STATISTICAL, SPECTRAL AND STOCHASTIC CHARACTERISTICS OF MUSIC

Declan Quinn and Jacqueline Walker

Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland
phone: + (353) 61-202780, fax: + (353) 61-338176, email: jacqueline.walker@ul.ie

ABSTRACT

The goal of this paper is to explore some of the principal spectral and statistical characteristics of music that are relevant to the problem of Blind Source Separation (BSS). Some BSS algorithms require some *a priori* knowledge of statistical characteristics of the mixtures under examination, if only to more accurately establish initial estimates. Furthermore, theoretical investigations depend upon the assumption of Wide-Sense Stationarity - the extent to which this assumption holds is investigated.

1. INTRODUCTION

The problem of Blind Source Separation has received much attention in recent years. Investigations, however, have typically been focused on the separation of mixtures of speech, with less attention paid to the separation of mixtures of music. The focus of this paper is to investigate some of the characteristics of musical samples pertinent to the BSS problem, and so facilitate the application of BSS techniques to these mixtures of lower temporal and spectral separation (due to the tendency of musical instruments to be played in a coherent manner).

2. MATHEMATICAL DEFINITIONS

Definition 1 *Given the expectation operator $E(.)$, a stochastic process is* **wide-sense stationary (WSS)** *if:*

1. *the mean function $m_x(t)$ of the process is a constant $m_x \forall t$.*
2. *the autocorrelation function is independent of time, and depends only on a time-shift τ*

$$E\{x(t)x(t-\tau)\} = r_x(\tau)\,\forall t$$

3. *the mean-square value $r_x(0) = E\left\{x(t)^2\right\}$ of the process is constant.*

Denoting $\bar{x}$ as the mean and σ as the standard deviation, respectively, of a sequence of discrete values, we define:

Definition 2 *The* **skewness** *of a sequence of N discrete values $x = \{x_i\}_{i=1}^N$ is given by*

$$Skew(x) = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{x_j - \bar{x}}{\sigma}\right)^3$$

Definition 3 *The* **kurtosis** *of a sequence of N discrete values $x = \{x_i\}_{i=1}^N$ is given by*

$$Kurt(x) = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{x_j - \bar{x}}{\sigma}\right)^4 - 3$$

The kurtosis is thus 0 for a normal (i.e. Gaussian) distribution; historically, this quantity has been referred to as the **kurtosis excess** - following [1], many authors now take this as the definition of kurtosis.

3. STATISTICAL CHARACTERISTICS

In order to investigate the typical characteristics of music, a large number of samples were extracted from a corpus of commercially available CD recordings (listed in Appendix A) and subject to analysis for this and the subsequent sections. The excerpts were each $10s$ in length (for a total of 5320 samples, each sampled at $44.1kHz$), cover a wide variety of musical styles, and include both live recordings and heavily-mixed studio recordings.

Shown in Figures 1 and 2 are, respectively, the skewness and kurtosis values of sequential 10-second samples drawn from all excerpts from all the tracks of the albums listed in Appendix A.

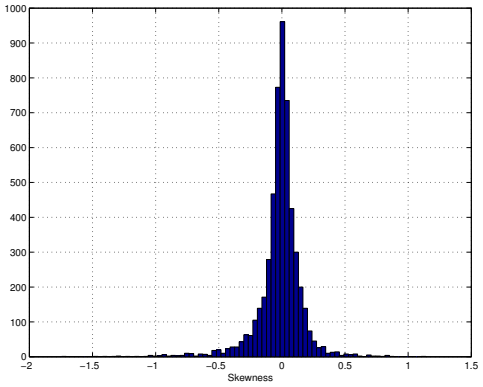

Figure 1 - Histogram of skewness values from music samples.

As may be seen from Figure 1, music tends to be slightly negatively skewed (with an average skewness of -0.0149), whereas Figure 2, displaying an average kurtosis of 1.9214, indicates that the samples are strongly **leptokurtic** or **supergaussian** (i.e. have kurtosis greater than 0). For certain iterative ICA algorithms, it is useful to know whether the sample under investigation is **leptokurtic** or **platykurtic** (i.e. **subgaussian**) in order to more efficiently set the initial estimate of a parameter (see [5]).

