Chapter 5 Data analysis

5.1 Introduction

Analysis of human motion can be interpreted on a number of levels. Low level analysis includes parameters such as limb position and orientation. High level analysis includes posture, gesture and expression analysis. Human motion has been studied for many years on both high and low level. For the purpose of this study it is convenient to represent the data as a discrete-time stochastic process, and the human motion is viewed as low level digital waveform representations. Mathematically tractable and precise engineering approaches can be used to analyze and characterize the waveforms. Specific attention will be given to the analysis of joint angles, since this is the primary source of information that will be compressed and coded.

There are fundamentally two approaches in determining the statistics of human motion. The first is to look at the *driving force* or *process* behind the motion and to analyze the motion from a purely analytical perspective. The other approach is to look at an *infinite* amount of stored motion data and to interpret it purely *numerically*. Both of these methods are fraught with difficulties. A useable mathematical model might not always exist for every human motion variable to be analyzed. Even with appropriate models there are still too many unknowns, which tend to undermine an analytical approach. On the other hand, it is impossible to store and process an infinite amount of data, which raises the question whether the sample used is representative of the population. This research approaches the analysis of the motion numerically, provided it is understood that the results are only applicable to the few *types* of motion discussed here.

5.2 Numerical analysis

It is convenient to assume that the human motion waveforms can be represented by an ergodic random process. Although this is a gross simplification, such a statistical point of view can yield useful results. The random process in question here is applicable only to *specific* types of motion (i.e. to the examples presented here) – we do not attempt to derive statistics for human motion in general. Such a task would be almost impossible. It is occasionally convenient to group a number of DOFs together to avoid tedious repetition and to clarify results. We assume that there are a number of such groups that are independent of each other, and that each has different characteristics. Clearly, foot movement does not depend on hand movement for normal human behaviour. Where appropriate, the characteristics of head movement, torso movement, arm movement, leg movement and finger movement will be jointly investigated. Table 5-1 shows this in more detail.

Table 5-1: Joint and segment grouping

Group	Reference number	Number of joints	Number of segments	Number of DOFs
Root and torso	0	2	2	7
Neck and head	1	2	2	5
Left arm	2	3	3	8
Left hand	3	14	14	19
Right arm	4	3	3	8
Right hand	5	14	14	19
Left leg	6	3	3	7
Right leg	7	3	3	7

Denote the jth DOF of the ith joint as a sequence $\{\theta_{i,j}(n)\}$ of a discrete-time random process. From table 2-1 and figure 2-4 it can be seen that $0 \leq i \leq 48$ and $1 \leq j \leq 6$. The same notation can be used for a group of DOFs, with the subscript i indicating the group number, and in this case we have from table 5-1 that $1 \leq i \leq 8$ and $1 \leq j \leq 19$. Refer to Appendix I for additional information. For the purposes of this research it is adequate to characterize the motion signal and its derivative in terms of its first order probability density, and in terms of its autocorrelation and power spectral density functions. These methods will be discussed in the following sections.

5.2.1 Examples

Three examples of human motion will be used for statistical analysis. The first example is obtained from general conversational movements, the second is obtained from fast dance movements and the third from hand gestures. The latter is used specifically for analysis of finger movement – general body activity often lacks detailed hand gestures. We assume that most common motion will fall between the extremes represented by these examples. Figure 5-1a shows a 10 second segment of conversational movement for the left arm, figure 5-1b a 10 second segment of dance movement for the left arm, and figure 5-1c a 10 second segment of finger movement. Figure 5-1d depicts an image of 1 second's worth of overlaid 3D rendered frames for the dance sequence (skeleton only). The length of the original sequences is 300 seconds each. For clarity these figures show but a fraction of the available DOFs and sequence lengths. The full motion sequences are available on request. The motion sequences were captured using the techniques described in chapter 3, at a sampling rate of 30 Hz. We therefore assume that the frequency content of the motion is less than 15 Hz to satisfy the Nyquist criterion. It will later be shown that this is indeed the case, except possibly for extremely fast motion such as found in sport activities.

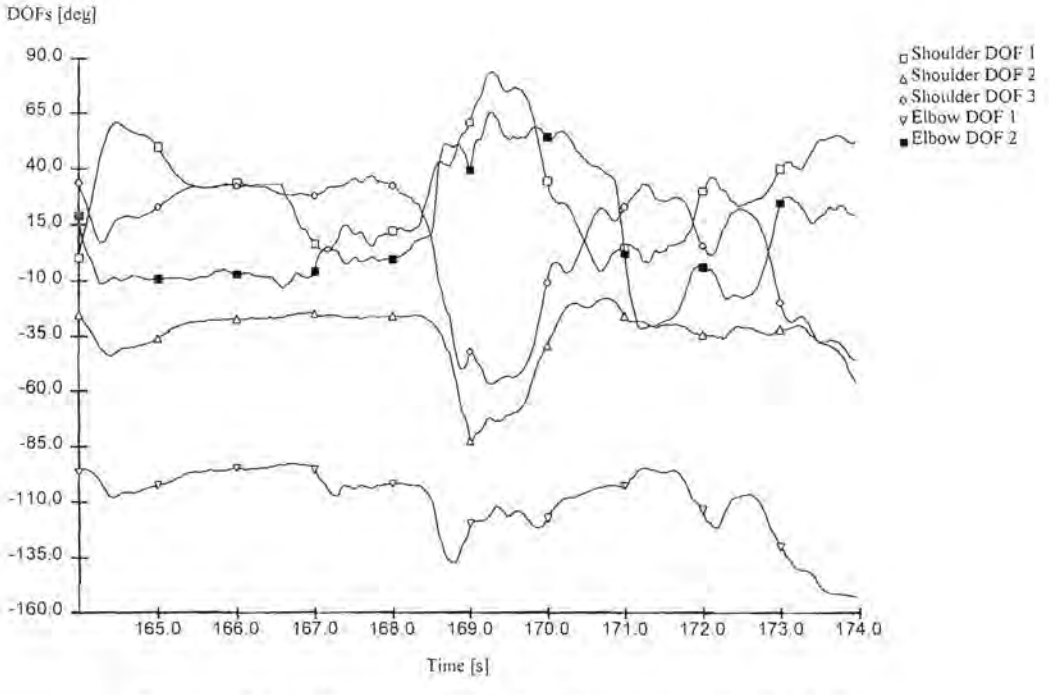

Figure 5-1a: Conversational motion example for left arm.

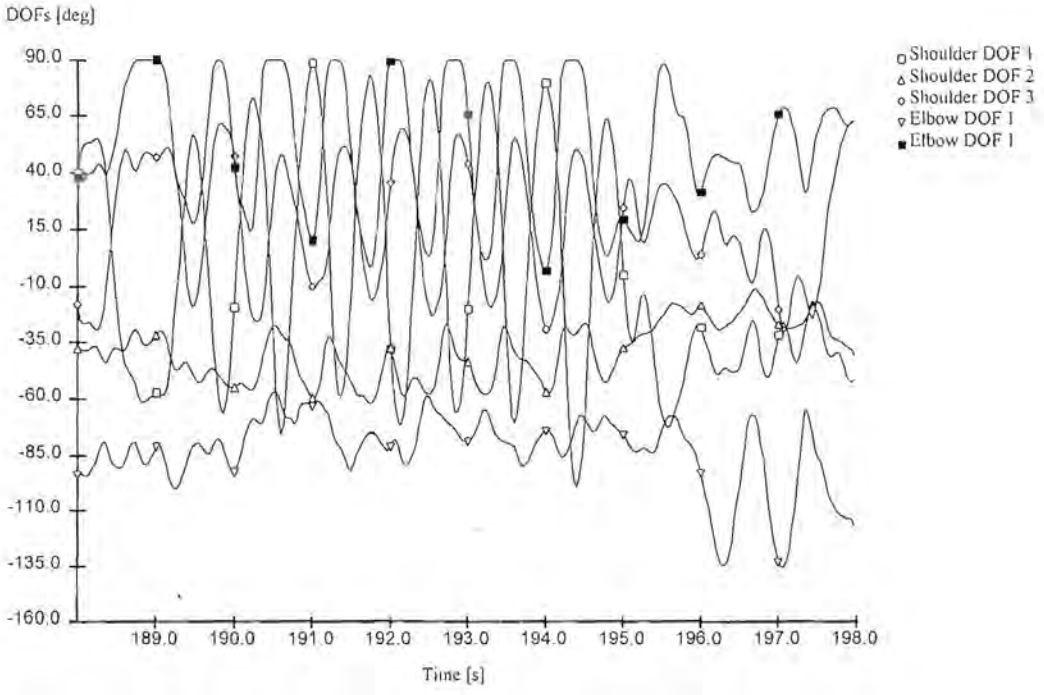

Figure 5-1b: Dance motion example for left arm.

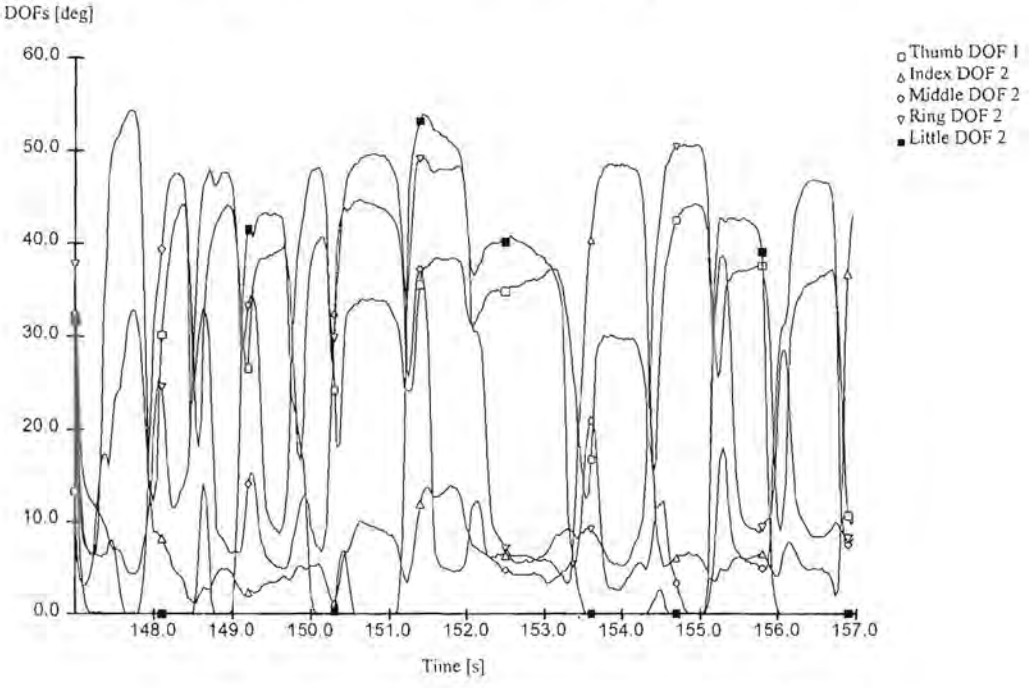

Figure 5-1c: Finger gesture example for left hand.

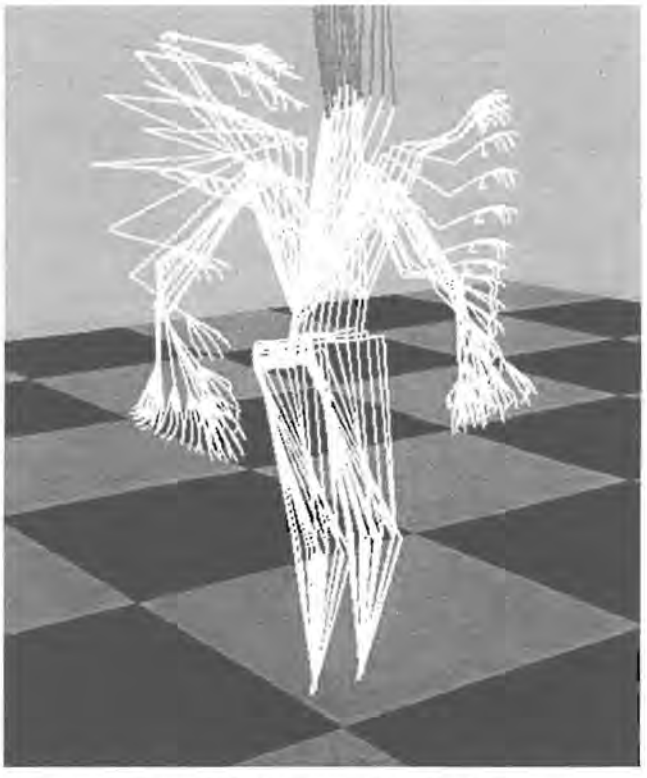

Figure 5-1d: Overlaid frames for the dance sequence.

5.2.2 Spatial content

Ranges

The range of each degree of freedom is mainly a function of the theoretical joint limits as discussed in chapter 2. However, table 5-2 summarizes the practical ranges as obtained from the motion sequences of figure 5-1, excluding the hands. Table 5-3 summarizes the characteristics of the finger movement for the gesture sequence. Also shown in the tables are the mean and standard deviation for each DOF.

Table 5-2: Summary of body DOF characteristics.

