# ANALYSIS OF SOUND TEXTURES IN MUSICAL AND MACHINE SOUNDS BY MEANS OF HIGHER ORDER STATISTICAL FEATURES.

*Shlomo Dubnov*[1] *Naftali Tishby*[2]

[1] Analysis/Synthesis Team, IRCAM, Paris 75004, France
[2] Institute for Computer Science, Hebrew University, Jerusalem 91904, Israel

## ABSTRACT

In this paper we describe a sound classification method, which seems to be applicable to a broad domain of stationary, non-musical sounds, such as machine noises and other man made non periodic sounds. The method is based on matching higher order spectra (HOS) of the acoustic signals and it generalizes our earlier results on classification of sustained musical sounds by higher order moments. An efficient "decorrelated matched filter" implemetation is presented. The results show good sound classification statistics and a comparison to spectral matching methods is also discussed.

## 1. INTRODUCTION

Sound textures can be considered as stationary acoustical phenomena that obtain their acoustical effect from internal variations in the sound structure, such as micro-fluctuations in the harmonics of a pitched sound or statistical properties of a random excitation source in an acoustic system. Signal acoustic models usually describe the structure of slowly varying partials such as the overall structure induced by spectral envelope of resonant chambers in musical instruments or the formant filters in speech. Besides these long-time ($\sim$ 50ms)) characteristics there are short-time ($\sim$ 10ms) fluctuations in frequency that contribute significantly to the timbre of a pitched (voiced) sound by effecting the sound harmonicity and coherence. In noise-like sound such as engine noise and other man made un-harmonic signals, the characterizing spectrum is continuous and a significant component of the sounds structure is due to features of the probability distribution function of the prewhitened signal.

The first step beyond analyzing the spectral amplitude distribution is to look at the statistical properties of the excitation signal. The effect of the spectral envelope can be removed by inverse filtering of the signal, thus obtaining a spectrally flat (white) *residual* signal. The higher order statistical (HOS) properties of this residual are closely related to the non-linearities of the excitation source [4][1]. We argued that for musical signals this effect can be modeled as a frequency modulating jitter of the harmonics [2]. In this work we would like to consider a unifying scheme were the bispectral and trispectral signatures of the sounds serve as features for analysis and classification of pitched (musical) signals and noise-like (machine) sounds with continuous spectrum.

## 2. MOMENTS SPACE REPRESENTATION

In an earlier work we have analyzed a set of sounds of musical instruments, by looking at the higher order moments of decorrelated signals. The moments of order $k$ in the time domain are related to the integral of the $k$-th order spectra in the frequency domain. For instance, for $k = 3$

$$m_3 = \lim_{T\to\infty} \int_0^T x^3(t)dt \approx \int\int B_x(\omega_1, \omega_2)d\omega_1 d\omega_2 \quad (1)$$

Using a source-filter model, the decorrelated signal of a stable periodic sound can be regarded as a stochastic version of a pulse train, with variations occurring due to some random frequency jitter applied to its harmonics. By looking at the *skewness* $\gamma_3 = m_3/\sigma^3$ of the signal and the *kurtosis* $\gamma_4 = m_4/\sigma^4$ we have shown that these statistics measure the amount of harmonicity among triplets and larger groups of partials apparent in the signal [2].

For musical sounds the moments show a clear distinction between string, woodwind and brass sounds. Representing the sounds as coordinates in 'moments space' locates the instrumental groups on 'orbits' with various distances around the origin, very much according to the traditional, orchestration handbook practice.

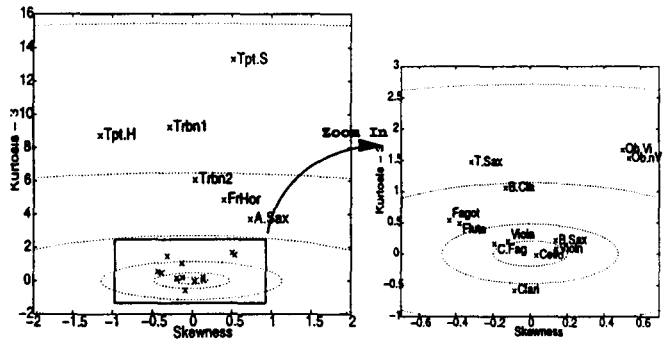


**Figure 1. Location of sounds in the 3rd and 4th normalized moments plane. The value 3 is subtracted from the kurtosis so that the origin would correspond to a perfect Gaussian signal. Brass sounds are on the perimeter. Strings are in the center.**