4. SPECTRAL CHARACTERISTICS

Shown in Figure 3 are percentile plots of signal power versus frequency for the music samples. The percentile plots, which do not depend on any assumptions upon the statistical properties of the underlying data, provide a robust indication of

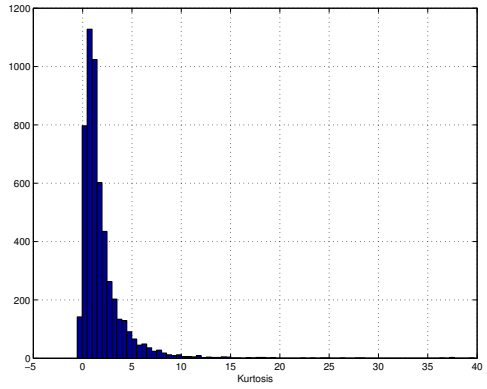

Figure 2 - Histogram of kurtosis values from music samples.

how signal-power is typically spread over frequency bands for typical musical recordings. The power spectrum was determined using Welch's overlapped periodogram method (with a window length of $64\,ms$).

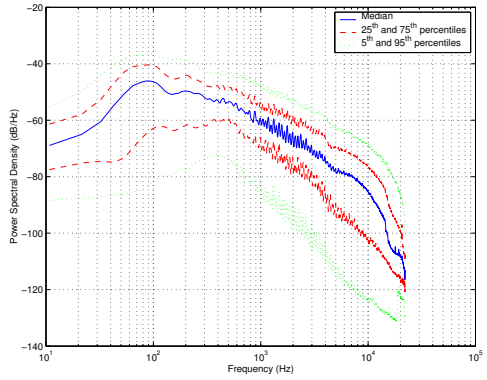

Figure 3 - Median and 5^{th}, 25^{th}, 75^{th} and 95^{th} percentiles of power-spectral densities of music samples (computed using Welch's averaged, modified periodogram method).

As may be clearly seen, the bulk of the signal power is contained within the $20Hz - 1kHz$ range. This suggests that it is reasonable to apply decimation of the input data as a pre-processing step when performing any type of audio analysis in cases where the number of samples to be analysed is excessive (so long as a high-fidelity reproduction is not required directly from the processed data). Should samples under investigation be so decimated, it is suggested that the samples first be appropriately low-pass (or band-pass) filtered prior to selection of a subset of the data. Furthermore, the filter should be applied in both a forward and a backward sense in order to ensure that there is no phase-shifting.

As one may expect, speech is concentrated in a narrower spectral range than music (corresponding to the more limited range of frequencies that the human voice can produce). Examination of median PSD values of Figure 3 shows that, in the case of music, the frequency range $35 - 1000\,Hz$ lies above a threshold of $-60\,dB$. Similar analysis of a number of speech samples (taken from the TIMIT archive [3], not shown) suggests a corresponding band, for human speech, of $100 - 700\,Hz$.

5. WIDE-SENSE STATIONARITY

The assumption of Wide-Sense Stationarity (also referred to as **Weak-Sense Stationarity**) is widely employed in signal processing algorithms; essentially, it permits treatment of time-series entirely in the frequency domain. Thus, whether this assumption is approximately true of music is of interest. In most studies that have been made of audio streams (both music and speech), the stationarity of the underlying process is assumed. In only a few studies in the audio domain (see, e.g. [7] and [6]) has the non-stationarity of samples been investigated; non-stationarity has been more extensively investigated in other applications of ICA (e.g. in the biomedical sciences [9]).

Following [2], one may define a measure of the extent to which a signal is Wide-Sense Stationary in terms of elementary statistics[1].

Splitting a data-set into S segments of equal length N, the means $\hat{\mu}_i$ and variances $\hat{\sigma}_i^2$ may be determined for $i \in \{1,\ldots,S\}$.