Description		Conversational sequence					Dance sequence				
Joint name	DOF	Min	Max	Range	Mean	Std. dev	Min	Max	Range	Mean	Std. dev.
Root	$\theta_{0,1}$	-71.4	59.7	131.2	-0.1	18.16	-63.6	51.5	115.1	-1.8	14.75
Root	$\theta_{0,2}$	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00
Root	$\theta_{0,3}$	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00
Root	$\theta_{0,4}$	-0.100	0.165	0.265	0.016	0.046	-0.163	0.247	0.410	0.034	0.088
Root	$\theta_{0,5}$	0.860	1.068	0.208	0.997	0.010	0.804	1.037	0.233	0.965	0.026

Root	$\theta_{0,6}$	-0.118	0.085	0.203	-0.011	0.028	-0.280	0.060	0.340	-0.060	0.049
Torso	$\theta_{1,1}$	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00
Torso	$\theta_{1,2}$	-16.3	31.9	48.2	6.5	6.05	-8.2	30.0	38.2	10.7	5.26
Torso	$\theta_{1,3}$	-29.6	20.4	50.0	-1.7	5.35	-26.8	28.9	55.7	1.1	6.61
Neck	$\theta_{2,1}$	-24.1	22.2	46.4	0.5	7.99	-29.1	22.8	51.9	-6.1	7.63
Neck	$\theta_{2,2}$	-15.0	15.0	30.0	0.7	4.97	-15.6	15.4	30.9	2.4	6.58
Head	$\theta_{3,1}$	-76.3	82.9	159.3	9.5	25.92	-88.5	89.2	177.7	-9.9	30.56
Head	$\theta_{3,2}$	-24.1	22.2	46.4	0.5	7.99	-31.4	22.8	54.2	-6.1	7.65
Head	$\theta_{3,3}$	-15.0	15.0	30.0	0.7	4.97	-15.6	15.4	30.9	2.4	6.58
Left shoulder	$\theta_{5,1}$	-82.8	131.8	214.5	29.5	35.66	-100.0	100.9	200.9	-7.5	30.10
Left shoulder	$\theta_{5,2}$	-110.3	-6.5	103.8	-27.0	11.65	-91.1	-3.9	87.2	-32.4	11.77
Left shoulder	$\theta_{5,3}$	-167.9	169.7	337.6	29.0	55.34	-95.4	85.0	180.4	8.8	30.63
Left elbow	$\theta_{6,1}$	-153.0	-4.0	149.0	-66.7	45.69	-148.8	-4.3	144.6	-86.0	22.42
Left elbow	$\theta_{6,2}$	-91.0	110.7	201.7	0.5	45.68	-38.7	90.4	129.0	45.8	22.83
Left wrist	$\theta_{7,1}$	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00
Left wrist	$\theta_{7,2}$	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00
Right shoulder	$\theta_{22,1}$	-113.6	104.8	218.4	-8.1	38.75	-103.5	104.5	208.0	1.8	29.79
Right shoulder	$\theta_{22,2}$	6.0	137.0	131.0	23.6	18.83	3.3	129.4	126.0	32.9	15.28
Right shoulder	$\theta_{22,3}$	-112.2	132.9	245.1	-7.5	39.55	-90.3	167.7	258.0	-4.7	28.92
Right elbow	$\theta_{23,1}$	-156.2	-4.0	152.1	-77.1	46.30	-161.4	-4.3	157.1	-74.4	29.41
Right elbow	$\theta_{23,2}$	-90.3	90.0	180.3	-15.3	31.93	-90.4	89.8	180.2	-24.4	33.60
Right wrist	$\theta_{24,1}$	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00
Right wrist	$\theta_{24,2}$	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00
Left hip	$\theta_{39,1}$	-32.6	7.2	39.8	-5.4	3.57	-42.6	2.7	45.4	-17.4	7.90
Left hip	$\theta_{39,2}$	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00
Left hip	$\theta_{39,3}$	-10.7	7.6	18.3	-0.9	3.18	-18.8	12.0	30.8	-2.2	5.95
Left knee	$\theta_{40,1}$	0.1	68.0	67.9	8.4	5.69	-0.8	80.4	81.2	25.1	13.92
Left ankle	$\theta_{41,1}$	-30.0	14.5	44.6	-2.9	4.15	-30.0	19.9	49.9	-7.7	8.27
Left ankle	$\theta_{41,2}$	-15.0	15.0	30.0	0.5	10.40	-15.1	15.8	30.9	1.3	9.80
Right hip	$\theta_{43,1}$	-32.4	4.8	37.2	-5.5	3.41	-41.9	4.0	45.9	-17.2	7.82
Right hip	$\theta_{43,2}$	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00
Right hip	$\theta_{43,3}$	-11.1	6.6	17.7	-1.3	3.22	-15.0	12.1	27.1	-2.5	5.88
Right knee	$\theta_{44,1}$	0.1	68.0	67.9	8.5	5.71	-0.8	80.5	81.3	25.4	13.91
Right ankle	$\theta_{45,1}$	-30.0	10.2	40.2	-3.0	3.84	-30.0	21.5	51.6	-8.1	8.06
Right ankle	$\theta_{45,2}$	-15.0	15.0	30.0	0.5	10.40	-15.1	15.8	30.9	1.3	9.80

Table 5-3: Summary of finger DOF characteristics.

Joint name	DOF	Min	Max	Range	Mean	Std. dev.
Left thumb 1	$\theta_{8,1}$	-0.0	30.0	30.1	11.7	4.73
Left thumb 2	$\theta_{9,1}$	-0.1	80.1	80.2	31.1	12.62
Left index 1	$\theta_{10,1}$	0.0	0.0	0.0	0.0	0.00
Left index 1	$\theta_{10,2}$	-0.1	74.9	75.0	27.4	16.74
Left index 2	$\theta_{11,1}$	-0.1	74.9	75.0	27.4	16.74
Left index 3	$\theta_{12,1}$	-0.1	60.0	60.0	21.9	13.39
Left middle 1	$\theta_{13,1}$	0.0	0.0	0.0	0.0	0.00
Left middle 1	$\theta_{13,2}$	-0.0	75.0	75.1	27.6	17.29
Left middle 2	$\theta_{14,1}$	-0.0	75.0	75.1	27.6	17.29
Left middle 3	$\theta_{15,1}$	-0.0	60.0	60.1	22.1	13.83
Left ring 1	$\theta_{16,1}$	0.0	0.0	0.0	0.0	0.00
Left ring 1	$\theta_{16,2}$	-0.1	75.1	75.2	24.5	16.08
Left ring 2	$\theta_{17,1}$	-0.1	75.1	75.2	24.5	16.08
Left ring 3	$\theta_{18,1}$	-0.0	60.1	60.1	19.6	12.86
Left little 1	$\theta_{19,1}$	0.0	0.0	0.0	0.0	0.00
Left little 1	$\theta_{19,2}$	-0.1	75.1	75.2	20.6	15.87
Left little 2	$\theta_{20,1}$	-0.1	75.1	75.2	20.6	15.87
Left little 3	$\theta_{21,1}$	-0.1	60.1	60.1	16.5	12.70
Right thumb 1	$\theta_{25,1}$	-0.0	30.1	30.1	11.7	6.44
Right thumb 2	$\theta_{26,1}$	-0.0	80.2	80.2	31.3	17.18
Right index 1	$\theta_{27,1}$	0.0	0.0	0.0	0.0	0.00
Right index 1	$\theta_{27,2}$	-0.0	75.1	75.1	21.7	20.26
Right index 2	$\theta_{28,1}$	-0.0	75.1	75.1	21.7	20.26
Right index 3	$\theta_{29,1}$	-0.0	60.1	60.1	17.4	16.21
Right middle 1	$\theta_{30,1}$	0.0	0.0	0.0	0.0	0.00
Right middle 1	$\theta_{30,2}$	-0.1	75.0	75.1	26.3	21.49
Right middle 2	$\theta_{31,1}$	-0.1	75.0	75.1	26.3	21.49
Right middle 3	$\theta_{32,1}$	-0.1	60.0	60.1	21.0	17.19
Right ring 1	$\theta_{33,1}$	0.0	0.0	0.0	0.0	0.00
Right ring 1	$\theta_{33,2}$	-0.1	75.1	75.1	29.6	20.58
Right ring 2	$\theta_{34,1}$	-0.1	75.1	75.1	29.6	20.58
Right ring 3	$\theta_{35,1}$	-0.1	60.0	60.1	23.7	16.46
Right little 1	$\theta_{36,1}$	0.0	0.0	0.0	0.0	0.00
Right little 1	$\theta_{36,2}$	-0.1	75.2	75.3	26.9	22.49
Right little 2	$\theta_{37,1}$	-0.1	75.2	75.3	26.9	22.49
Right little 3	$\theta_{38,1}$	-0.1	60.2	60.2	21.5	17.99

Resolution

The resolution of each DOF is a function of the resolution of the input device used to digitize the motion. As described in chapter 3, we use 6 DOF electromagnetic sensors for body tracking and fiber optical data gloves for finger flexure sensing. The electromagnetic sensors output a 16-bit value for a ±3 m and a ±180° range respectively [74], while the glove device outputs an 8-bit value for full flexure [75]. The output of both of these

devices is converted to a floating-point representation for internal use. The useable hardware resolution of the electromagnetic sensors is a function of the distance from the transmitter, as was indicated in figure 4-1. The useable glove sensor hardware resolution is constant under all circumstances. The total useable resolution in bits is a function of both the hardware and range resolution, denoted by N_h and N_r respectively. Using the noise power σ_p^2 from figure 4-1, we find that the maximum number of useable bits for the 3 position DOFs is given by

$$\begin{aligned} N_p &= 16-(N_h+N_r) \\ &= 16-\left(\log_2\left(\frac{65536.\sigma_p}{6}\right)+\log_2\left(\frac{6}{R_p}\right)\right), \end{aligned} \tag{5-1}$$

where R_p is the range as found in table 5-2. The bits for the 3 angular DOFs is given by

$$\begin{aligned} N_a &= 16-(N_h+N_r) \\ &= 16-\left(\log_2\left(\frac{65536.\sigma_r}{360}\right)+\log_2\left(\frac{360}{R_a}\right)\right), \end{aligned} \tag{5-2}$$

where R_a is the range as found in table 5-2, and σ_a^2 is the noise power from figure 4-1. Similarly, if we assume a full finger flexure of roughly 70°, from chapter 4 we find that the useable number of flexion bits is given by

$$N_f = 8-\log_2\left(\frac{256.\sigma_f}{70}\right). \tag{5-3}$$

Assuming that the non-linear inverse kinematics calculations do not adversely influence resolution and range, the useable number of bits for each DOF can be found using the above mentioned equations. Table 5-4 shows the bit quantities with respect to the test sequences discussed above. A maximum, typical and minimum resolution is presented. Note that the number of finger bits is constant, and is only shown once. Throughout the

rest of the text, the *typical* value will be used for comparison purposes. Table 5-4 shows only the DOFs that we can actually measure, as described in chapter 4. In all fairness, only these DOFs should be used to calculate the raw, uncompressed bit-rate requirement. From table 5-4, it can be seen that the reduced skeleton model typically requires 345 bits/frame for the body and 168 bits/frame for the hands. At a sampling rate of 30 Hz, the bit-rate requirement is 15390 bits/second.

Table 5-4: Resolution in bits for each DOF

Joint name	DOF	Range	Maximum	Typical	Minimum
Root	$\theta_{0,1}$	131.2	14	12	9
Root	$\theta_{0,4}$	0.410	14	11	7
Root	$\theta_{0,5}$	0.264	13	11	7
Root	$\theta_{0,6}$	0.364	13	11	7
Torso	$\theta_{1,2}$	48.2	13	10	7
Torso	$\theta_{1,3}$	58.4	13	11	8
Neck	$\theta_{2,1}$	51.9	13	10	8
Neck	$\theta_{2,2}$	30.9	12	10	7
Head	$\theta_{3,1}$	177.7	15	12	9
Head	$\theta_{3,2}$	54.2	13	11	8
Head	$\theta_{3,3}$	30.9	12	10	7
Left shoulder	$\theta_{5,1}$	231.8	15	13	10
Left shoulder	$\theta_{5,2}$	106.4	14	12	9
Left shoulder	$\theta_{5,3}$	337.6	16	13	10
Left elbow	$\theta_{6,1}$	149.0	14	12	9
Left elbow	$\theta_{6,2}$	201.7	15	12	10
Right shoulder	$\theta_{22,1}$	218.4	15	13	10
Right shoulder	$\theta_{22,2}$	133.6	14	12	9
Right shoulder	$\theta_{22,3}$	279.9	15	13	10
Right elbow	$\theta_{23,1}$	157.3	15	12	9
Right elbow	$\theta_{23,2}$	180.4	15	12	9
Left hip	$\theta_{39,1}$	49.8	13	10	8
Left hip	$\theta_{39,3}$	30.8	12	10	7
Left knee	$\theta_{40,1}$	81.2	14	11	8
Left ankle	$\theta_{41,1}$	49.9	13	10	8
Left ankle	$\theta_{41,2}$	30.9	12	10	7
Right hip	$\theta_{43,1}$	46.7	13	10	7
Right hip	$\theta_{43,3}$	27.1	12	10	7
Right knee	$\theta_{44,1}$	81.3	14	11	8
Right ankle	$\theta_{45,1}$	51.6	13	10	8
Right ankle	$\theta_{45,2}$	30.9	12	10	7
Fingers	14 DOFs per hand	70	8	6	6

5.2.3 Temporal content and statistics

Average

The average of $\{\theta_{i,j}(n)\}$ is given by

$$\eta_{i,j} = E[\theta_{i,j}(n)] = \frac{1}{N}\sum_{n=1}^{N}\theta_{i,j}(n), \tag{5-4}$$

for a sequence of length N. For a stochastic process to be ergodic, we must be able to prove that the time averages are equal to the ensemble or probability averages. Although we do not prove it explicitly here, it is reasonable to assume that this is the case and that $\{\theta_{i,j}(n)\}$ is ergodic. The results from the following sections also give strong indications that this assumption is reasonable.

Variance

The variance of $\{\theta_{i,j}(n)\}$ is given by

$$\sigma_{i,j} = E\left[\left(\theta_{i,j}(n)-\eta_{i,j}\right)^2\right] = \frac{1}{N}\sum_{n=1}^{N}\left(\theta_{i,j}(n)-\eta_{i,j}\right)^2, \tag{5-5}$$

for a sequence of length N.

Autocorrelation

The autocorrelation function of $\{\theta_{i,j}(n)\}$ is given by

$$r_{i,j}(k,l) = E[\theta_{i,j}(k)\theta_{i,j}(l)]. \tag{5-6}$$

A stochastic process is wide sense stationary (WSS) if its mean is constant, i.e. $\eta_{i,j}$ is not a function of n, and its autocorrelation function depends only on the lag or time difference $m = k - l$. In this case, the autocorrelation function can be written as

$$r_{i,j}(m) = E[\theta_{i,j}(n)\theta_{i,j}(n+m)] = \frac{1}{N-m}\sum_{n=1}^{N-m}\theta_{i,j}(n)\theta_{i,j}(n+m), \qquad (5\text{-}7)$$

for a sequence of length N. We have evaluated the mean and autocorrelation functions of the example sequences for arbitrary DOFs and various time origins to test the validity of a WSS process.

Figure 5-3 shows the mean of the arbitrarily chosen head yaw angle $\theta_{3,0}$, the left shoulder elevation angle $\theta_{5,1}$, the index finger flexion $\theta_{27,1}$ and the middle finger flexion $\theta_{30,1}$ for the test sequences as a function of sample origin n. Although there is a small variation, the mean can comfortably be approximated by a constant. Figure 5-4a shows a number of overlaid autocorrelation functions of the head pitch angle $\theta_{3,1}$ for the conversational test sequence, evaluated at a number of arbitrary sample origins n. Figure 5-4b shows the same functions for the dance sequence. Similarly, figure 5-4c and 5-4d depict the autocorrelation functions for the left shoulder twist angle $\theta_{5,2}$. Figure 5-4e shows the overlaid autocorrelation functions for the right index finger flexion angles $\theta_{27,1}$. It is clear from the results that these functions depend little on the sample origin n, and are mainly a function of the lag m. The autocorrelation functions and mean of all the other DOFs exhibit similar behaviour, and it is therefore reasonable to assume that the stochastic process $\{\theta_{i,j}(n)\}$ is wide sense stationary. It should be noted that this is true *only* if all of $\{\theta_{i,j}(n)\}$ is within the *same type of motion*, and the concept of WSS cannot be extended to include human motion in general.