In this work we try to extend the above results to deal with non periodic sounds, specifically dealing with sounds derived from machine and other man made noises. For the noise-like signals, we have investigated 11 recordings

of sounds such as people talking in a room, engine noises of various kinds, recordings of factory sounds and etc. Since *skewness* is zero for symmetric signals[1], we ignored the 3rd order moment. In figure 2 we show the kurtosis values of the 11 different sound types, estimated using 2 sec. long segments. The moments space representation locates pulse like, "jagged" or "rough" sounds, such as talking voices (babble), machinery crashes (fac1, fac2 - factory noises) and noisy engines (m109, dops - destroyer operations room recording) to be higher in the moments space, while the "smoothly" running engines being near the origin (kurtosis = 3), and thus closer to a Gaussian model. As one can see, the variance of the estimate is significant and does not allow for detailed discrimination between sounds.

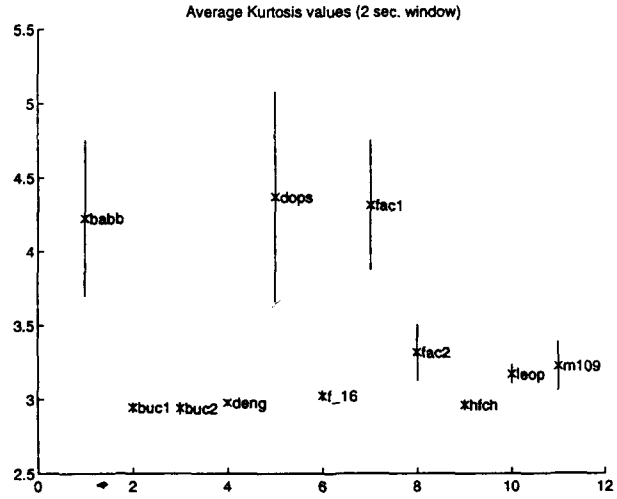


**Figure 2. Kurtosis mean values and standard deviation for the various samples. Sounds with many transitory elements have higher kurtosis values. Smooth sounds are closer to 3.**

## 3. A COMPLETE HOS FEATURE SPACE REPRESENTATION

The 'moments space' representation, although computationally attractive, actually detects only the average value of the signals polyspectral planes. In order to get more information out of the polyspectral contents of a signal, we would like to consider further features based on HOS.

As will be explained in the following, we are interested in the HOS of decorrelated signals. The differences in bispectral contents of decorrelated signals can be visually observed in figure (3). These figures visually demonstrate the bispectral signatures for the following sounds: recordings of a people talking (babble), "crashing" factory noise (fac2) and a rather "smooth" buccaneer engine noise (buc1).

One can see that the bispectral amplitudes of the signals differ significantly.

The matching in polyspectral domain is performed by

[1] i.e. signals whose amplitude distribution is symmetric with respect to the mean.

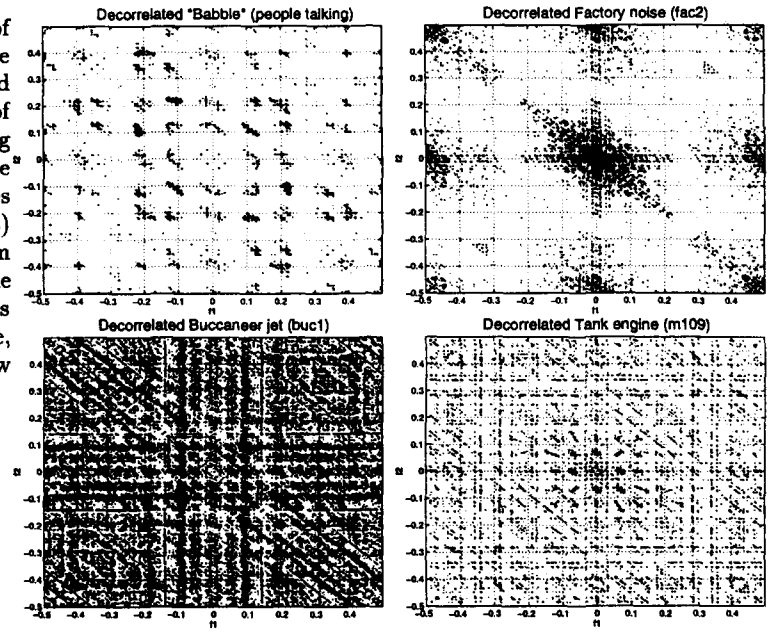


**Figure 3. Bispectral amplitudes for people talking, car production factory noise, buccaneer engine and m109 tank sounds (left to right).**