For two blocks i and j, the T and F tests are, respectively, the statistical tests for the equality of means and variances:

$$T = (\hat{\mu}_i - \hat{\mu}_j)\sqrt{\frac{N-1}{\hat{\sigma}_i^2 + \hat{\sigma}_j^2}}$$

$$F = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_j^2}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ signify the *estimates* of the mean and variance respectively. T has a Student-t distribution with $\nu = 2N-2$ degrees of freedom and F has an F distribution with $\nu_1 = \nu_2 = N-1$ degrees of freedom (see [10] and [1]). For each of the tests, we use a two-tailed test (as we are checking for a difference, and not a particular direction).

Taking

$$p_{ij} = \begin{cases} 1, & \text{if } |T| \le t_{\nu,\frac{\alpha}{2}} \text{ and } F_{\nu_1,\nu_2,1-\frac{\alpha}{2}} \le F \le F_{\nu_1,\nu_2,\frac{\alpha}{2}} \\ 0, & \text{otherwise} \end{cases}$$

where $t_{\nu,\frac{\alpha}{2}}$ is the critical value for the T test, and $F_{\nu_1,\nu_2,1-\frac{\alpha}{2}}$ and $F_{\nu_1,\nu_2,\frac{\alpha}{2}}$ are the critical values for the F test; both test statistics are computed at a confidence level of α. We arrive at

Definition 4 *The* **Wide-Sense Stationarity Quotient** *is defined as*

$$W_N = \frac{2}{S_N(S_N-1)} \sum_{i=1}^{S_N-1} \sum_{j=i+1}^{S_N} p_{ij}$$

For both tests, we take a confidence level of 95% (i.e. $\alpha = 0.05$). Considering the worst possible case, where both components of the p_{ij} indicator are erroneously 0, a WSS quotient as low as 0.9 may be expected for purely stationary data (given that the 5% confidence-level applies to both criteria).

In the development of the above metric, specific forms are assumed for the distributions that the samples to be analysed should follow (i.e. that the distribution followed is

[1] We base our statistics on *unbiased* estimates.

Gaussian). However, from the results of section 3, we know that the distribution of music samples is not entirely Gaussian. It should also be possible to develop a non-parametric version of the WSS-Quotient based on Levene's test and the Mann-Whitney test (see [4]).

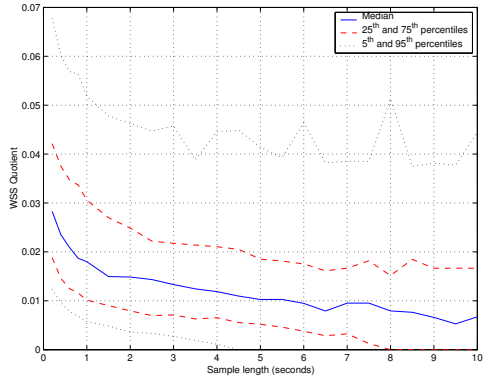

Figure 4 - Median and 5^{th}, 25^{th}, 75^{th} and 95^{th} percentiles of the WSS quotient for samples of varying length taken from each track.

Shown in Figure 4 are the median values of the WSS quotient (and corresponding percentile bounds) for subsamples of different lengths of the music samples. For shorter sample lengths, we see that the computed WSS quotient values are higher; this is encouraging, given that one would expect most demixing algorithms to be applied to short sample lengths (and even to have an on-line version for real-time applications).

In order to further explore typical WSS-Quotient values for short sample lengths, further analysis was performed. All the music tracks sampled above were divided into 10s blocks, and each of these were sub-divided into blocks of $\{0.01, 0.02, 0.04, 0.05, 0.1, 0.2, 0.25, 0.5, 1.0\}$ seconds. For each of the blocks, WSS-Quotient values were then determined for each subblock length; for presentation purposes, the median and 5^{th}, 25^{th}, 75^{th} and 95^{th} percentiles of values at each subblock length were computed. These are shown in Figure 5.