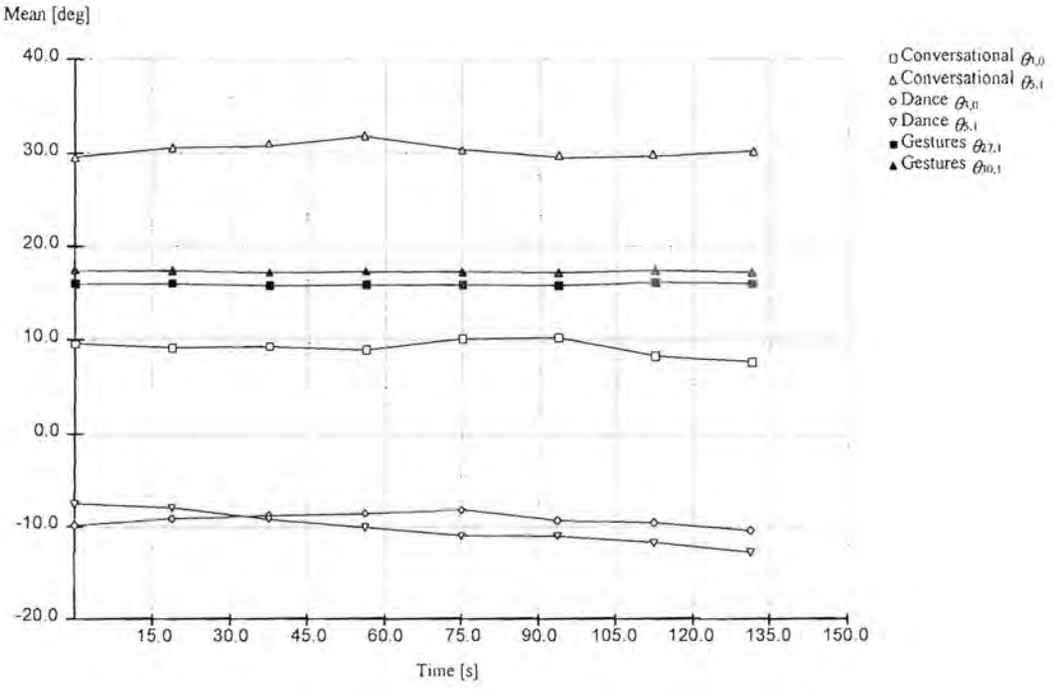

Figure 5-3: Mean as a function of time.

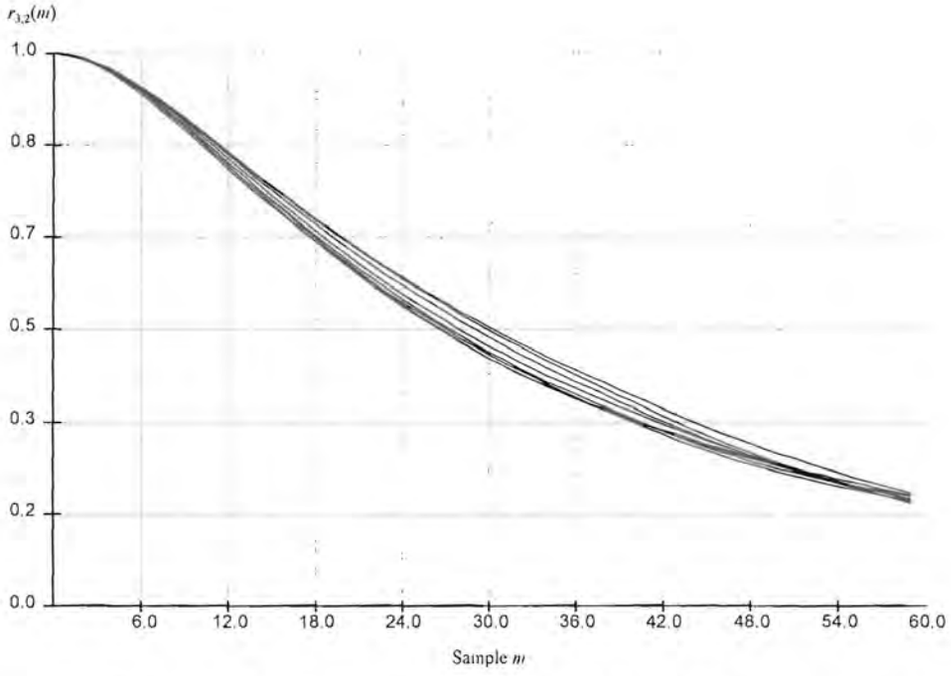

Figure 5-4a: Autocorrelation function of head pitch $\theta_{3,1}$ for conversational motion.

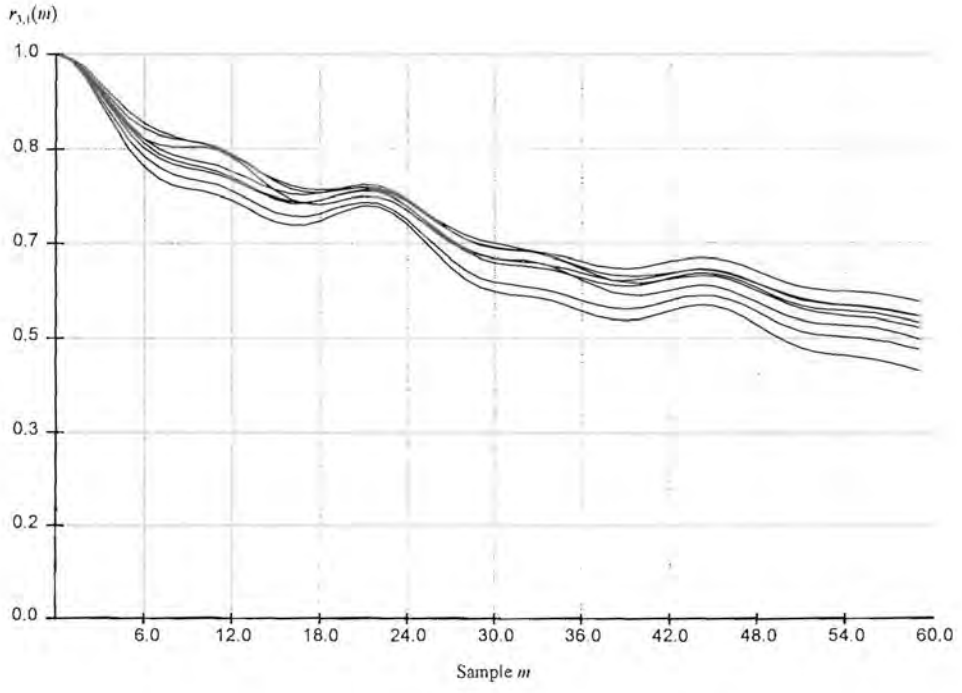

Figure 5-4b: Autocorrelation function of head pitch $\theta_{3,1}$ for dance motion.

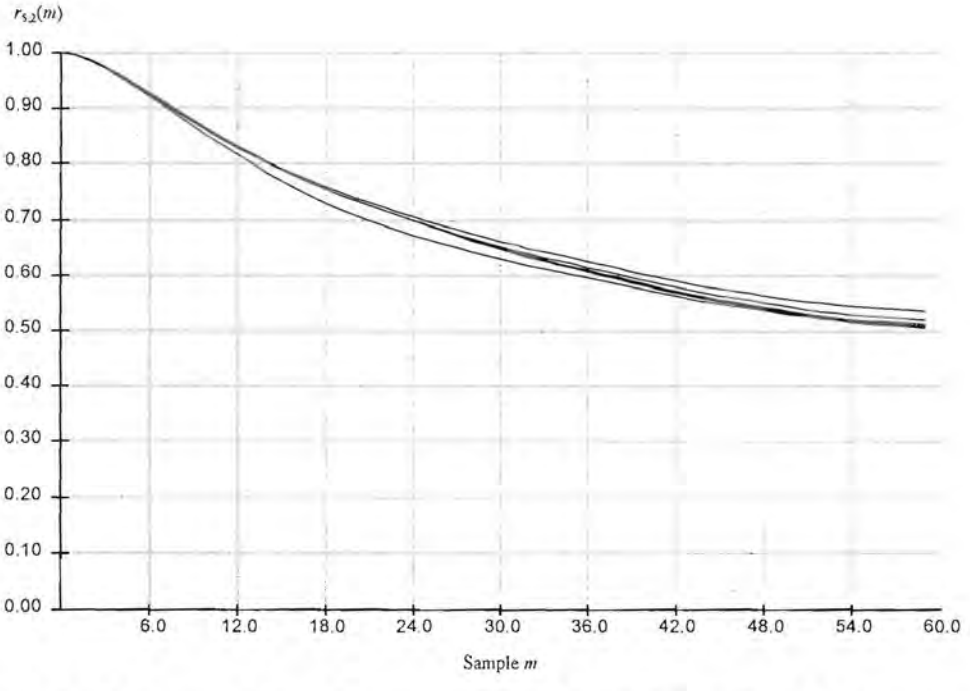

Figure 5-4c: Autocorrelation function of shoulder twist angle $\theta_{5,2}$ for conversational motion.

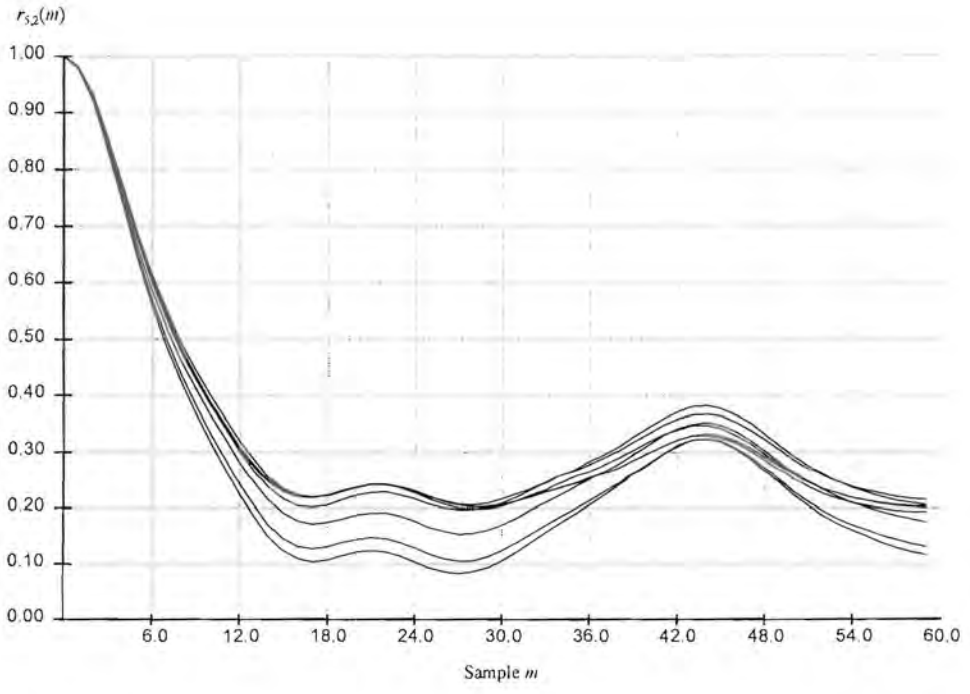

Figure 5-4d: Autocorrelation function of shoulder twist angle $\theta_{5,2}$ for conversational motion.

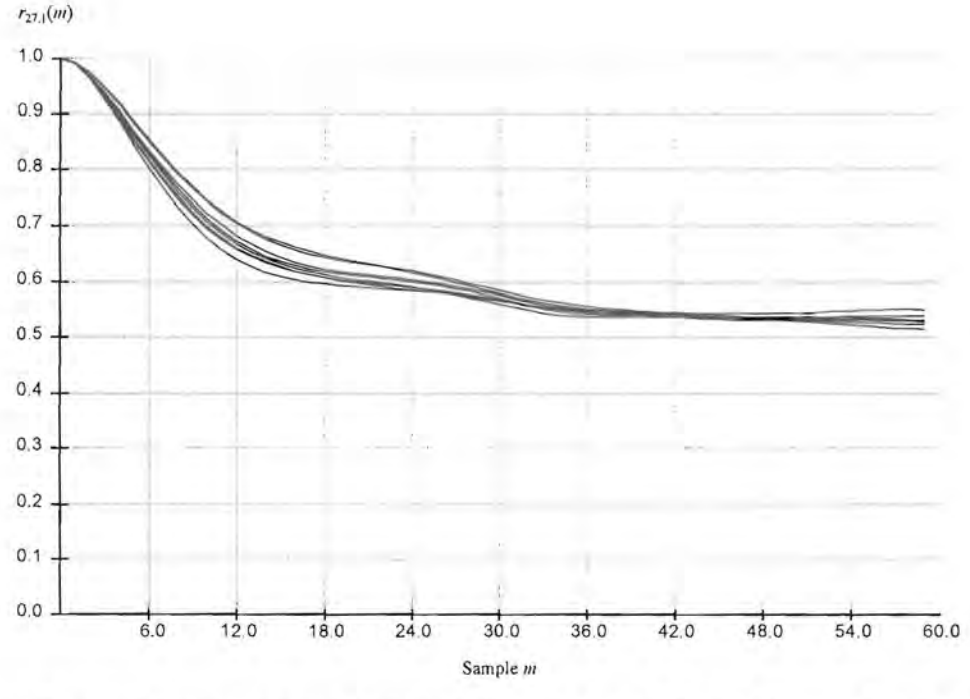

Figure 5-4e: Autocorrelation function of right index finger flexion angle $\theta_{27,1}$ for gesture motion.

It is often convenient to model a time signal as a Markov process, for which the correlation between samples is proportional to their time difference. The autocorrelation function of a discrete, second order, zero-mean Markov process can be written as

$$\phi(m) = \lambda e^{-\alpha^2 m^2}, \tag{5-8}$$

where λ and α are scaling constants. Figure 5-5a depicts the autocorrelation function for the left shoulder elevation angle $\theta_{5,1}$ for the conversational test sequence, with $\phi(m)$ where $\lambda = 1$ and $\alpha = 2.5\text{e-}4$. Figure 5-5b shows the same function for the dance sequence with $\alpha = 0.015$. It can be seen that there is a close match for $m < 10$, and the assumption that the motion can be modeled as a Markov process is reasonable. The other DOFs exhibit similar behaviour.