using a Maximum Likelihood (ML) classifier [3]

$$d(y,x) = -\int\int \frac{|B_x(\omega_1,\omega_2) - B_y(\omega_1,\omega_2)|^2}{S_x(\omega_1)S_x(\omega_2)S_x(\omega_1+\omega_2)} d\omega_1 d\omega_2 \qquad (2)$$

which is an expression for likelihood of seeing signal $y$ under the hypothesis $x$ (i.e. a system which has power-spectra $S_x$ and bispectra $B_x$).

The mathematical assumptions behind this classifier are: 1) Gaussianity of the bispectral estimator, which is observed under some mild mixing conditions of the random process; 2) for discrete signals, when assuming an ARMA model, a dense enough sampling of the frequency domain ensures a unique classification.

### 3.1. Implementation Notes

Our implementation of the polyspectral matching is based on the equivalence between bispectrum of decorrelated signal and a "spectrally normalized" version of bispectrum, called usually *bicoherence* function

$$b(\omega_1,\omega_2) = \frac{B_x(\omega_1,\omega_2)}{\sqrt{S_x(\omega_1)S_x(\omega_2)S_x(\omega_1+\omega_2)}} \qquad (3)$$

We summarize this in the following Lemma,

*Lemma*: For non-Gaussian, linear process $x(t)$, the bicoherence function $b_x(\omega_1,\omega_2)$ is equal to the bispectrum $B_{\tilde{x}}(\omega_1,\omega_2)$ of a decorrelated signal $\tilde{x}_{S_x}$. The decorrelation is done by inverse filtering of $x(t)$ through a filter that has power-spectrum equal to the spectrum $S_x(\omega)$ of the original signal $x$.

*Proof:*

Assuming the signal $x(t)$ is described by a non-Gaussian,

independent and identically distributed (i.i.d.) excitation signal $u(t)$ passing through a linear filter $h(t)$.

$$x(t) = \int_{-\infty}^{\infty} h(t')u(t-t')dt' \quad , \tag{4}$$

the spectra and bispectra of $x(t)$ are given by

$$\begin{aligned} S_x(\omega) &= \mu_2|H(\omega)|^2 \\ B_x(\omega_1,\omega_2) &= \mu_3 H(\omega_1)H(\omega_2)H^*(\omega_1+\omega_2) \end{aligned} \tag{5}$$

(with a similar equation for the trispectra), with $\mu_2$, $\mu_3$ (and $\mu_4$) being the second, third (and fourth) order cumulants of $U_t$. For convenience we shall assume $\mu_2 = 1$. Taking a spectrally matching filter $\hat{H}(\omega)$

$$S_x(\omega) = |\hat{H}(\omega)|^2 \tag{6}$$

the inverse filtering of $x(t)$ with $\hat{H}^{-1}(\omega)$ gives a residual $\tilde{x}(t)$. The bispectrum of $\tilde{x}(t)$ is

$$B_{\tilde{x}}(\omega_1,\omega_2) = \frac{B_x(\omega_1,\omega_2)}{\hat{H}(\omega_1)\hat{H}(\omega_2)\hat{H}^*(\omega_1+\omega_2)} \quad . \tag{7}$$

Using the spectral matching property of $\hat{H}$, we arrive at the equivalence of the right-hand side of the above equation and the definition of the bicoherence function of the original signal $x(t)$.
□

*3.1.1. "Matched filter" implementation*

One of the biggest problems with using polyspectral features for signal matching is the long signal duration needed for averaging the estimates in order to overcome their variance. This might be extremely demanding in terms of the memory and computation power requirements. In our implementation, we have used a "matched filter" variant, which is described below, to do the classifier calculation.

