \quad (14)$$

where $thr_{s,w}$ is the white noise threshold for f_s measure and it is set to the mean value of f_s for white noise, i.e. 0.8. γ is a parameter that considers how much f_d is involved in the calculation of F_M. It is calculated as $\gamma(m) = 0.375 + 0.125 \tanh\left(2 \times \|\mathbf{K}_m\| - 4\right)$. γ reduces the role of f_d in calculating F_M when the number of the detected isolated spectral peaks is less than 3.

8. EXPERIMENTAL EVALUATION

In this section the performance of the musicalness measure (F_M) is evaluated using the residual noise of power spectral subtraction. Spectral subtraction is used because it naturally produces musical noise and the amount of the musical noise can be easily controlled by two parameters. The power spectral subtraction for generating musical noise is formulated as

$$|N_r(m,k)|^2 = \max(|N(m,k)|^2 - osc \times \sigma_N^2(k), flrco \times |N(m,k)|^2) \quad (15)$$

in which $N(m,k)$ and $N_r(m,k)$ are the short-time spectra of the noise and the residual noise, respectively. In (15) $\sigma_N(k)$ represents the standard deviation of the noise spectrum in the kth frequency bin. *osc* and *flrco* are over-subtraction and floor coefficients, respectively. These coefficients are used to control the amount of the residual noise as well as its musicalness. White Gaussian noise, taken from NOISEX-92 database and re-sampled to the sampling frequency of 16 kHz, is used in experiments.

Table 2. Musicalness Subjective Scores

Score	Musicalness Level
0	No Musical
1	Very Low
2	Low
3	Medium
4	High
5	Very High

Two separate listening tests were constructed to examine how well F_M measures the amount of the perceived musical noise. 10 listeners participated in the tests. The tests are about 5-7 minutes long with a break of 3-5 minutes between them. The break is used to familiarize the listeners with the second test. The test signals are the residual noise signals of white Gaussian noise after spectral subtraction; each with the length of 48000 samples (3s). The test signals were played by headphone to the listeners with a 0.5 to 1 second gap between them. The listeners were asked to only comment on the musicalness of the residual noise. To avoid the contribution of the signal power to the listeners' opinions, all the signals were normalized to the active level of -26dBov using ITU-T P.56 [8] standard.

The performance of the proposed objective measure in predicting the amount of the perceived musical noise were evaluated using the Pearson product-moment correlation coefficient, defined as

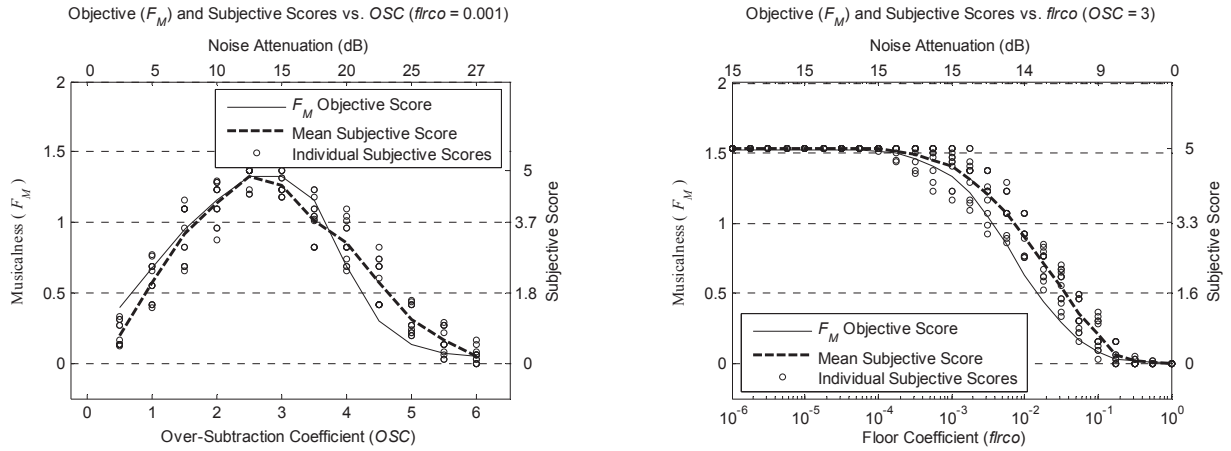
$$\rho = \frac{\sum_{m=1}^{M}(X_m - \overline{X})(Y_m - \overline{Y})}{\sqrt{\sum_{m=1}^{M}(X_m - \overline{X})^2 \sum_{m=1}^{M}(Y_m - \overline{Y})^2}}, \quad (16)$$

in which X_m are the subjective scores and Y_m are the corresponding objective scores. $\overline{X}$ and $\overline{Y}$ are the mean values of X_m and Y_m, respectively.