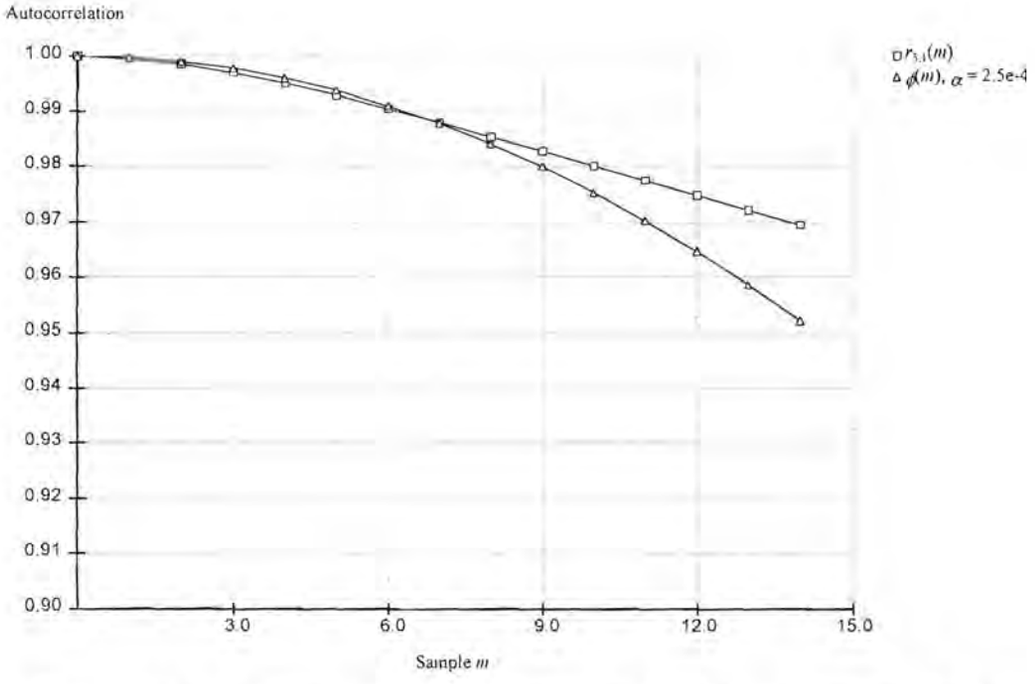

Figure 5-5a: Autocorrelation comparison with a Markov process for the conversational sequence.

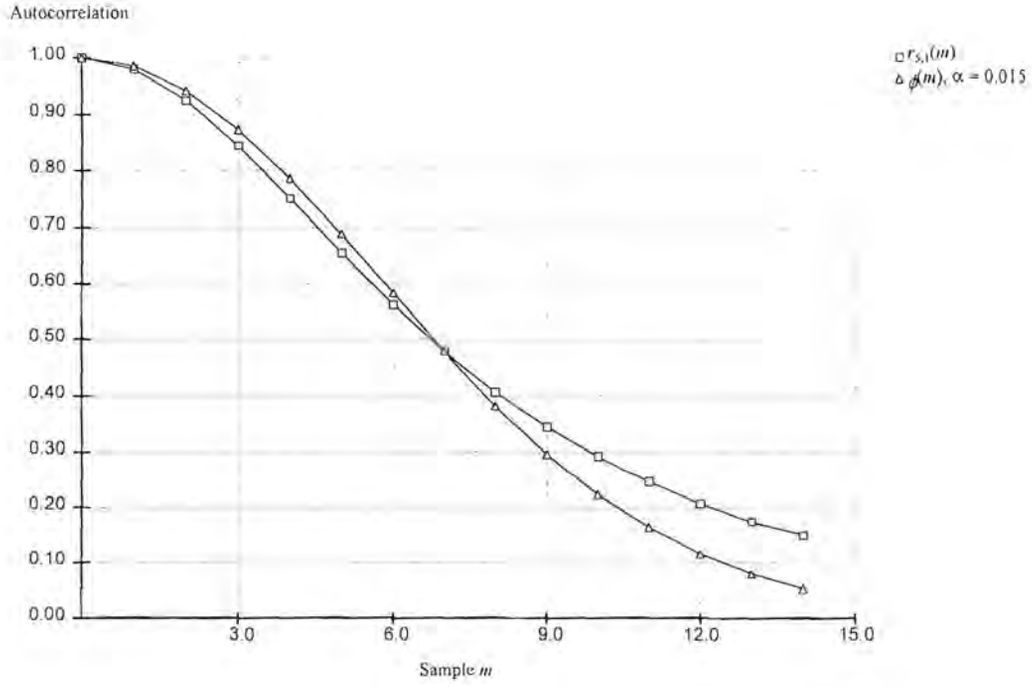

Figure 5-5b: Autocorrelation comparison with a Markov process for the dance sequence.

Probability density function (PDF)

From tables 5-1 and 5-2 it can be seen that most of the DOFs have vastly different ranges and characteristics, and that it would be necessary to obtain separate PDFs for each DOF. For clarity we only show examples of four such PDFs in figure 5-6a to 5-6d, which represent the arbitrary DOFs for the head, arm, leg and finger sections. It is clear that the graphs peak at the orientation favoured by the specific motion sequence, and are also an indication of the mean value. The shoulder angle shows two distinct peaks for the conversational motion. This is an indication that the person who performed the actions favoured two separate postures. The finger PDFs indicates either an open or closed gesture. Some of the body PDFs resembles a Gaussian-like distribution, except for those with strong peaks, in which case the sum of a number of distributions would be more appropriate. The finger PDFs resembles a one sided Rayleigh-like distribution for the open handed gesture and a Gaussian-like distribution for the other gestures.

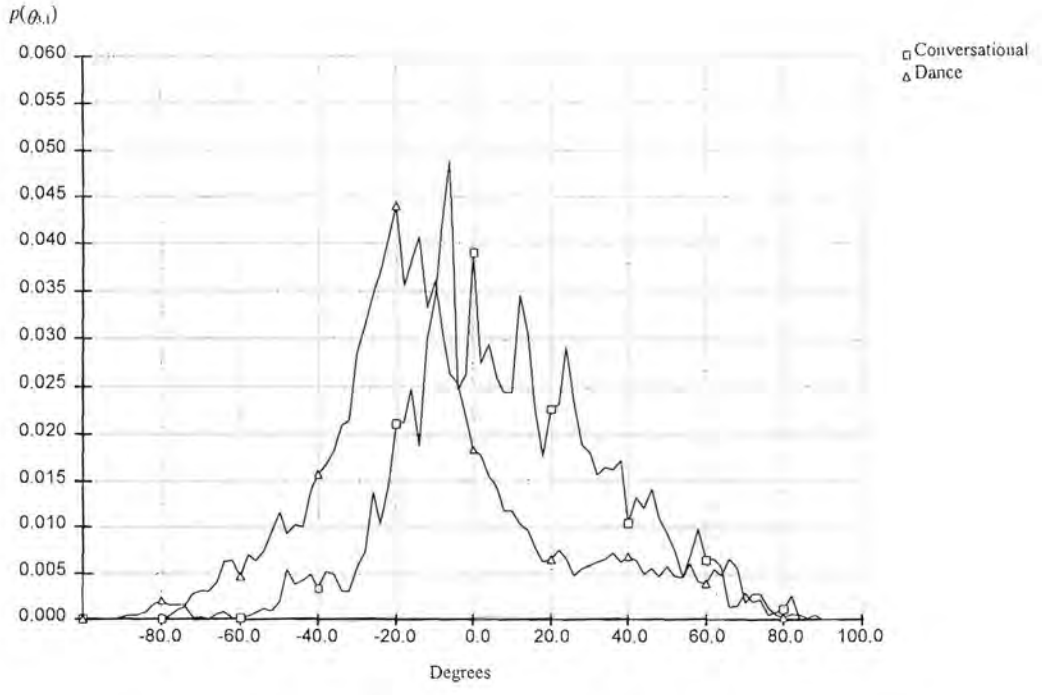

Figure 5-6a: PDF of the head angle $\theta_{3,1}$ for the conversational and dance motion.

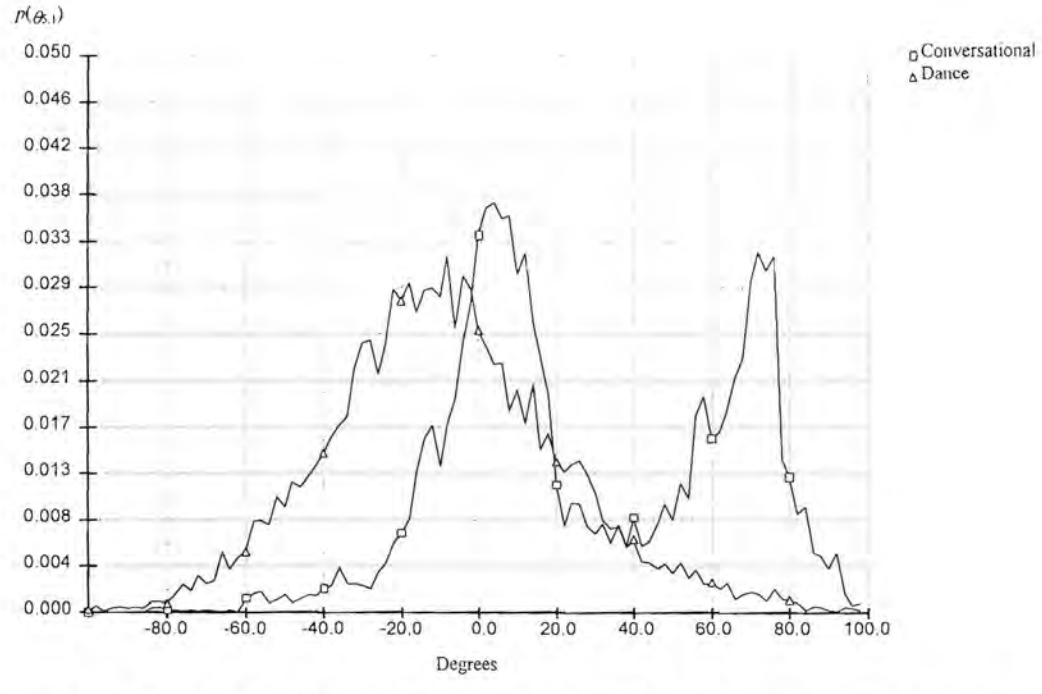

Figure 5-6b: PDF of the shoulder angle $\theta_{5,1}$ for the conversational and dance motion.

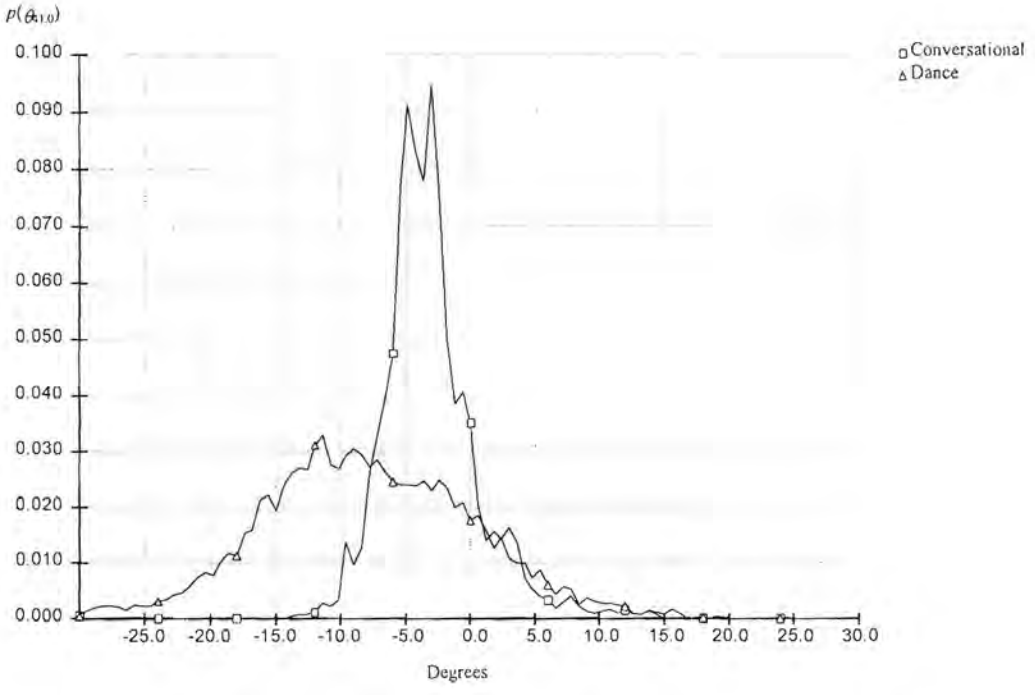

Figure 5-6c: PDF of the ankle angle $\theta_{41,0}$ for the conversational and dance motion.

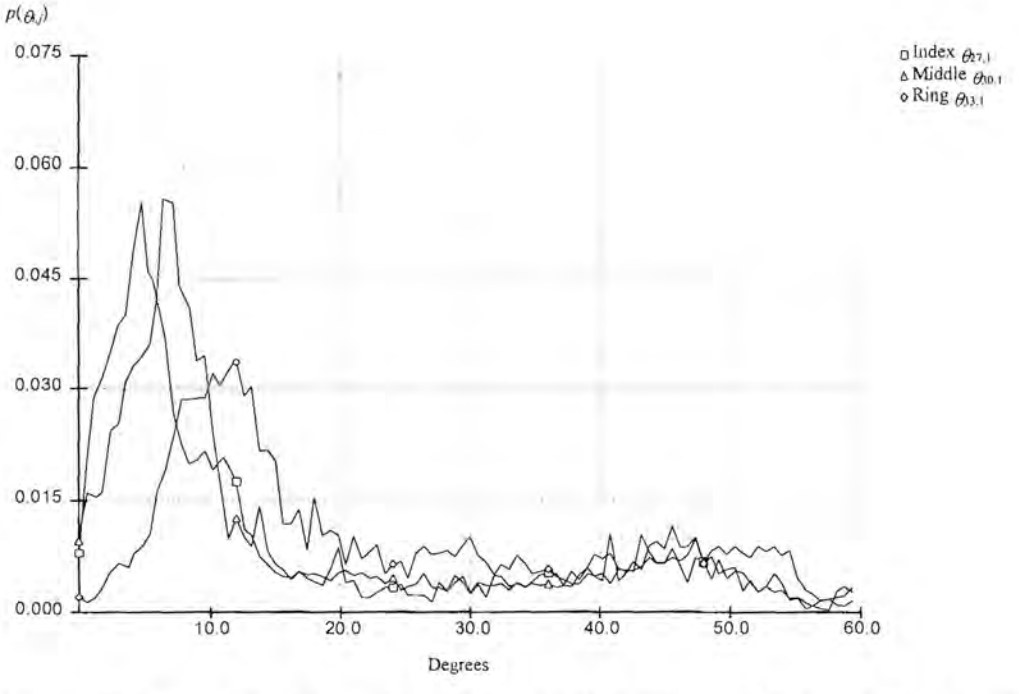

Figure 5-6d: PDF of the index, middle and ring finger flexion for the gesture sequence.

Derivative PDF

Linear predictive coders generally calculate a difference signal that is subsequently coded and transmitted. It is convenient to obtain a PDF for the first order difference $d_{i,j}(n) = \theta_{i,j}(n) - \theta_{i,j}(n\text{-}1)$ as an indication of the range and statistics of this difference signal. Under the assumption that there are a number of DOFs that exhibit similar difference behaviour, such as the arm DOFs or leg DOFs, it is possible to obtain a combined PDF by grouping these together. Figure 5-7 a-d depict the first difference PDFs for the head, arm, leg and finger groups. It can be seen that the first difference is a zero mean sequence, with considerably less variance than the DOF PDFs. Most of the difference PDFs resembles a Laplace-like distribution.