Let us look at a signal $z(t)$ which is the result of a convolution between two signals $x(t)$ and $y(-t)$.

$$z(t) = x(t) \otimes y(-t) \tag{8}$$

In case where $x(t)$ and $y(t)$ are statistically independent, the bispectrum of $z(t)$ is

$$B_z(\omega_1,\omega_2) = B_x(\omega_1,\omega_2) \cdot B_y^*(\omega_1,\omega_2) \tag{9}$$

On the other hand, if $y(t) = x(t)$, one can show that

$$B_z(\omega_1,\omega_2) > |B_x(\omega_1,\omega_2)|^2 \tag{10}$$

Moreover, it is important to recall that $k$-th order order moments are equal to the integral over the $k$-spectra. Using these two properties and Lemma, we implemented a variant of equation (2)

$$\begin{aligned} \tilde{d}(y,x) &= \int\int |B_{\tilde{x}}(\omega_1,\omega_2) - B_{\tilde{y}}(\omega_1,\omega_2)|^2 \\ &\leq \int (\tilde{x}(t) \otimes \tilde{x}(-t))^3 dt - 2\int (\tilde{x}(t) \otimes \tilde{y}(-t))^3 dt \\ &+ \int (\tilde{y}(t) \otimes \tilde{y}(-t))^3 dt \end{aligned} \tag{11}$$

Note that $y(t)$ is decorrelated by a filter that matches $S_x$.

As we have already mentioned, the skewness is zero for symmetric signals. An analogous derivation holds for trispectra, and in our simulations we have used powers of four in the moments calculation.

## 4. RESULTS

The data that we have used in our simulations comes from a Signal Processing Information Base (SPIB), at http://spib.rice.edu/spib/select_noise.html.

For simulations we have used 11 different recordings (sound files), with 1 template and 25 test segments taken from each file. The first 200 milliseconds of each file were used as a training segment (template). The test segments (200 msec. each) where drawn, at 1 second time lags, from the same recordings, starting from time of two seconds. Thus, we had in total 11 classes of sounds represented by 11 templates and a test set of 25 examples of each class.

The classification results for 25 test samples of each class are summarized in the table 1.

## 5. DISCUSSION

A natural feature for matching sound signal is the spectrum. In order to evaluate the spectral similarity among the signals, we have looked at the coherence function

$$C_{xy}(\omega) = \frac{|P_{xy}(\omega)|^2}{P_{xx}(\omega) \cdot P_{yy}(\omega)} \tag{12}$$

where $P_{xx}(\omega)$, $P_{yy}(\omega)$ are the Power Spectral Densities of $x(t)$ and $y(t)$ respectively, and $P_{xy}(\omega)$ is the Cross Spectral Density of $x$ and $y$. In figure(4) we show the integral over all frequencies, of the coherence function calculated between various test set samples and a 'babb' template. The coherence function was taken with fft of order 512, which gives spectral resolution of approximately 40 Hz.

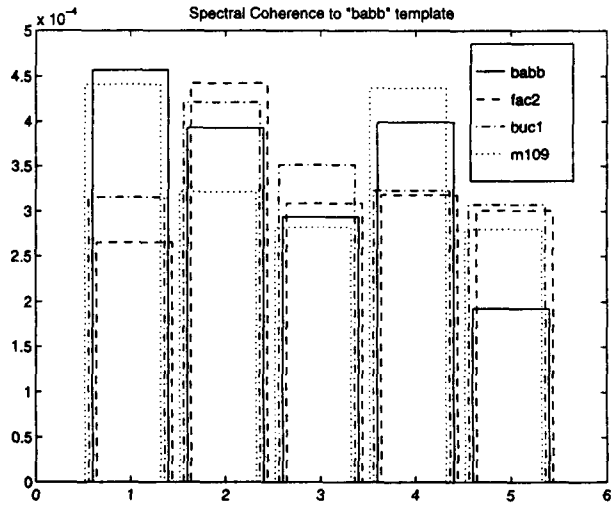


**Figure 4. Integral of spectral coherence function between five samples of four test classes (total 24 test samples) and a 'babb' template. Note that the result for test classes 'babb' is not better then for the other classes.**

| | babb | buc1 | buc2 | deng | dops | f_16 | fac1 | fac2 | hfch | leop | m109 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| babb | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| buc1 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| buc2 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| deng | 1 | 0 | 0 | 19 | 0 | 0 | 0 | 0 | 5 | 0 | 0 |
| dops | 0 | 0 | 0 | 0 | 24 | 0 | 1 | 0 | 0 | 0 | 0 |
| f_16 | 0 | 0 | 0 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 |
| fac1 | 0 | 1 | 0 | 0 | 0 | 3 | 17 | 0 | 0 | 4 | 0 |
| fac2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 22 | 0 | 2 | 0 |
| hfch | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 | 0 |
| leop | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 |
| m109 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 |