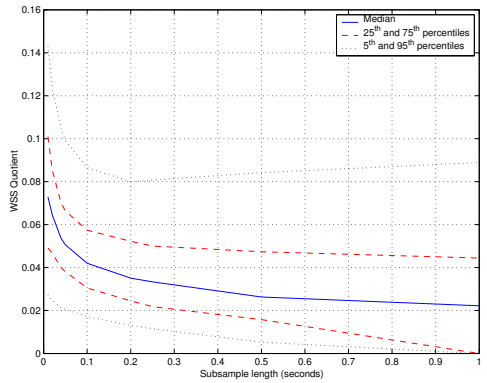

Figure 5 - Median and 5^{th}, 25^{th}, 75^{th} and 95^{th} percentiles of the WSS quotient for sub-samples of varying length taken from 10s blocks.

Once more, we observe that the WSS-Quotient values calculated fall far short of the theoretical ideal of 0.9, thus suggesting that the WSS assumption is unnecessarily restrictive in the theoretical formulation of blind-source separation methodologies. Taking shorter window-lengths (of 5s and 2s - not shown) yields slightly higher median values for the WSS-Quotient (particularly for the shorter subblock lengths), but values remain of the same order. For subblock lengths ranging from 8 ms to 128 ms (across all window lengths), the WSS-Quotient values are also extremely low (mean values between 0.06 and 0.075 approx.); these subblock lengths correspond to common choices for the length of the (e.g. Hamming) window in the Short-Time Fourier Transform. It may be that the low WSS-Quotient can be attributed to the highly self-similar[2] nature of music (both within the blocks under examination and across the sample blocks - see [2]). These results suggest that it may be more appropriate to make some extremely short-term (rather than global) assumption of wide-sense stationarity.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. U.S. Department of Commerce, 1972.

[2] S. Bates, "Traffic characterisation and modelling for call admission control schemes on asynchronous transfer mode networks," Ph.D. dissertation, http://www.see.ed.ac.uk/∼sasg/index.html, Univ. of Edinburgh, Scotland, 1997.

[3] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus (LDC93S1)," http://www.ldc.upenn.edu/Catalog/, Linguistic Data Consortium, Philadelphia, 1990.

[4] M. Hollander and D. A. Wolfe, *Nonparametric statistical methods*. Wiley, 1973.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley-Interscience, 2001.

[6] S. Larbi and M. Jaidane, "Watermarking influence on the stationarity of audio signals," vol. 6, no. 2, Apr. 2003.

[7] S. Larbi and M. Jaidane-Saidane, "Audio watermarking: a way to stationnarize audio signals," *IEEE Trans. Signal Processing*, vol. 53, no. 2, pp. 816–823, Feb. 2005.

[8] B. Mandelbrot and R. L. Hudson, *The (Mis)Behavior of Markets, A Fractal View of Risk, Ruin and Reward*. Perseus, 2006.

[9] J. A. McEwen and G. B. Anderson, "Modeling the stationarity and gaussianity of spontaneous electroencephalographic activity," *IEEE Trans. Biomed. Eng.*, vol. 22, no. 5, pp. 361–369, 1975.

[10] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C and the Art of Scientific Computing*. Cambridge University Press, 1992.

A. CD SOURCES

The musical samples used for the preceeding analysis were drawn from the following CDs:

[2]The extent of self-similarity may be investigated through determination of the **Hurst exponent** - see [8].

- Live in Paris : Diana Krall (13 tracks)
- Rolling Back The Years Vols. 1 & 2 (40 tracks)
- Musiques pour Jeanne la Folle - 1479-1555 Espagne: La Nef (26 tracks)
- Requiem : W. A. Mozart (18 tracks)
- The Tempest *or* The Enchanted Island : Henry Purcell (20 tracks)
- Delius - Royal Philharmonic Orchestra : Frederick Delius (9 tracks)
- Symphony in G minor & Sinfonietta : Ernest John Moeran (13 tracks)
- Long Journey Home (16 tracks)
- The Rough Guide to the music of Canada (21 tracks)
- Oxfam Salsa (14 tracks)
- The Dance Music of Ireland: Jigs & Reels (14 tracks)
- Traditional Music of Ireland (14 tracks)