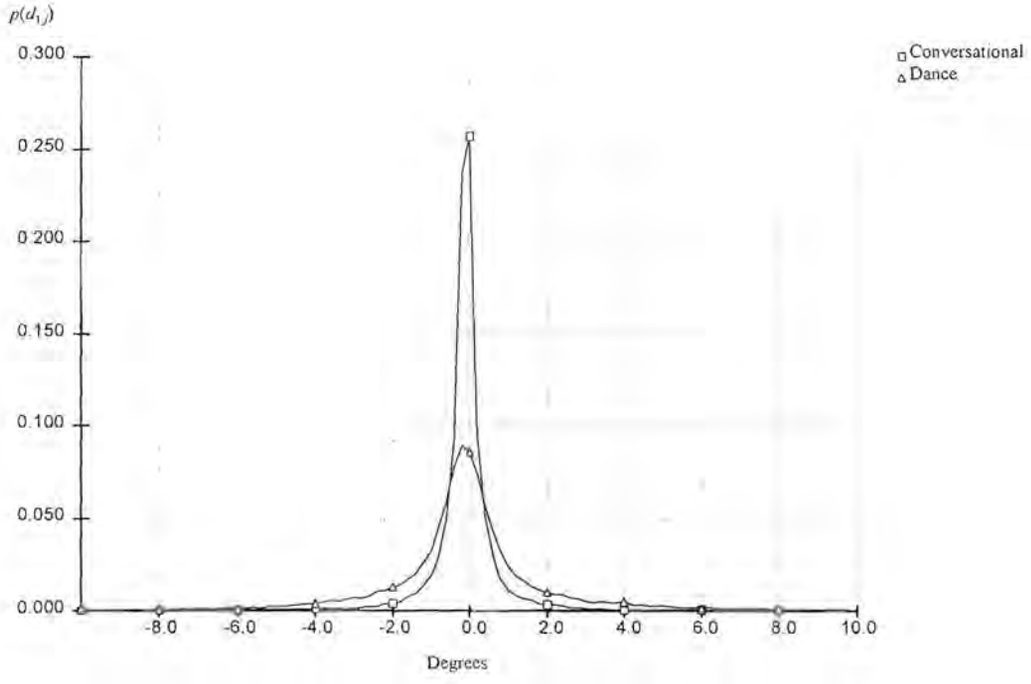

Figure 5-7a: PDF of the head joint group difference angle.

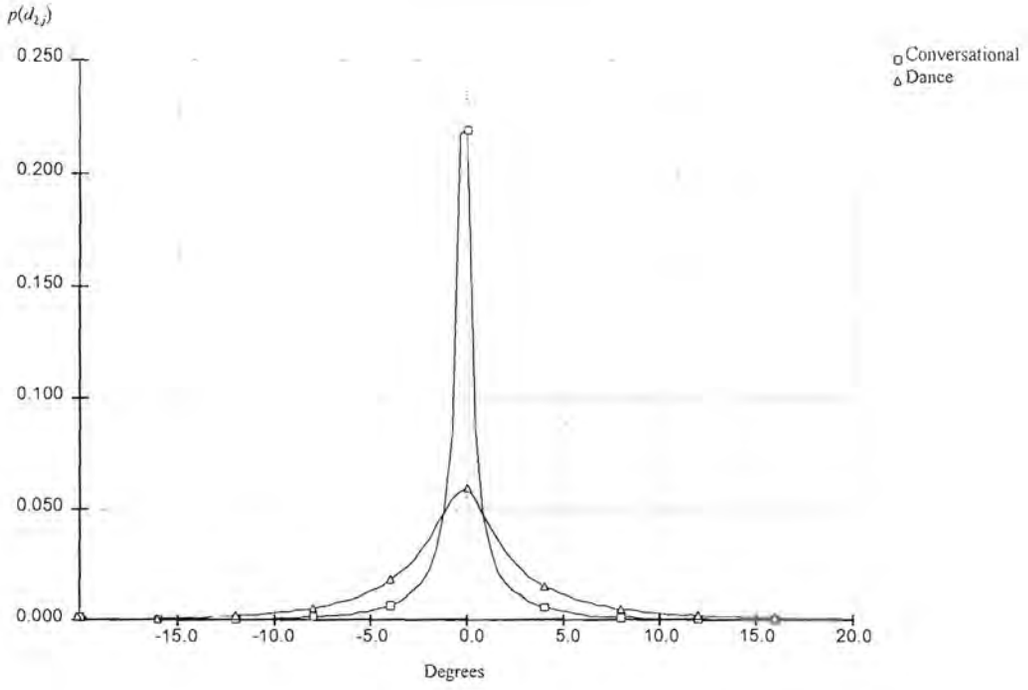

Figure 5-7b: PDF of the left arm joint group difference angle.

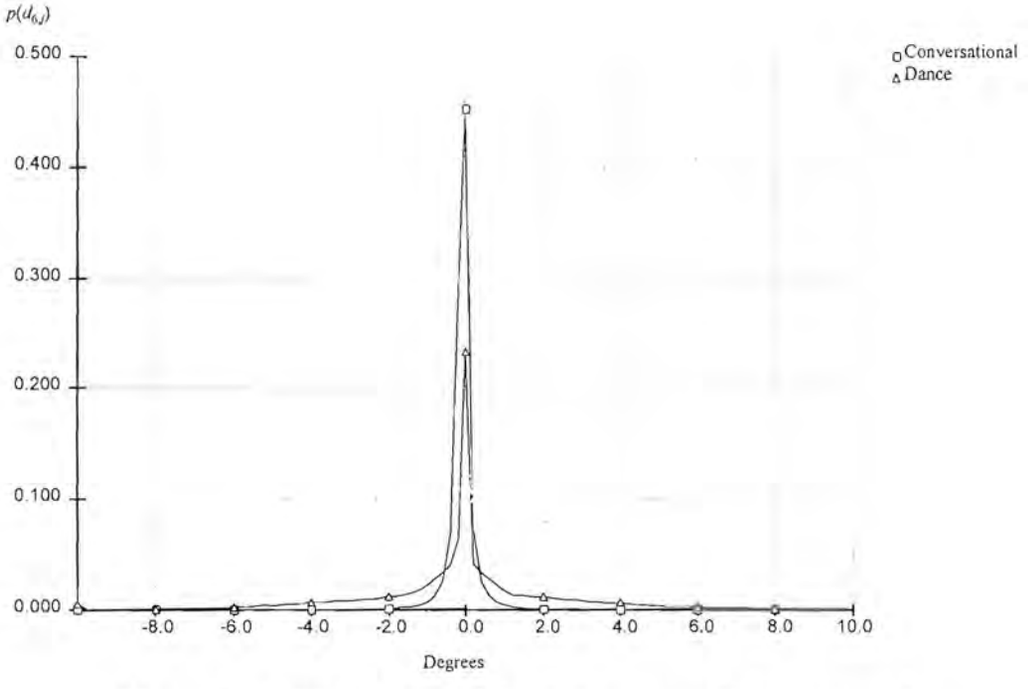

Figure 5-7c: PDF of the left leg joint group difference angle.

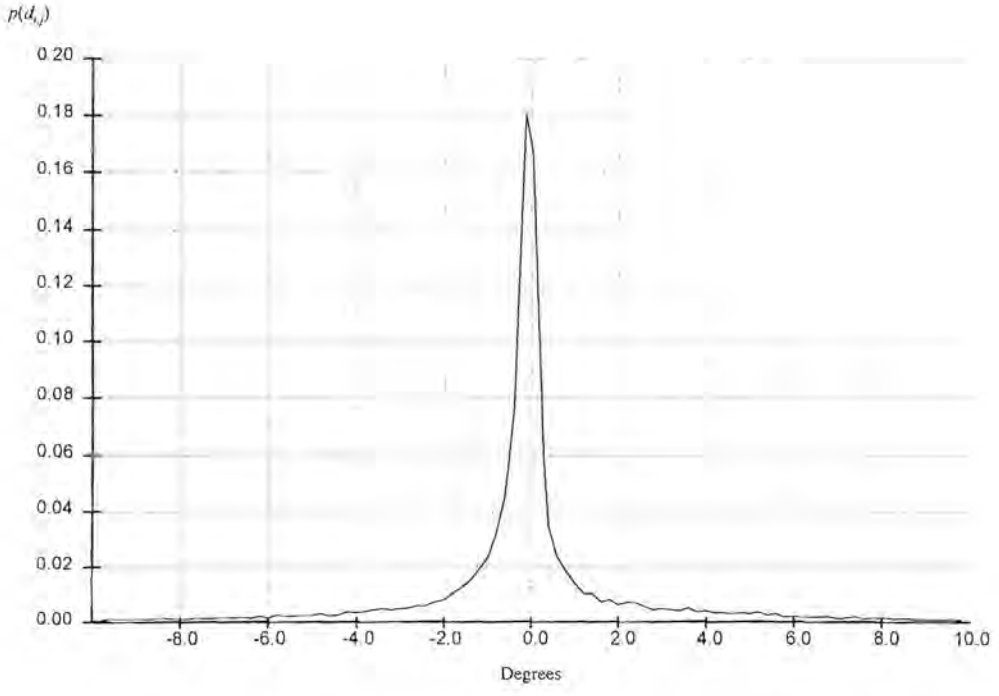

Figure 5-7d: PDF of the right hand finger joint group difference angle for the gesture sequence.

Joint cross dependence

There are a number of coding techniques that rely on the correlation or dependence between two or more joints. It is reasonable to assume that joints or DOFs from completely different body sections, such as the legs and arms, will have very little correlation. Therefore, only the correlation between the DOFs belonging to the body, head, arm, leg and finger sections will be investigated. One way of visualizing the relationship between DOFs is to plot them on a phase space diagram. Figure 5-8 shows such a plot of the left shoulder elevation angle $\theta_{5,1}$ against the other four arm DOFs for the dance sequence. There is a clear clustering behaviour, and it can be concluded that these DOFs are indeed dependant on $\theta_{5,1}$. It becomes tedious to plot every DOF against all the others in a group, and the graph quickly becomes cluttered, especially for long sequences. The rest of the body sections exhibit similar behaviour, and the results are not shown here.

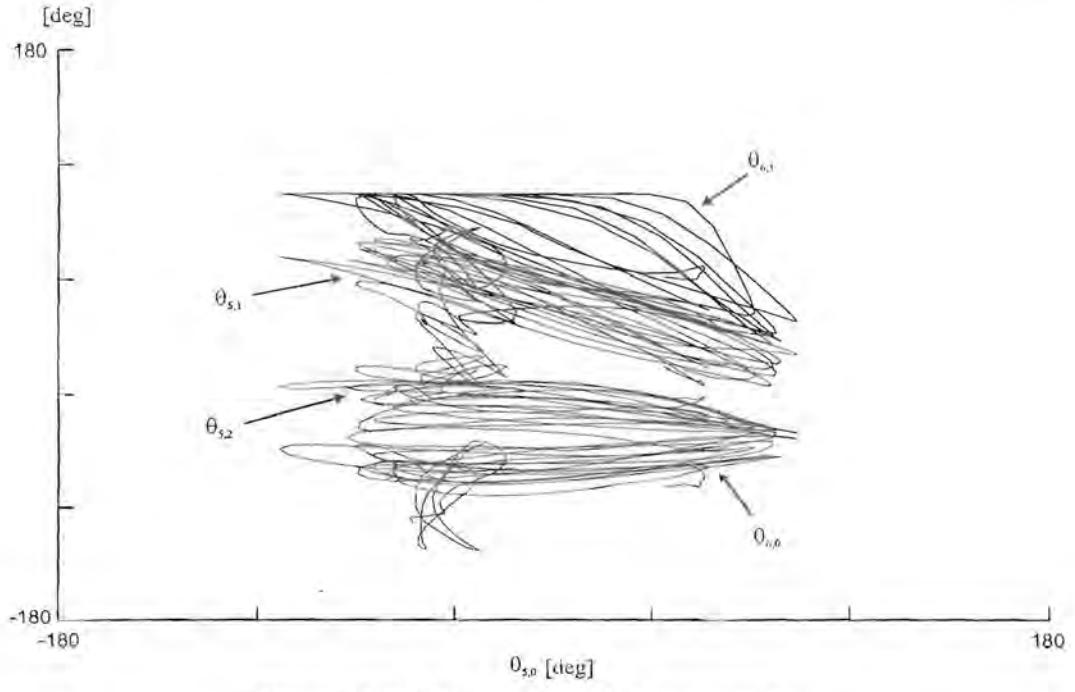

Figure 5-8: Phase plot of left arm angles.

A statistically more correct method is to calculate the joint PDF (where it is understood that joint does not mean physical human joint) for the DOFs in the relevant section. The visual results are kept to two-dimensional PDFs, as it is difficult to visualize more than that in a three-dimensional world. Figures 5-9 a-d show various PDFs of arbitrarily chosen body DOF pairs for the conversational test sequence, while figures 5-10 a-d similar PDFs for the dance sequence. The PDFs are shown as gray scale bitmaps, with a darker value indicating a higher occurrence. They all range from -180° to 180° on both axis. A single point or line on the PDF indicates that one or both of the DOFs is constant, while a large round cluster indicates not much of a cross correlation. However, it is clear from the images that the DOFs are indeed dependent on each other. Figures 5-11a and 5-11b show joint PDFs of arbitrarily chosen finger DOF pairs for the gesture sequence. The range is 0° to 60° on both axis. The finger DOFs are extremely correlated, and this fact will be used later to achieve higher compression ratios.

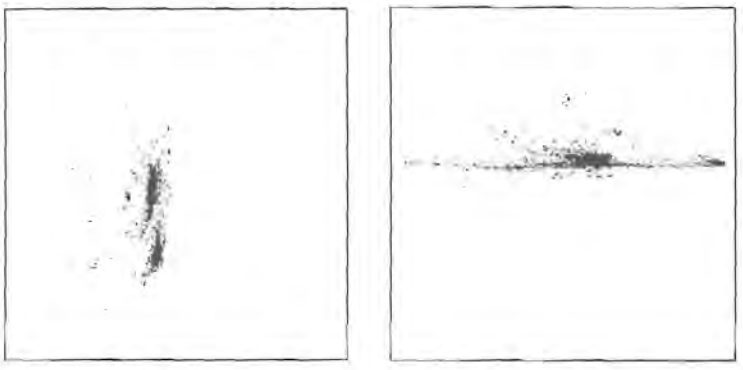

Figure 5-9: a) Joint PDF for $\theta_{5,0}$ and $\theta_{5,1}$ and b) Joint PDF for $\theta_{5,1}$ and $\theta_{5,2}$ for the conversational sequence.