**Table 1. Classification results for order 4 (trispectral) matching. The set contains 11 classes of signals of 25 test segments each.**

The results show a very poor coherence of a 'babb' template to the various test signals (order of $10^{-4}$). Moreover, no preference exists for matching the test samples of the 'babb' class compared to other test signals.

Testing a matched filter classifier on the same set of signals[2] we find unsatisfactory results for most of the signals ( $\leq$ 44% classification rate for seven signals, 64% for 'm109' and 100% for 'buc2', 'dops' and 'hfch').

The correct classification probability depends on the credibility of the features. Naturally, longer segments give less variance of the estimators and improve the classification results. We summarize the dependence of the classification probability on segment length in figure 5. Due to the large fluctuations in the classification results for short segment sizes, we plot an order two regression graph for times shorter then 0.2 seconds.

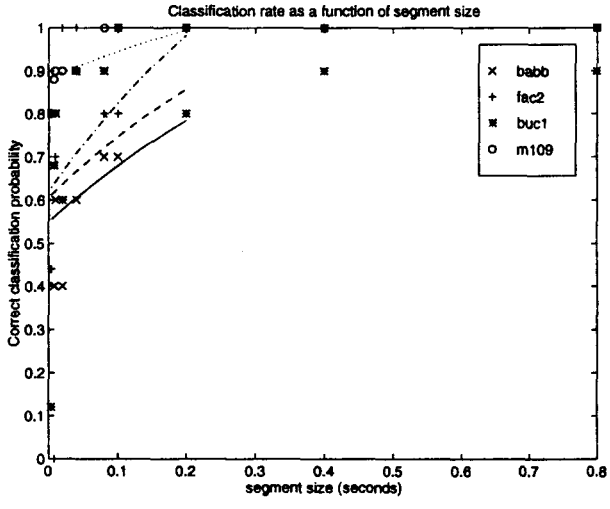


**Figure 5. The correct classification probabilities as afunction of segment length, tested for four signals. See text for more details.**

[2] This classification is done by choosing a template which maximizes the square sum (energy) of the convolution signal between the original (not decorrelated) template and an original test segment.

## 6. SUMMARY AND CONCLUSION

In this work we have presented a method for sound classification which is based on matching HOS of acoustic signals. The method was applied to sounds from diverse sources, such as machine noises and other non periodic, stationary, man made sounds. This type of acoustical phenomena seem to acquire their acoustical effect to a large extend due to the micro-fluctuations of their spectral components. These fluctuations seem to be successfully detected by polyspectral analysis methods.

We have also presented an efficient implementation of the polyspectral matching by means of a matched filter between decorrelated signals. The results show that polyspectral features are superior in their discriminatory properties to the power spectral features. This matching requires the knowledge of 1) spectral envelope filter coefficients and of 2) vector of decorrelated signals samples, which is used as a template for the decorrelated matching.

### Acknowledgments

The authors are most grateful to Xavier Rodet for the many important remarks and fruitful discussions of this work.

## REFERENCES

[1] S.Dubnov, N.Tishby *Testing for Non linearity and Gaussianity in sustained portion of musical signals*, Proceedings of the Journees d'Informatique Musicale, Caen, 1996.

[2] S.Dubnov, N.Tishby, D.Cohen, *Influence of frequency modulating jitter on the higher order moments of musical signals*, Proceeding of the International Computer Music Conference, Hong-Kong 1996.

[3] G.B.Giannakis and M.K.Tsatsanis, *A Unifying Maximum-Likelihood View of Cumulant and Polyspectral Measures for Non-Gaussian Signal Classification and Estimation*, IEEE Transactions on Information Theory, Vol.38, No.2, march 1992.

[4] M.J. Hinich, *Testing for Gaussainity and Linearity of a Stationary Time Series*, Journal of Time Series Analysis, Vol. 3, No.3, 1982.