8.1 Musicalness vs. over-subtraction coefficient

In the first test, the increasing and decreasing behavior of F_M by varying *osc* was studied. In Figure 1(a) the average values of the F_M measure are plotted versus *osc* for *flrco*=0.001 (solid line). The values of F_M in this plot correspond to the *osc*=0.5 to *osc*=6. First, the residual noise signals were played for the listeners in the increasing order of the *osc* and they were asked to comment on the increasing or decreasing behavior of the musical noise. In fact, all listeners stated the amount of the musical noise increased until it reached a maximum point and then decreased. They were requested to find out which signal is the most musical one. The purpose of this stage was not only finding the most musical signal, but also familiarizing the listeners with the test signals.

In the second stage, listeners were instructed to rate the musicalness of the signals on a scale of 0 to 5 as described in Table 2. The maximum musical noise was scored 5 and the reference white noise was scored 0. The average of the subjective scores is plotted in Figure 1(a) by the dashed thick line. It is seen that the F_M values are very close to the average subjective scores. In this experiment, the correlation coefficient between the objective and the subjective scores becomes as high as 0.96. The peak of the curve was perceived at the *osc* of 2.5, by the most of the listeners. In Figure 1(a), the amount of the attenuation is also shown on the top axis. As expected, noise attenuation increases when the *osc* is increased.



(a) (b)

Figure 1. The musicalness of the spectral subtraction residual noise (a) for different value of over-subtraction coefficient at floor coefficient of 10^{-3} and (b) for different values of floor coefficient at over-subtraction coefficient of 3. (The mean objective score (F_M) is the solid thin line and the mean subjective score is the thick dashed line.)

8.2 Musicalness vs. subtraction floor coefficient

In Figure 1(b), the average values of the F_M measure are plotted versus *flrco* at *osc*=3 (solid line). In the second listening test, the residual noise signals for *flrco*=10^{-6} to 1 were played for the listeners. All the listeners stated that as the value of *flrco* decreased the amount of the musical noise increased and no more increase could be perceived after a certain range of *flrco*. As a matter of fact the listeners were asked to find the saturation point and give the score of 5 to all the signals that consequently had the same quality.

Afterwards, listeners were instructed to rate the remaining noise signals on the scale of Table 2. To reduce the listeners' inaccuracies they were instructed to find the mid-point first and then rate the other signals. The mean subjective score is plotted in Figure 1(b) by the dashed thick line. The difference between the subjective and the objective scores results from the fact that some listeners detected the saturation point at a point different from that of the objective scores curve. This is natural and it is because of the different sensitivity of the different people's ears. This difference does not affect the final correlation coefficient between the objective and subjective scores. The Pearson product correlation coefficient for this experiment becomes as high as 0.986 which shows the merit of the very high performance. It is necessary to note that the tails of the curves are excluded in the correlation coefficient calculation; otherwise the correlation coefficient would have become 0.989.

9. CONCLUSION

In this paper a novel method for detecting and evaluating the musical noise in the acoustic signal was proposed. The proposed method used three features, extracted from the short-time spectrum, in order to detect and evaluate the musical noise. These features included the distinctiveness of the isolated non-harmonic spectral peaks, and the similarity between consecutive frames spectra. The proposed objective measure used these measures to predict the amount of the perceived musical noise in the ear. In order to evaluate the proposed objective measure, subjective listening tests were performed using the residual noise of white Gaussian noise in a spectral subtraction noise reduction system. In the experiments, the correlation coefficient between the objective and the subjective test results became higher than 0.96, which shows the very high correlation of the proposed objective measure with the perceived subjective quality of the residual noise.

10. ACKNOWLEDGEMENT

The authors sincerely thank all those who patiently participated in the listening tests and made a big contribution to this work by their precious opinions.

REFERENCES

[1] Y. Hu, and P. Loizou, "Evaluation of Objective Measures for Speech Enhancement," in *Proc. of INTERSPEECH'06,* pp. 1447-1450, Pittsburg, PA, 17-21 Sep. 2006.

[2] ITU-T P.835, *Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithms,* ITU-T Recommendation P.835, Nov. 2003.

[3] P. Dreiseitel, and G. Schmidt, "Evaluation of algorithms for speech enhancement," in *Topics in Acoustic Echo and Noise Control,* E. Hänsler and G. Schmidt (Eds.), Berlin: Springer-Verlag, 2006, pp. 431-484.

[4] S. Ben Jebara, "A perceptual approach to reduce musical noise phenomenon with wiener denoising technique," in *Proc. of IEEE Int. Conf. on Acoust. Speech & Signal Processing (ICASSP'06),* vol. 3, pp. 49-52, Toulouse, France, 14-19 May 2006.

[5] A. Ben Aicha, and S. Ben Jebara, "Perceptual musical noise reduction using critical bands tonality coefficients and masking thresholds," in *Proc. of INTERSPEECH'07,* pp. 822-825, Antwerp, Belgium, 27-31 August, 2007.

[6] J. D. Johnston, "Transform coding of audio signal using perceptual noise criteria," *IEEE J. Select. Areas Commun.,* vol. 6, pp. 314–323, February 1988.

[7] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Tran. Audio Speech & Signal Processing,* vol. 27, no. 2, pp. 113-120, April 1979.

[8] ITU-T P.56, Objective Measuring Apparatus, Objective Measurement of Active Speech Level, ITU-T Recommendation P.56, March 1993.