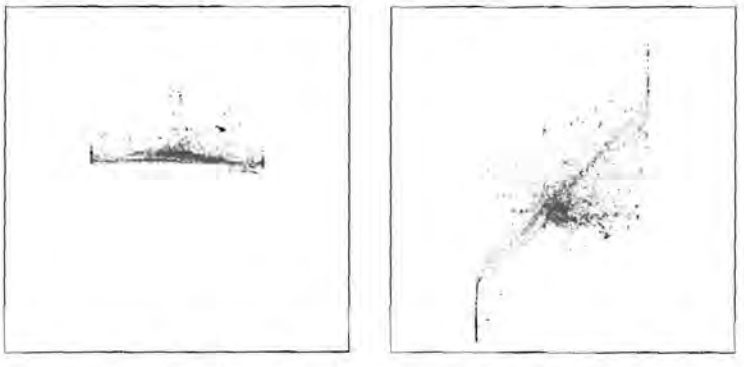

Figure 5-9: c) Joint PDF for $\theta_{5,1}$ and $\theta_{6,1}$ and d) Joint PDF for $\theta_{5,2}$ and $\theta_{6,1}$ for the conversational sequence.

Figure 5-10: a) Joint PDF for $\theta_{5,0}$ and $\theta_{5,1}$ and b) Joint PDF for $\theta_{5,1}$ and $\theta_{5,2}$ for the dance sequence.

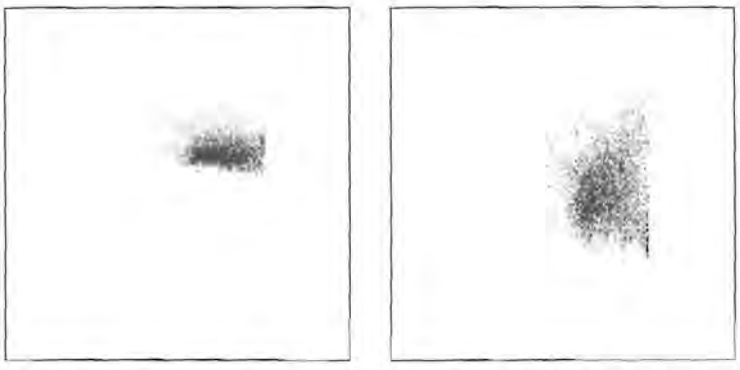

Figure 5-10: c) Joint PDF for $\theta_{5,1}$ and $\theta_{6,1}$ and d) Joint PDF for $\theta_{5,2}$ and $\theta_{6,1}$ for dance sequence.

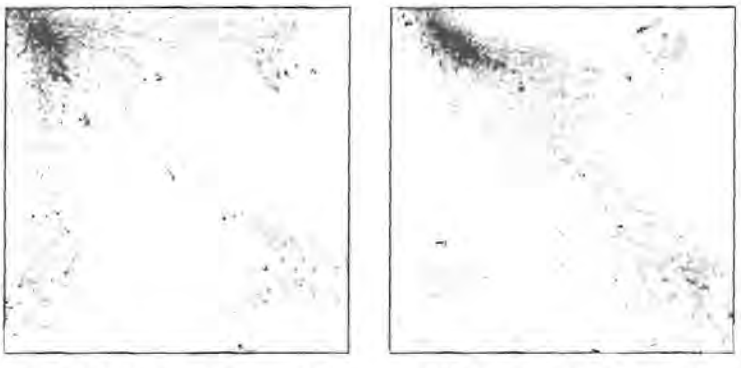

Figure 5-11: a) Joint PDF for $\theta_{27,1}$ and $\theta_{30,1}$ and b) Joint PDF for $\theta_{30,1}$ and $\theta_{33,1}$ for the gesture sequence.

5.2.4 Frequency content

A very important measure of motion information can be found by investigating the frequency content of the motion. It can be used to conclude the minimum sample rate required to capture, process or display human motion. The power spectrum or power spectral density (PSD) of a WSS process $\{\theta_{i,j}(n)\}$ is defined as the Fourier transform of its autocorrelation function. To avoid an impulse at the origin in the case of a process where the mean $\eta_{i,j}$ is non-zero, it is often more convenient to use the autocovariance of the process, which is given by

$$c_{i,j}(m) = E\left[\left(\theta_{i,j}(n) - \eta_{i,j}\right)\left(\theta_{i,j}(n+m) - \eta_{i,j}\right)\right] \tag{5-9}$$

The PSD is then defined by

$$P_{i,j}(\omega) = \sum_{n=-\infty}^{n=\infty} c_{i,j}(n) e^{-j\omega n} . \qquad (5\text{-}10)$$

For a finite length sequence, only an estimate of the PSD can be made, but the term will be used anyway. There are a number of efficient algorithms that can be used to evaluate equation (5-10). We use the Blackman-Tukey [47] method for general PSD calculations and a parametric model based approach for smooth spectra. In the latter case we assume that human motion spectra have broadband characteristics, and autoregressive (AR) parameters are obtained using the autocorrelation method [47].

It is convenient to assume that there are DOFs that exhibit similar frequency behaviour, and to group them together to obtain a combined PSD for the relevant body section. Similar to groupings done elsewhere, we calculate PSDs for the body, head, arm, leg and finger sections. Figures 5-12 a-d show long-term PSDs (i.e. the average PSD of the *whole* sequence) for both the conversational and dance sequences. Figure 5-12e shows the PSD of the fingers for the gesture motion sequence, together with the PSD for arm movement of the same sequence. If a suppression of 50 dB is taken as the cut-off threshold for perceptible motion, then it is clear that the average frequency content is limited to roughly 3 Hz and 6 Hz for the conversational and dance sequences respectively. The finger content is slightly more, which is to be expected since the inertial forces are the smallest on the fingers. However, these results do not imply that the short-term content will necessarily follow the same pattern. Figure 5-13 shows a short term PSD of arm movement at various time intervals for the conversational test sequence, and figure 5-14 a similar PSD for the dance sequence. It can be seen that the short term frequency content stays relatively constant with time. The higher mid-frequencies can clearly be seen across the time range for the dance sequence.

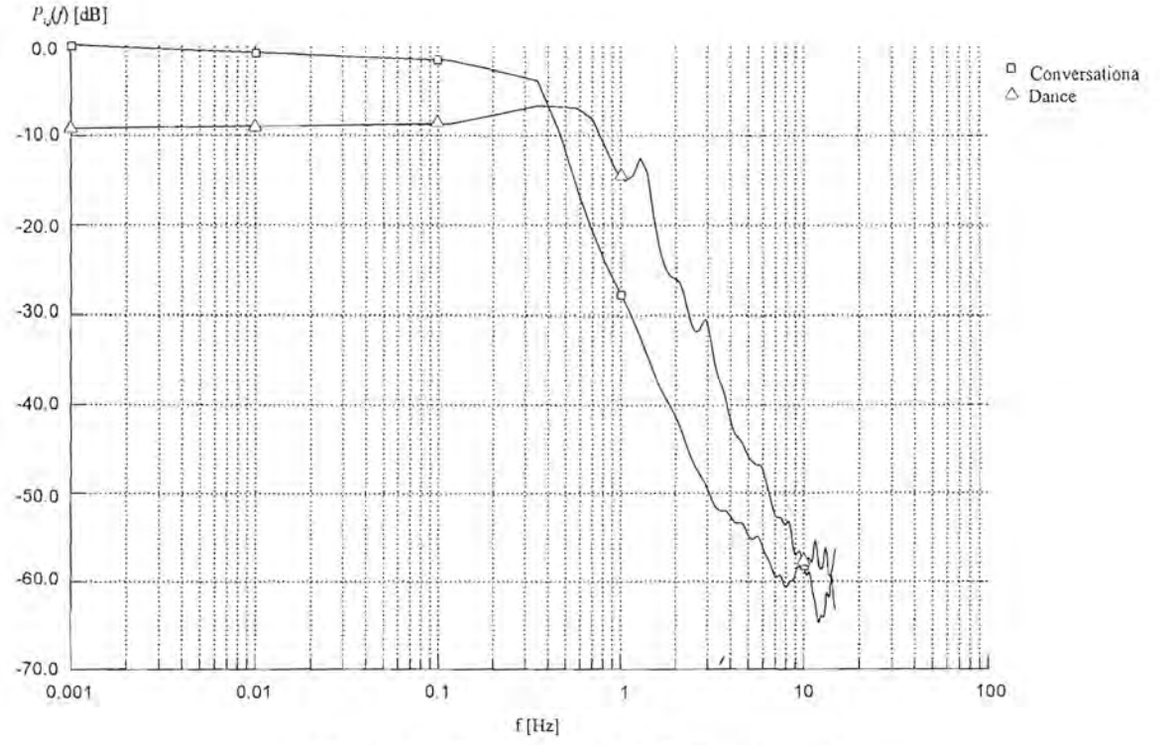

Figure 5-12a: PSD for body movement.

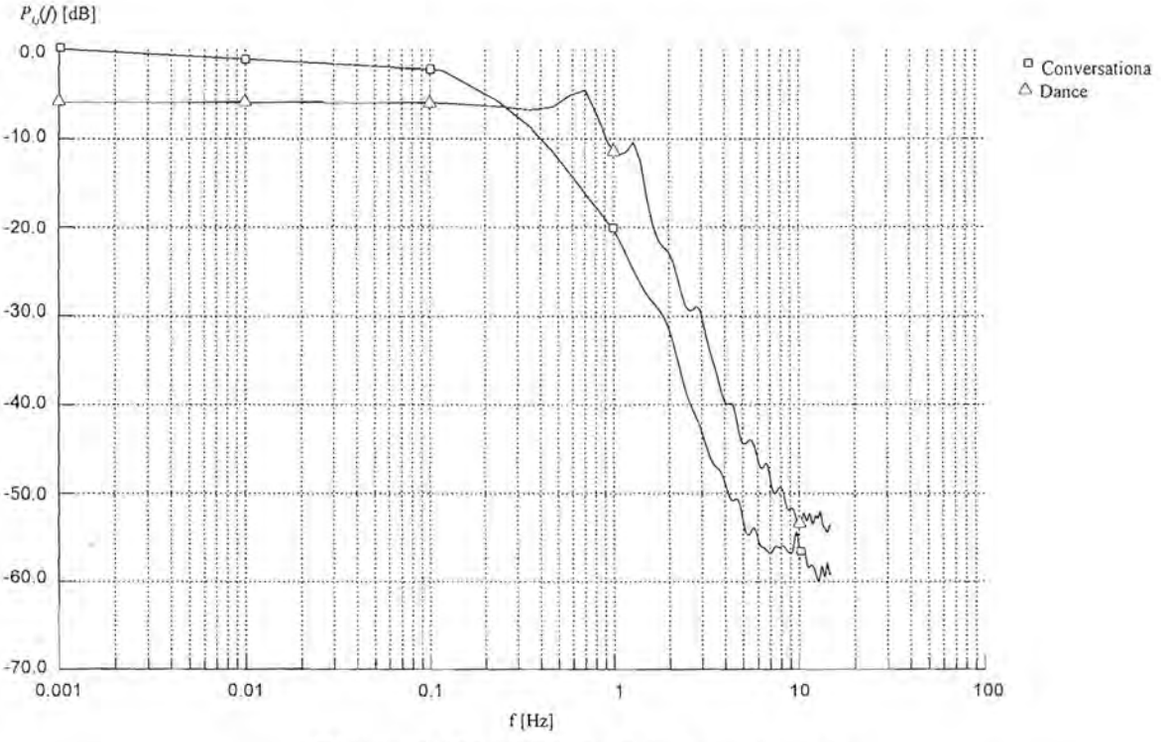

Figure 5-12b: PSD for head movement.

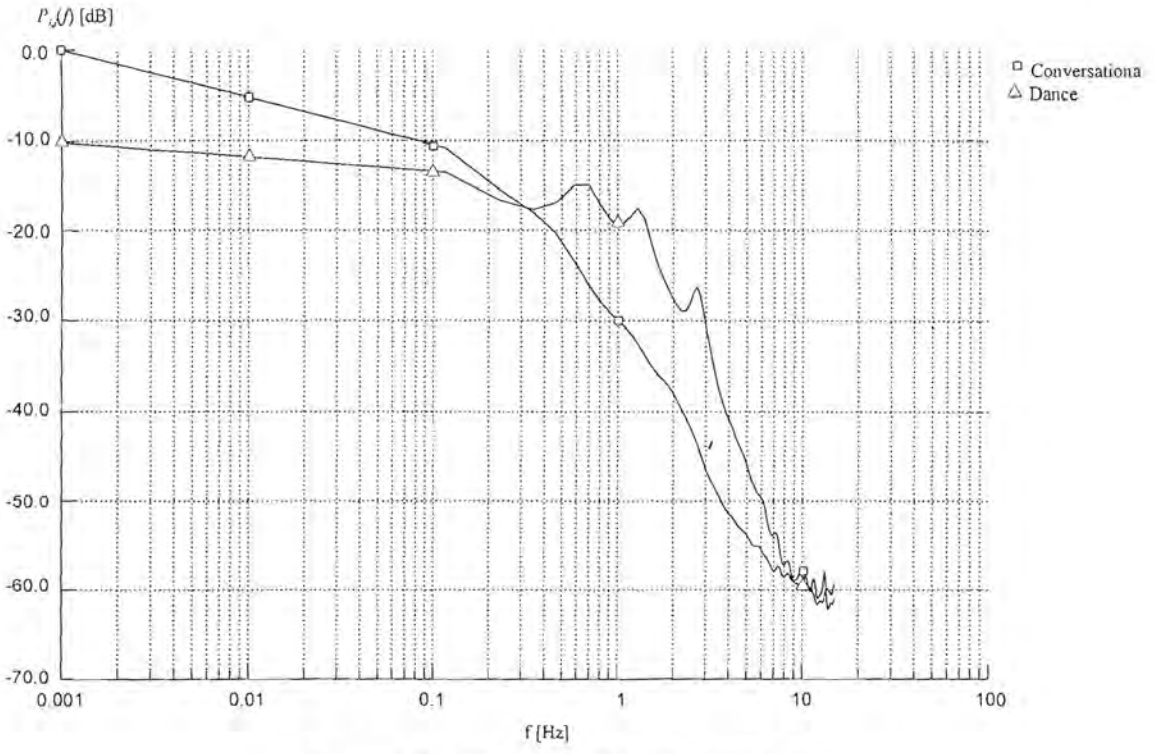

Figure 5-12c: PSD for arm movement.

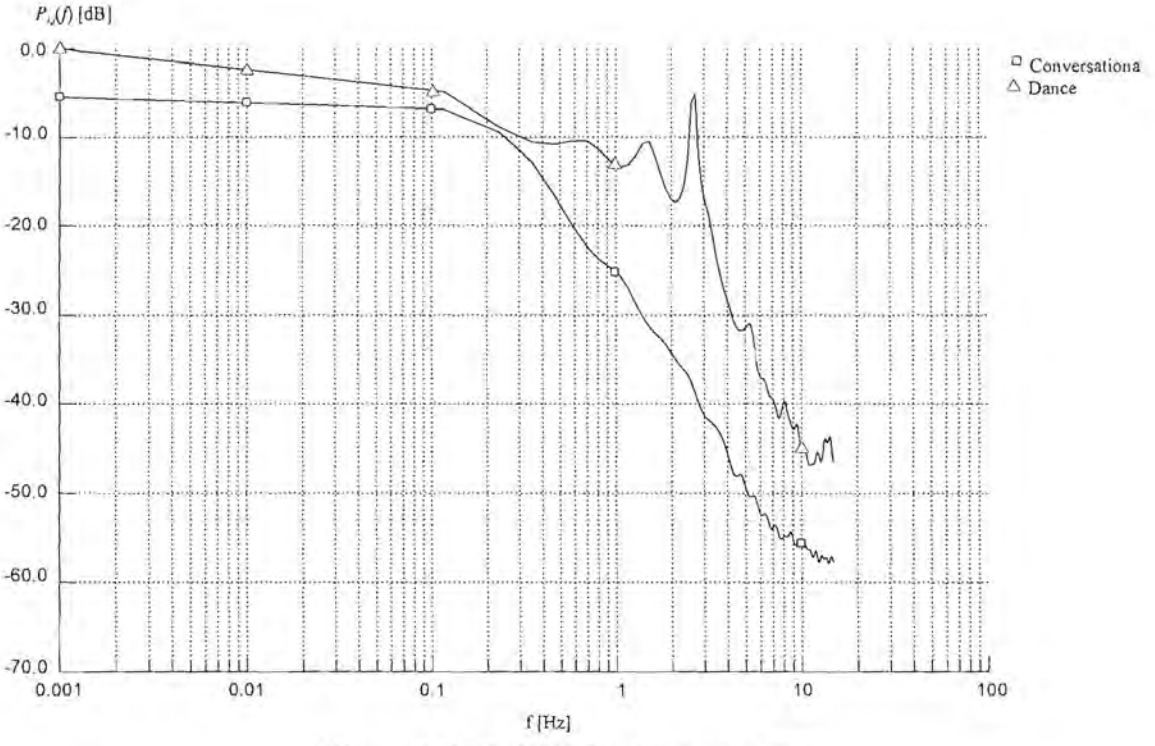

Figure 5-12d: PSD for leg movement.

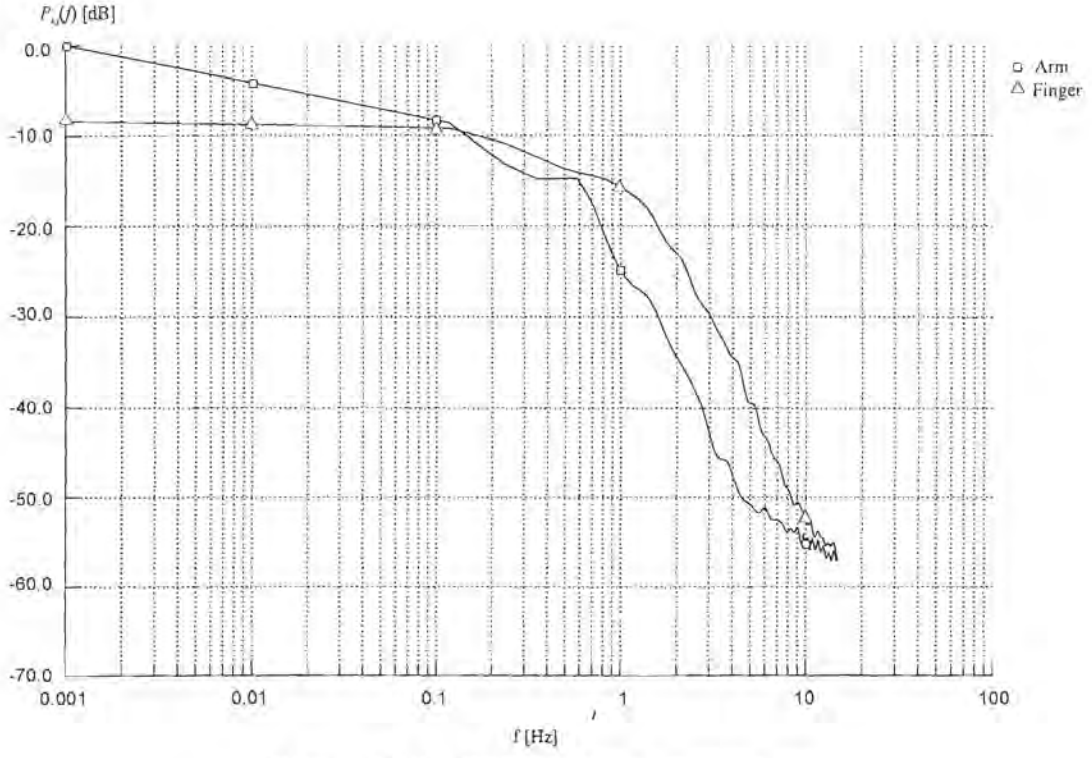

Figure 5-12e: PSD for finger movement.

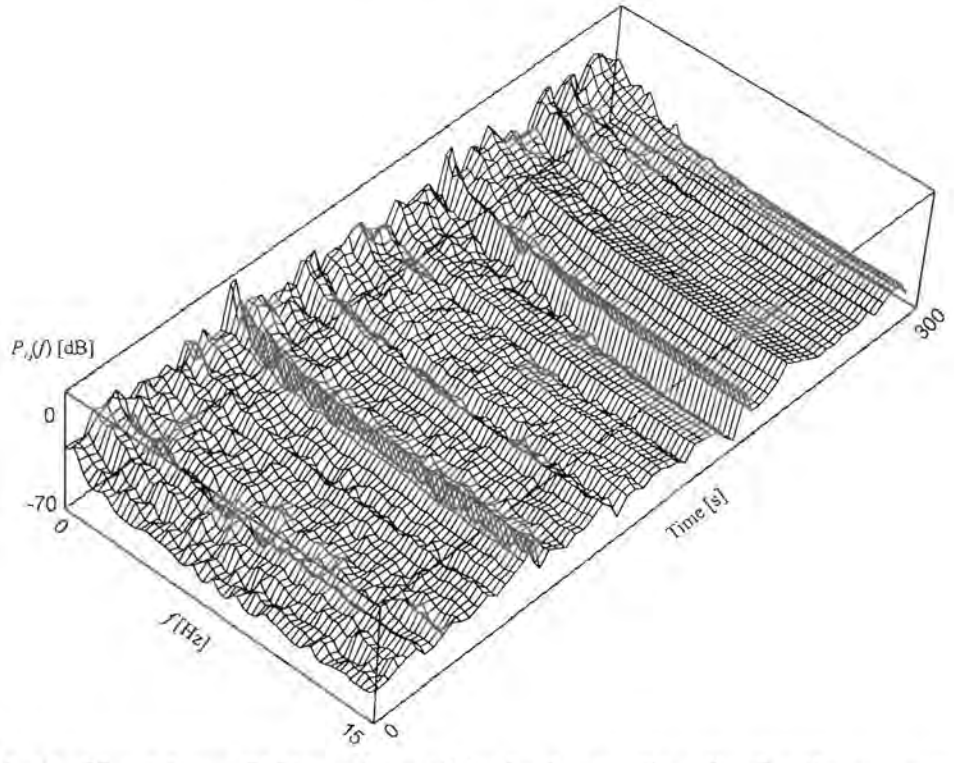

Figure 5-13: Short term PSD vs. time of the body position $\theta_{0,4}$ for the conversational sequence.

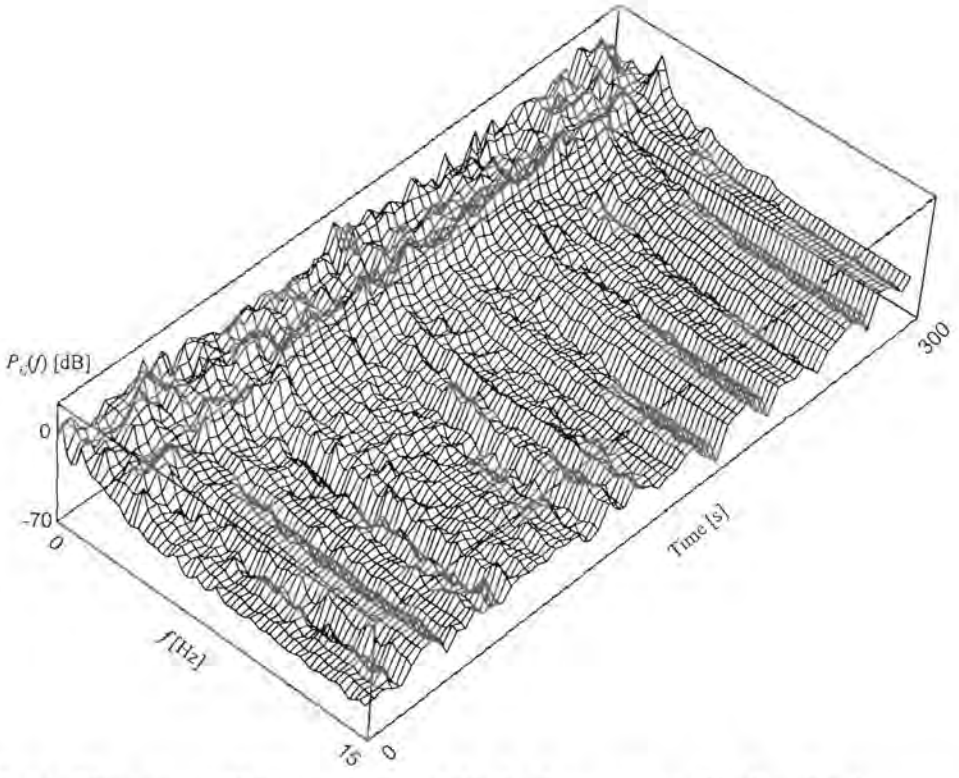

Figure 5-14: Short term PSD vs. time of the body position $\theta_{0,\theta}$ for the dance sequence.

5.3 Summary

This chapter presented a detailed statistical analysis of the human motion captured by the techniques described in chapter 4. The spatial content in terms of range and resolution was investigated, and it was found that these quantities rely on both the nature of the motion as well as the performance of the capturing hardware. Temporal content and statistics were investigated and it was found that it is reasonable to assume the motion data to be ergodic and wide sense stationary. Probability density studies revealed similarities in joint angle behaviour and indicated potential for predictive coding methods. Frequency content analysis indicated that human motion in general is rather band-limited, with the exception of a few peculiar movements. 40 dB cut-off was achieved at as low as 3 Hz for relaxed movement and the frequency content almost never exceeded 8 Hz, even for the dance motion.

Chapter 6 Error measurement

An error measure is a quantitative *or* qualitative indication of the amount of dissimilarity or distortion between two processes. Quantitative measures can be expressed mathematically and the result is some numerical value. Qualitative measures are a bit more difficult to pin down. They are usually given in some descriptive form, such as "it looks horrible". In the following sections, we will give an analysis of a number of error measurement techniques, and their applicability to human motion. Low level coding methods, such as waveform coding, require quantitative error measures that are mathematically tractable. High level methods, such as model based coding play havoc with strictly quantitative error measurements, since there is usually not a one-to-one relationship between the original and coded motion. The best that one can do for model based coding is to define some long-term measurement that will give an indication of the visual quality, or to develop subjective testing mechanisms.

6.1 Quantitative measures

6.1.1 MS error

One of the most common and well-known error measurements is the Mean Square (MS) error. Assume a sequence of values (or degrees of freedom) $\{\theta(n)\}$, and a processed sequence $\{\theta'(n)\}$, which is an approximation of $\{\theta(n)\}$. For clarity the subscript i,j is dropped, and it is understood that the sequence $\{\theta(n)\}$ can represent any DOF. The mean square error is given by

$$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}\left(\theta(n)-\theta'(n)\right)^2, \tag{6-1}$$

for a sequence of length N. The square root of the MS error is sometimes more convenient, and is given by

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}. \tag{6-2}$$

Some useful variations on the MSE are the normalized MSE

$$\mathrm{NMSE} = \frac{\sum_{n=1}^{N}\left(\theta(n) - \theta'(n)\right)^2}{\sum_{n=1}^{N}\theta(n)^2} \tag{6-3}$$

and the peak MSE

$$\mathrm{PMSE} = \frac{\frac{1}{N}\sum_{n=1}^{N}\left(\theta(n) - \theta'(n)\right)^2}{R^2}, \tag{6-4}$$

where R is the range of $\{\theta(n)\}$. The mean square error is often described in logarithmic or decibel form as an equivalent signal-to-noise ratio (SNR)

$$\mathrm{NSNR} = -10\log_{10}(\mathrm{NMSE}), \tag{6-5}$$

or

$$\mathrm{PSNR} = -10\log_{10}(\mathrm{PMSE}). \tag{6-6}$$

Mean square error measurements are generally used as an evaluation tool after some process or operation has been completed, i.e. it is performed on a whole sequence of values.

6.1.2 Instantaneous error

Many of the compression algorithms require an error measurement that is applicable to the current frame or update. The best that can be done in this case is to use a distance metric as the error measurement. If $\theta(n)$ is a value at the nth sample, and $\theta'(n)$ is an approximation of $\theta(n)$, the distance is simply given by

$$d(n) = |\theta(n) - \theta'(n)|. \tag{6-7}$$

Sometimes it is desirable to use a metric that is mathematically more tractable (such as being easily differentiable), and we can use

$$d(n) = (\theta(n) - \theta'(n))^2. \tag{6-8}$$

6.1.3 Vector error

Although the distance errors specified in equation (6-7) and (6-8) are useful on their own, it is often convenient to group a number of dependent variables together as a vector and use their combined error (the reasons for doing so are explained in more detail in chapters 6 and 7). Mathematically it serves no purpose to group independent variables, as the uncorrelated result will be meaningless to the compression algorithm. Table 6-1 repeats the grouping scheme, together with the number of joints, segments and DOFs for each group. Refer to the human skeleton representation in figure 3-4 for details.

Table 6-1: Joint and segment grouping

Group	Reference number	Number of joints	Number of segments	Number of DOFs
Root and torso	0	2	2	7
Neck and head	1	2	2	5
Left arm	2	3	3	8
Left hand	3	14	14	19
Right arm	4	3	3	8
Right hand	5	14	14	19
Left leg	6	3	3	7
Right leg	7	3	3	7

Although we do not attempt to prove it here, it is reasonable to assume that the above combined joints or variables are correlated to some extent. The body specification of MPEG-4 uses a similar grouping scheme [39].

We denote the sequence of a group or vector of DOFs by $\{\theta_i(n)\}$, and the vector of the approximated DOFs by $\{\theta'_i(n)\}$, where $\{i = 0...7\}$. Individual components of the vector are denoted by $\{\theta_{i,j}(n)\}$ or $\{\theta'_{i,j}(n)\}$, where $\{i = 0...7, j = 0...K\text{-}1\}$ and K is the number of DOFs of the ith vector as given in table 6-1. Using this notation, we define the normalized weighted vector error of the ith group for the nth sample as

$$w_i(n) = \frac{1}{K}\sum_{j=0}^{K-1}\frac{a_{i,j}\left(\theta_{i,j}(n)-\theta'_{i,j}(n)\right)^2}{b_{i,j}^2}, \tag{6-9}$$

where K is the number of DOFs and $b_{i,j}$ is the range of the jth DOF. The quantity $a_{i,j}$ is a weighing coefficient that defines the contribution of the jth DOF to the error. If $a_{i,j}$ is in [0, 1], then $w_i(n)$ will be in [0, 1] with lower values indicating a good match. The values for $a_{i,j}$ and $b_{i,j}$ can also be defined in such a manner that the quantity $w_i(n)$ has meaningful units, such as [deg^2].

We are often interested in the *maximum* error for a group of joints, instead of a linear combination of errors. The maximum normalized weighted error for the nth sample of the ith group is given by

$$m_i(n) = MAX\left(\frac{a_{i,j}\left|\theta_{i,j}(n)-\theta'_{i,j}(n)\right|}{b_{i,j}}\right), \quad 0 \le j < K, \tag{6-10}$$

where K is the number of DOFs and $b_{i,j}$ is the range of the jth DOF. The quantity $a_{i,j}$ is a weighing coefficient that defines the contribution of the jth DOF to the error. If $a_{i,j}$ is in [0, 1], then $m_i(n)$ will be in [0, 1] with lower values indicating a good match. It should be

noted that equation (6-10) is not easily differentiable (compared to equation (6-9)), and is not very useful in error minimizing algorithms.

Equations (6-7) to (6-10) can also be used on a sequence of values similar to the definition of MS error and its variants. For example, we can write

$$W_i = \frac{1}{N}\sum_{n=1}^{N} w_i(n) \tag{6-11}$$

as the normalized weighted error on a whole sequence of length N. The maximum error can be redefined in a similar manner.

6.1.4 Joint and segment errors

Cases of special interest in human motion analysis are those of joint and segment position and/or orientation error, which are often geometrically more meaningful and intuitive than individual joint angle errors. By taking three-dimensional volume displacement into consideration, we get a bit closer to visual based comparisons between various body postures. We denote a sequence of joint positions by $\{\mathbf{u}_{i,j}(n)\}$, and that of the approximated joints by $\{\mathbf{u}'_{i,j}(n)\}$, where $\{i = 0...7, j = 0...K\text{-}1\}$. K is the number of joints for the ith group, and is given in table 6-1. The joint position error for the ith group of joints is given by

$$p_i(n) = \frac{1}{K}\sum_{j=0}^{K-1} \frac{a_{i,j}\left\|\mathbf{u}_{i,j}(n) - \mathbf{u}'_{i,j}(n)\right\|}{b_{i,j}}, \tag{6-12}$$

where $a_{i,j}$ and $b_{i,j}$ are weighing and normalizing coefficients similar to equation (6-9) and (6-10). It is often more meaningful to define the coefficients such that $p_i(n)$ has units of meters. The coefficients $a_{i,j}$ can also be defined as an impulse function to obtain the error for a single joint in the group.

When the axis of rotation is parallel to the rotated segment, the use of equation (6-12) on its own can sometimes result in complete failure to detect a rotation error. An example of this is the upper and lower arm twisting motion, both of which can result in a constant elbow or wrist joint position. To satisfy both position and rotation errors in a single generalized equation, we define additional DOFs for each group. The original and additional DOFs are grouped together in a *configuration* vector **c**. Configuration vectors describe both joint position and rotation. For example, the 14-dimensional configuration vector for the left arm group would be given by

$$\mathbf{c}_2 = [e_x\ e_y\ e_z\ w_x\ w_y\ w_z\ \theta_{2,0}\ \theta_{2,1}\ \theta_{2,2}\ \theta_{2,3}\ \theta_{2,4}\ \theta_{2,5}\ \theta_{2,6}\ \theta_{2,7}], \tag{6-13}$$

assuming that the shoulder is fixed at the world origin [0 0 0]. The additional DOFs are the elbow position, which is given by $[e_x\ e_y\ e_z]$, and the wrist position, which is given by $[w_x\ w_y\ w_z]$. A sequence of configuration vectors for the ith group is written as $\{\mathbf{c}_i(n)\}$, and individual components as $\{c_{i,j}(n)\}$. Using similar notation as in equation (6-9), the generalized error for the ith group can be written as

$$\varepsilon_i(n) = \frac{1}{K}\sum_{j=0}^{K-1}\frac{a_{i,j}\left(c_{i,j}(n) - c'_{i,j}(n)\right)^2}{b_{i,j}^2}, \tag{6-14}$$

where K is the number of elements in the configuration vector. Given proper coefficients, equation (6-14) is a useful error measure under many conditions. We obtained suitable values for $a_{i,j}$ and $b_{i,j}$ for equations (6-9), (6-10), (6-12) and (6-14) using heuristic methods and subjective testing.

6.2 Visual measures

Visual error measurement implies a method that will tell us whether the visual posture and motion of the human figure are acceptable, and if possible, to what extent. It should be noted that there is often a vast difference between a visual measure and a strictly mathematical measure. If the animation has natural and pleasing motion, it does not

necessarily mean it has the correct original position or orientation. Visual measurement techniques often rely on subjective tests by a panel of viewers. Many parameters of our compression techniques were obtained in this manner. However, it need not be done only subjectively. In fact, it would be desirable to define an objective visual measure that is mathematically tractable. When seeking such a solution, there is often no clear mathematical relationship between the original quantity and distorted quantity, and we are forced to look at the characteristics of these quantities separately.

6.2.1 Natural movement

One method of identifying visual artifacts is by evaluating the joint angles and their first and second derivatives for discontinuities or abnormally large values. Naturally, if both the original and coded values contain such anomalies not much can be said about the error. However, if the decoded motion exhibits values that are out of bounds compared to the original, it is reasonable to assume that something had gone wrong in the coding process. A more advanced method than simply identifying discontinuities is to compare the decoded human motion with dynamically simulated motion. One way of doing this is to calculate the metabolic energy spent in performing a motion, and to compare it to the original. It has been established that humans try to accomplish movement using the least amount of energy [44]. Abnormally large values indicate unnatural movement, and can be considered as an error in the coding process. Unfortunately, the methods described above rely primarily on the decoded sequence. We need at least some reference to the original sequence, otherwise the error between completely different original and decoded actions will be pronounced acceptable.

Discontinuities and unnatural movement aside, common errors on a waveform level are primarily due to *phase* and *amplitude* differences[1]. Phase errors are usually generated by coding delay and motion interpolation approximations. Amplitude errors are primarily generated by quantization in the spatial, temporal and frequency domains. We have found

[1] Not to be confused with the actual amplitude and phase functions of the original signal.

that phase errors are visually more tolerable than amplitude errors, especially high frequency amplitude errors. For example, spatial quantization generates high frequency discontinuities and jerkiness, and the differentiating characteristics of the human visual system causes such errors to be perceived as visually annoying. It is common practice to compensate for the (known) coding delay when calculating quantitative errors. The remaining phase error is therefore primarily a function of the compression method. These errors vary relatively slowly over time compared to quantization errors, which can occur at every sample. Phase errors in general result in fewer high frequency discontinuities and artifacts.

6.2.2 Visual MS error (VMSE)

The observed low and high frequency relationship between phase and amplitude errors led us to develop the visual mean square error, or VMSE. Figure 6-1 shows a conceptual diagram of the method. The difference between the original and coded signal (i.e. the error) is divided into a number of frequency bands, each is assigned a certain weight, and the results are combined again. By adjusting the coefficients α_i, the importance of various visual dissimilarities and artifacts that exist between the original and coded sequences can be set. Naturally, by setting all of the coefficients to unity, the VMSE measurement reduces to the normal MSE measurement. Similar to the MS error defined in equation (6-1), it is understood that by *signal* we mean any DOF, and that the VMSE of the total figure is given by the sum of the VMS errors of some or all of the DOFs. Mathematically, the VMSE can be written as

$$VMSE = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{m=1}^{M}\alpha_m e_m(n)\right)^2, \qquad (6\text{-}15)$$

for a sequence of length N, with M frequency bands. The quantity $e_m(n)$ is the output of the mth bandpass filter. The filtering can be implemented in any number of convenient ways. Similar to equations (6-4) and (6-6), the peak visual mean square error is defined as

$$VPMSE = \frac{VMSE}{R^2}, \tag{6-16}$$

where R is the range of the original input sequence. The equivalent peak signal-to-noise ratio is defined as

$$VPSNR = -10\log_{10}(VPMSE). \tag{6-17}$$

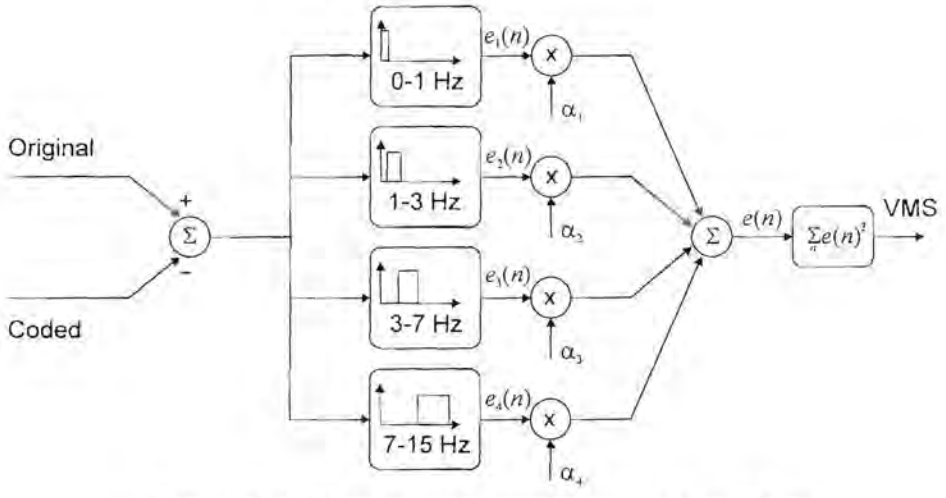

Figure 6-1: Visual mean square error algorithm.

Figure 6-2 shows a comparison between the normal MS measurement and the visual MS measurement (using the peak signal-to-noise ratio variation). The rate axis indicates a dimensionless quantity chosen for convenience. We simulate noisy amplitude errors by quantizing a signal to various levels, and phase errors by shifting a signal in time by various amounts. The amplitude errors are visually quite obvious, while the phase errors are indistinguishable without reference to the original sequence. Although this is an oversimplification of errors encountered from real compression methods, it gives an indication of what to expect from best and worst case scenarios. We use the filter banks as shown in figure 6-1, i.e. the error signal is divided into four consecutive frequency bands, with bandwidth increments by a power of two starting at one. The coefficients were heuristically chosen as $\alpha_i = \{0.25, 0.5, 1, 2.25\}$, i.e. low frequency and mean errors are subdued while high frequency errors are emphasized. In chapter 5 it was shown that the

original signal contains very little or no high frequency components. It is therefore in order to set α_4 to quite a high value, since errors in this band can originate only from the compression method. The coefficients defined above clearly forms a high-pass filter and equation (6-15) could have been implemented as such. However, we have found it more intuitive to work with a number of discrete frequency bands, each representing a certain type of visual artifact. For example, the lowest frequency band contains the general gist of the motion, while the middle frequency bands add *emotion* to the movement. High frequency bands contain jerky behaviour, which is often a result from quantization errors.

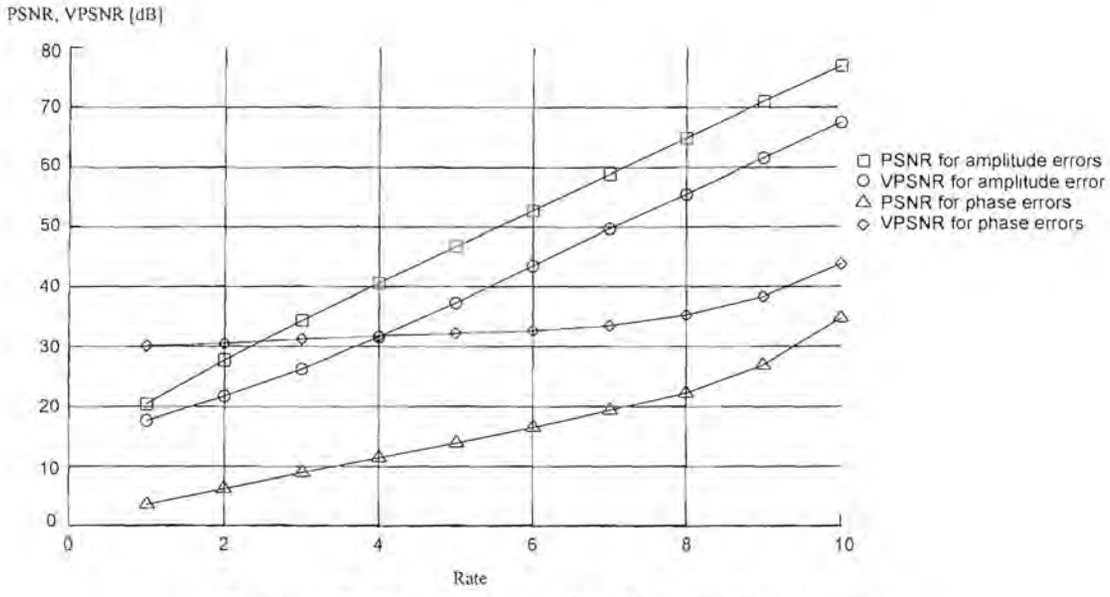

Figure 6-2: PSNR vs. VPSNR for simulated errors.

It is clear from figure 6-2 that the VMSE measure consistently indicates a lower SNR compared to the MSE measure for high frequency amplitude errors. Severe quantization results in long constant values with occasional high frequency jumps to adjacent levels. In this case, it can be seen that the VMSE starts to favour the low frequency errors introduced by these constant values. As is to be expected, at high quantization levels the error diminishes (i.e. the coding becomes lossless), and the two measures converge (not shown). In the case of phase errors, the MSE measure starts failing even for moderate errors. In this case the VMSE in figure 6-2 shows a clear advantage, which is consistent with the visual appearance of the errors.

6.3 Summary

This chapter presented a number of error measurement techniques. A distinction was made between purely quantitative methods such as the naive mean square error (MSE) measure, and qualitatively motivated methods such as the newly proposed *visual* mean square error (VMSE). Quantitative methods such as the MSE and its variants are easy to implement, are mathematically tractable and are suitable for direct implementation in a wide variety of compression algorithms. However, these methods clearly failed to distinguish acceptable error artifacts from annoying visual errors such as severe quantization noise. In order to accommodate visual errors the VMSE was introduced, which is similar in concept to the noise-shaping error measures used in speech coding.